UNIVERSITY OF TWENTE

MASTER THESIS

The Effect of Voice and Embodiment on the Gender Perception of Speaking Social Robots

Author:
Sjoerd van Veen

Supervisors:
Dr. Khiet Truong
Dr. Cesco Willemse
Hideki Garcia Goo
Ella Velner

Human Media Interaction Group
Faculty of Electrical Engineering, Mathematics, and Computer Science

December 6, 2024

UNIVERSITY OF TWENTE

Abstract

Faculty of Electrical Engineering, Mathematics, and Computer Science

Master of Science

The Effect of Voice and Embodiment on the Gender Perception of Speaking Social Robots

by Sjoerd van Veen

Synthetic computer voices and social robots are seeing increased use as (virtual) assistants, yet research shows that these technologies reinforce harmful gender stereotypes, prompting new research into the gender perception of technology to allow its designers to make more informed decisions about its perceived gender. However, this research has mainly focused on computer voices or robots in isolation, rarely combining the two. As such, it remains unclear how the voice and embodiment (appearance) of a speaking social robot influence its perceived gender. This study addresses this gap by investigating the effect of computer voices and embodiments on the gender perception of speaking social robots, and each other. Additionally, it investigates different methods for the creation of ambiguously gendered speaking social robots. An online survey was conducted, in which robots and voices of ambiguous, feminine, and masculine gender were combined into robot-voice combinations and then tested for their perceived gender. The results indicate dominance of the voice over the embodiment for the gender perception of speaking social robots, as the voices showed greater effects on the gender scores than the robots. However, this is partially dependent on the gender of the voice, as ambiguously gendered stimuli, whether voice or embodiment, are found to have little impact when combined with a binary gendered stimulus. Combinations of masculine and feminine stimuli are found to score high on ambiguity, similar to combinations consisting solely of ambiguous stimuli. However, remarks from participants indicate unfavourable attitudes towards such combinations, suggesting additional research is necessary to determine if combining masculine and feminine stimuli is a viable method for the creation of gender ambiguous speaking social robots.

Acknowledgements

First and foremost, I would like to thank Khiet, for introducing me to this topic and being my main supervisor. I also want to thank her together with my other supervisors, Hideki, Ella, and Cesco for all their help, support, and feedback throughout this entire project. Without them, this would not have been possible. I have to say a big thank you to Tim as well, for always helping, supporting and motivating me, and for all the time we spent together while working on our own projects. I would like to thank my parents and my brother for always supporting me and making this possible. Last but not least, I would like to acknowledge and extend my appreciation to everyone I met throughout my exchange semester at Umeå University during this project, for inspiring and motivating me even if they did not realise it, with a special thanks to Jonathan, Mirjam, Lea, Sunna, Antonia, and Aelia.

Contents

Abstract iii

Acknowledgements v

1 Introduction 1

2 Related Work 3
- 2.1 Terminology . . . 3
  - 2.1.1 ‘Gender’ . . . 3
  - 2.1.2 ‘Gender ambiguous’ . . . 3
- 2.2 Anthropomorphism . . . 4
  - 2.2.1 Defining Anthropomorphism . . . 5
  - 2.2.2 Anthropomorphism & Stereotypes . . . 5
  - 2.2.3 Anthropomorphism & Gender . . . 6
- 2.3 Gendering Social Robots . . . 7
- 2.4 Gendering Computer Voices . . . 8
- 2.5 Auditory vs. Visual Stimuli . . . 9
- 2.6 Measuring Gender Perception . . . 10
  - 2.6.1 Multiple-Choice . . . 10
  - 2.6.2 Single Scale . . . 10
  - 2.6.3 Multiple Scales . . . 11
  - 2.6.4 The Naming Technique . . . 11

3 Hypotheses 13

4 Pre-test 15
- 4.1 Methods & Materials . . . 15
  - 4.1.1 Participants . . . 15
  - 4.1.2 Robots . . . 15
  - 4.1.3 Computer Voices . . . 17
  - 4.1.4 Measures . . . 18
  - 4.1.5 Procedure . . . 19
  - 4.1.6 Analysis . . . 20
- 4.2 Results . . . 20
  - 4.2.1 Robots . . . 20
  - 4.2.2 Voices . . . 21
- 4.3 Discussion . . . 22
  - 4.3.1 Robots . . . 22
  - 4.3.2 Voices . . . 23
- 4.4 Conclusion . . . 25

5 Main Experiment **27**
5.1 Methods & Materials . . . 27
5.1.1 Participants . . . 27
5.1.2 Robots & Voices . . . 27
5.1.3 Robot-Voice Combinations . . . 28
5.1.4 Measures . . . 28
5.1.5 Procedure . . . 28
5.1.6 Analysis . . . 29
5.2 Results . . . 30
5.2.1 General Overview . . . 30
5.2.2 Gender Categorisation . . . 31
5.2.3 Masculine Base Models . . . 35
5.2.4 Feminine Base Models . . . 37
5.2.5 Ambiguous Base Models . . . 38

6 Discussion **43**
6.1 RQ1 — The influence of embodiment and voice on the gender perception of speaking social robots, and each other . . . 43
6.2 RQ2 — The most promising robot-voice combination for the creation of gender ambiguous speaking social robots . . . 44
6.3 Limitations . . . 46
6.4 Design Recommendations . . . 47

7 Conclusion **49**

A Ambiguous Robot Images **51**

B Feminine Robot Images **53**

C Masculine Robot Images **55**

D Voice Fundamental & Formant Frequencies **57**

E Pre-Test Survey: Voice Excerpt **59**

F Pre-Test Survey: Robot Excerpt **61**

G Pre-Test: Robot Gender Perception Results **63**

H Pre-Test: Voice Gender Perception Results **65**

I Robot Height Indication **67**

J Main Experiment Survey: Combination Excerpt **69**

K The Harmony Robot **71**

L Main Experiment: Isolated Gender Perception Results **73**

M Main Experiment: Combination Gender Perception Results **75**

N Main Experiment: Masculine Base Results **79**

O Main Experiment: Feminine Base Results **83**

P Main Experiment: Ambiguous Base Results 87

Bibliography 89

List of Figures

4.1 Gender perception questions for robots and voices 19
4.2 The masculine and feminine FGSs, and AbsDiffs of all the robots . . . 21
4.3 The masculine and feminine FGSs, and AbsDiffs of all the voices . . . 22

5.1 Masculine gender perception question for robot-voice combinations . . 28
5.2 The gender perception results of the voices and robots when presented in isolation. 31
5.3 Bar plot showing the mean change (with standard deviation) of the masculinity score of a base after a robot or voice addition. 32
5.4 Bar plot showing the mean change (with standard deviation) of the femininity score of a base after a robot or voice addition. 32
5.5 Bar plot showing the mean change (with standard deviation) of the neutrality score of a base after a robot or voice addition. 33
5.6 Bar plot showing the mean change (with standard deviation) of the AbsDiff of a base after a robot or voice addition. 33
5.7 Interaction plot with 95% confidence intervals of the masculinity-delta model of the masculine bases 36
5.8 Interaction plot with 95% confidence intervals of the femininity-delta model of the feminine bases 37
5.9 Interaction plot with 95% confidence intervals of the neutrality-delta model of the ambiguous bases 39
5.10 Interaction plot with 95% confidence intervals of the AbsDiff-delta model of the ambiguous bases 39

List of Tables

4.1 An overview of the robots and their placeholder names 20
4.2 Comparison between pre-test results and ROBO-GAP database for masculine and feminine robots 23
4.3 Comparison between pre-test results and ROBO-GAP database for ambiguous robots 23
4.4 The top 5 lowest AbsDiff and highest neutrality scores of the voices . . 24

5.1 The results from the gender categorisation of the robot-voice combinations 34
5.2 The linear mixed model results for the masculinity-delta of masculine bases 36
5.3 The linear mixed model results for the femininity-delta of feminine bases 38
5.4 The linear mixed model results for the neutrality-delta of ambiguous bases 40
5.5 The linear mixed model results for the AbsDiff-delta of ambiguous bases 40

Chapter 1

Introduction

In 2019, UNESCO published ‘I’d blush if I could,’ a report on the gender biases present in the field of digital technology [1]. The title refers to the response from Siri, Apple’s (female) voice assistant, to a user saying “hey Siri, you’re a bitch.” Although this response has been removed, it still serves as an example of what UNESCO refers to as the ‘obsequiousness’ of Siri and other ‘female’ digital voice assistants, that is, the servile compliance or prompt obedience as well as the submissiveness in the face of abuse, portrayed by these voice assistants.

While the boundless subservience of voice assistants seemingly encourages impolite and excessively direct speech [2], a more pressing issue is also presented. Both the representation, and the reinforcement of gender stereotypes through these behaviours from voice assistants, have been cause for concern [1], and while not explicitly mentioned in the given example, these same concerns apply to (social) robots, putting them in a similarly precarious position regarding the proliferation of gender stereotypes (e.g. [3]). To combat the presence of gender biases and stereotypes in digital technologies like voice assistants and robots, in the face of massively increasing use of both [4][5], UNESCO recommend, among other things, the exploration of the feasibility of a ‘machine gender’ that is neither obviously male nor female [1].

Whether this recommendation was the catalyst or not, recent years have seen an abundance of research on gender perception as a whole, as well as ‘gender neutrality’ or ‘ambiguity,’ in computer voices and (social) robots, though this has been quite fragmented. Research with computer voices (naturally) focuses on the sound of the voice (e.g. [6]), whereas research with social robots mainly focuses on physical appearance (e.g. [7]). Little effort has been made to consolidate these fields, despite many social robots having the ability to speak. Consequently, while it is clear how different voice parameters may impact the gender perception of a computer voice, or how different appearance factors may impact the gender perception of a robot, it remains unclear if and how these may interact, which could prove essential in understanding the gender perception of a speaking social robot.

Researchers in other fields have previously examined whether human faces or voices have greater importance in the process of gender perception (e.g. [8]). However, it is debatable whether such results can be generalised to speaking social robots, especially considering the difference in visual stimuli (human face vs. robot appearance). Thus, as similar research with robots has barely been conducted, a knowledge gap persists.

This knowledge gap in the context of speaking social robots is the main focus of the current research. Specifically, the aim is to find whether the embodiment (the appearance) or the voice has more influence on the gender perception of a speaking social robot. Additionally, the current research aims to find the most promising combination of gendered embodiment and voice, for the creation of gender ambiguous speaking social robots. The knowledge acquired here can aid in the creation of design guidelines, to ensure proper coordination of voice and embodiment design with respect to their impact on gender perception, thereby limiting the risk of voice and embodiment interfering with each other in pursuit of a gendered design goal. Correspondingly, this knowledge could also facilitate and support further research on the creation of gender ambiguous social robots.

From this aim emerge the main research questions:

RQ1 What is the influence of *a)* voice and *b)* embodiment on the gender perception of speaking social robots?

This can be split into two sub-questions: *(a)* how does voice affect the gender perception of the embodiment of a social robot? And *(b)* how does embodiment affect the gender perception of a synthetic voice? And:

RQ2 What combination of gendered embodiment and gendered voice yields the most promising approach for the creation of gender ambiguous speaking social robots?

This report first covers related works regarding anthropomorphism, gender in robots, gender in computer voices, and the agency of visual vs. auditory stimuli, to provide a background for the rest of the research. A pre-test is then conducted to select proper stimuli for the main experiment aimed at answering the research questions. Results from the main experiment are presented and discussed, followed by a conclusion of this report and recommendations for future research.

Chapter 2

Related Work

2.1 Terminology

To start, some key terminology used in this paper should be discussed and clarified. These terms may be familiar, but their definition can remain elusive. This section discusses what these terms will mean in the context of this paper.

2.1.1 'Gender'

Gender is defined by the WHO [9] as a social construct, referring to norms, behaviours, and roles associated with being a man or woman. It is distinctly different from sex, which refers to the different biological characteristics of men, women, and intersex [10]. Gender identity refers to a person's internal, individual experience of gender, which may or may not correspond to their sex.

In the context of technology, the term sex is not relevant. Still, people do give technologies a gender identity through the aforementioned social construct. This applied gender identity has been denoted as male/female, man/woman, or masculine/feminine in the past. For the remainder of this study, gender identity will simply be referred to as gender, as this is often already the case [11]. Gender will be denoted as masculine/feminine, though other terms may be used when referring to papers that used other terms.

2.1.2 'Gender ambiguous'

The term gender ambiguous was first introduced by Sutton [12]. She argues that voices cannot be genderless or gender neutral, rather they can only be gender ambiguous (def. *(1)* doubtful or uncertain, *(2)* open to more than one interpretation [13]). Referencing back to Mullennix et al. [14] who did several experiments on perception, she explains that the use of a 6-point scale to rate gender (with male and female on either end of the spectrum) did not reveal a clear boundary between male and female, instead displaying a steady transition between them, showing the presence of gender ambiguity in some voices. But when participants could only answer 'male', 'female', or 'other', barely anyone selected 'other', which essentially questions the existence of a non-binary voice altogether. Furthermore, while genderless implies gender is not present, and gender neutral implies gender has been removed or made ineffective in some way, Sutton argues it is highly questionable whether listeners will actually perceive a 'genderless' voice as neither male nor female. This is due to the predominant conceptualisation of gender being binary. Finally, she also brings up ethical concerns, stating that (implied) genderlessness of voices sidesteps issues of sexism in technology design. Envisioning a scenario where technology designers relinquish culpability for any negative consequences related to gender when they have selected a 'genderless' voice.

Somewhat similarly, questionnaires taken with members of LGBTQ communities found that non-binary people strongly believe that any voice can be non-binary [15]. Though, following the reasoning of Sutton [12] and the results from Mullennix et al. [14], it is highly likely that while any voice can be non-binary, they will not be initially perceived that way, and more likely be perceived as a binary gender, or as gender ambiguous, due to the predominant conceptualisation of gender being binary mentioned earlier. Additionally, as was also noted by the LGBTQ communities by Danielescu et al. [15], there is an inherent association between non-binary and androgyny, that is, the state of being neither specifically feminine nor masculine, or the combination of feminine and masculine characteristics [16]. In this sense, androgyny is very similar to gender ambiguity. And while non-binary includes much more than simply androgyny, it is a comparatively easy way of demonstrating non-binary identity in situations with very little context.

While the term 'gender ambiguous' will be mainly used throughout this paper to refer to anything that may not fit, or be perceived as a binary gender definition, other terms like genderless, gender neutral, or non-binary may still be used when referring to papers that used that specific terminology.

2.2 Anthropomorphism

This research largely takes place inside the space of social robotics. Where, despite growing research into the use of social robots in different contexts like healthcare, education, work environments, and at home [17], a universally agreed upon definition for social robots does not (yet) seem to exist [18]. Hegel et al. define a social robot as a robot plus a social interface, in which "a social interface is a metaphor which includes all social attributes by which an observer judges the robot as a social interaction partner" [19, pp. 174]. On the other hand, in their research about social robots, Leite et al. [17, pp. 291] use the functional definition "robots designed to socially interact with people or to evoke social responses from them." In early work by Breazeal [20], social robots are defined as those that people apply a social model to, in order to understand and interact with them.

Despite the lack of uniformity between these definitions, a clear returning theme within them is the presence of some form of human-robot interaction (HRI). In the same paper by Breazeal, four subclasses of social robots are identified: socially evocative, social interface, socially receptive, and sociable. Which all, to a different extent, encourage and partake in interactions with humans (HRI). When successively moving through the list, the subclasses describe robots that become increasingly capable of sustaining HRI in increasingly complex environments and scenarios, and contain increasingly more human characteristics. To the point where, eventually, both their ability to interact, and the reason why they interact become almost indistinguishable from humans. At this stage, these robots will have been fitted with many human-like characteristics to help create meaningful interactions and relationships with humans [20][21]. To understand why human-like characteristics embedded in technology have this effect, we must look at anthropomorphism.

2.2.1 Defining Anthropomorphism

Anthropomorphism is commonly defined as the internal attribution of human characteristics or traits to non-human agents [22], often emphasized within this definition, is the attribution of human mental capacities (the belief that a non-human agent is able to think and feel like humans can) [23]. A more simplified term that is frequently used instead of anthropomorphism is 'seeing human' [22][24][25], simply because we perceive an agent as more human when we attribute human characteristics to them. However, according to Frazer [26], the term anthropomorphism is also frequently used to refer to the design strategies/features used to encourage the perception of an agent as human-like (e.g. [27]). Frazer [26] calls these strategies 'operationalisations of anthropomorphism'. She found eight different operationalisation strategies used in research, of which physical appearance was most used (by far) followed by vocal features. These observable anthropomorphic design strategies and features are conceptualized to induce anthropomorphism of non-human agents, or in simpler terms, the presence of human-like design features is believed to lead to the internal attribution of human characteristics and traits to non-human agents [23][28].

2.2.2 Anthropomorphism & Stereotypes

In the context of social robots, anthropomorphism plays an important role. Research has found that anthropomorphized technology is generally favoured. It is more likely to be trusted [29], evaluated more favourably [30], and treated with greater care [31]. Robots with embedded human-like features are also assumed to allow for more intuitive interactions, as, following the Computers are Social Actors paradigm [32], it allows the human interactants to apply social models that are normally used for interactions with other humans [33][34]. These social models usually allow us to socially categorise people, to simplify our understanding of them and help us navigate social interactions efficiently [35]. However, these social categories also include stereotypes and biases, and as we apply them to social robots, so do we apply these stereotypes and biases [3][36][37]. Gender may be (one of) the most salient of these social categories; research shows it is significantly more likely to be ascribed to robots than other categories like race, religion, or age [25]. And when gender is ascribed, gender-specific stereotypes follow. The abundance of gender stereotypes can be divided into several components: traits, role behaviours, occupation, and physical appearance [38], several of which have been found to be applied to robots as well [39], possibly altering reactions towards, and perceptions of those robots (e.g. [40]–[47]). However, it should be noted that some studies do not find conforming results and instead suggest that the effect of gender stereotypes may be smaller than we anticipate [48][49].

Similarly to social robots, anthropomorphism also plays an important role in the context of computer voices and voice assistants. All voices, whether natural or synthetic, have vocal characteristics which contain certain features (mostly paralingual) that allow for rapid extraction of socially relevant information [50]. This is used to quickly form first impressions [51][52]. Following the Computers are Social Actors paradigm [32], similarly to the previous paragraph, this extraction of information to form an impression is not only applied to human voices but also to machine/computer voices [53][54]. Consequently, as first impressions are hypothesised to rely partly on over-generalisation [55][56], they also rely on stereotypes.

Resulting in those same stereotypes being applied to machine and computer voices by the user [54].

The implementation of gendered features, and thereby gender stereotypes, in technologies has been found to yield several positive results. It is seen as beneficial for user acceptance [57], as it ensures the design meets expectations and feels familiar to users [54][58]. Similarly, some argue that stereotypes are essential to efficiently process social information that enables us to interact with others [59]. Finally, the presence of stereotypes also suggests the potential for the accompanying stereotypic expectations to be violated [60], which could cause negative reactions [61], but is also seen as an opportunity to challenge these very stereotypes [37].

On the contrary, embedding stereotypes in technologies can also reinforce them [1]. An example of this is the (almost exclusive) use of feminine voices for the major voice assistants (Siri, Alexa, Cortana, Google Assistant). The commanding nature around the use of voice assistants reinforces the subservient assistant stereotype towards women, which leads to the reproduction of the commanding tone used for voice assistants, in interactions with real women [1][62]. One particular paper that provides significant insights into stereotypes in voice assistants is by Hwang et al. [58]. They specifically examined responses from several South Korean voice assistants to queries regarding relationships, personality, and appearance among others. They found the voice assistants typically described themselves as young women, with a tendency to value their bodily beauty (despite not having actual bodies). The assistants were intimate with the user, yet remained subordinate, even in the face of bad treatment and verbal abuse. They also embraced sexualization, even when the user's remarks could be construed as sexual harassment. The researchers concluded that the voice assistants represent victimized women who willingly embrace insults and sexual harassment, and noted the clear presence of a power dynamic between the user and the female voice assistant. Such behaviour from voice assistants with feminine voices can have a significant impact on the way actual humans with feminine voices are seen and treated.

Evidently, there is an inherent link between anthropomorphism and stereotypes, as anthropomorphism allows stereotypes to be transferred to non-human agents, and while this may lead to some beneficial outcomes, it could also reinforce potentially harmful stereotypes, and should thus be handled with care.

2.2.3 Anthropomorphism & Gender

Contrary to what was mentioned before, that increasing anthropomorphism seems to increase the ascription of (binary) gender [7], Martin & Mason [24][25] have recently suggested that gender ascription may not be just a consequence of anthropomorphism, but that it facilitates it. Firstly, they point out that features that enhance anthropomorphism are already entangled with gender, e.g. voices and faces [63]. Then, through a series of experiments, they support the assumption that gender is a defining feature of humanity [25], and show that ascribing gender increases the humanization of, and attachment to anthropomorphized technology [24]. More precisely, they found that male and female gendered voice assistants, autonomous vehicles, and robotic vacuums were ascribed more humanness than any of their genderless (non-gender-specific) counterparts [24]. Furthermore, they suggest that gender helps humanize 'targets' (anything that may be anthropomorphized/seen as human), and if these targets are described without gender, they are granted less humanity [25]. Important to note, the experiments conducted by Martin & Mason were mostly based on the participants' imagination, and very rarely included any physical products. For example, participants were asked to image aliens [25] and a voice assistant [24][25] with a certain gender, asked to form a first impression of an autonomous vehicle based on a short text [24], and asked to form a first impression of a 'person' based on two sentences [25]. Still, while it remains to be seen if the same results will be found when experiments are conducted with physical targets, their studies highlight the importance of gender as a social category, in anthropomorphism.

2.3 Gendering Social Robots

As described, the ascription of gender to a robot can have many consequences. This has led some to attempt to create 'gender neutral' robots, to bypass reinforcement of, or negative effects from, gender stereotypes (as previously mentioned in Chapter 1). Rather inconveniently, it is hypothesized and suggested by research findings that robots are perceived as male by default [7][37][64][65]. Similarly, it is suggested that gender neutrality is related to human-likeness, such that, the more human a robot looks, the more likely it is to be gendered (i.e. ascribed a binary gender; similar to what was described in Chapter 2.2) [7]. The ROBO-GAP database[1] [7], a database containing age and gender perception ratings of 251 anthropomorphic robots, which themselves were taken from the ABOT database[2] [66] which holds human-like appearance values for these robots, demonstrates what differently gendered robots may look like at many different levels of human-likeness.

The ROBO-GAP database [7] is an interesting database for those who want to research how the physical appearance of robots can affect their perceived gender, as well as their perception of traits or suitability for certain tasks. Due to the large number of robots in the database (251), it may indirectly serve as a conglomeration of much previous research on the effect of different visual stimuli on gender perception. Examples of such visual stimuli are hairstyles [3], chest-hip and waist-hip ratio [67]. Or more abstract stimuli, like curves and round shapes being more feminine, while sharp edges and straight lines are more masculine [68]–[70]. Also, aspects like texture [70], and colour (blue vs. pink) [71][72] can provide gender cues. The creators of the ROBO-GAP database themselves found that the presence of body manipulators (e.g. arms and legs) was exclusively associated with masculinity, while the presence of surface looks (e.g. eyelashes, head hair, or eyebrows) was exclusively associated with femininity [7].

Still, gender cues outside the physical realm are also frequently used. Most popular to manipulate, are the voice of the robot [40]–[48][73], the robot's name [42][43][46][48], or the pronouns used address the robot [37][49]. Research has even shown that when obvious gender cues are missing (e.g. physical appearance, voice, name, etc.), robots have been ascribed the gender associated with the task they were performing, i.e. the same robot would be described as feminine when cleaning, and masculine when playing computer games [74].

The amount and the diversity of these gender cues highlight the difficulty of creating the desired 'gender neutral' speaking social robot. While gender ambiguous robot embodiments have been successfully created, as evident in the ROBO-GAP database, ensuring the addition of voice, name, and pronouns does not affect the ambiguity of the robot, as well as ensuring the tasks it should perform and the context in which it operates, do not lead to binary gender ascription, seems to be both

[1] https://robo-gap.unisi.it/

[2] http://www.abotdatabase.info/

an arduous task, and limiting to the robot itself. To illustrate, one of the few studies that combined a gender ambiguous robot (Pepper[3]) with an ambiguous voice, found that only 30% of respondents classified it as ’neither male nor female,’ while 64% classified it as male [48].

2.4 Gendering Computer Voices

Gender is one of the most noticeable cues in voices [53], as it is one of the most dominant sources of variation [75]. Research shows, that any gender revealing cue, no matter how minor, can cause stereotypical responses [54]. As Lee et al. [53, pp. 290] mention, “the bizarre pronunciations and cadences of even the best TTS [*text-to-speech*] engines remain a constant reminder that a speaking computer is neither morphologically nor culturally gendered. Nonetheless, (...) individuals seem to automatically respond to the minimal infestations of gender in TTS as if they were interacting with a real person...” Note that this study was conducted in 2000, TTS engines have advanced significantly since then. Still, combined with the fact that any tiny suggestion of gender can cause stereotypical responses [54], it highlights the importance of considering voice design when it is implemented in any kind of computer system.

Possibly the most dominant gender cues in voices are the pitch and formant frequencies. The pitch in this case refers to the main underlying (fundamental) frequency of the voice, this is dependent on the vibrations of the vocal cords, whereas formant frequencies refer to the resonant frequencies of the vocal tract [76]. Formant frequencies can also be seen as band-pass filters created by the throat, mouth and nasal cavity, which means they emphasise some frequencies while suppressing others [77]. This allows us to hear different vowels, while also providing paralingual information. These (especially the pitch) are among the easiest gender cues to (digitally) manipulate, and as a result, much research regarding the gender perception of (computer) voices uses such manipulations.

While it is common knowledge that masculine voices on average have a lower pitch than feminine voices, the actual pitch range in terms of specific frequency values often differs between papers. Re et al. [78] note the pitch of male participants in their research ranged from 86-152 Hz, and the pitch of female participants ranged from 143-285 Hz. Biemans [79] noted that the pitch of her participants ranged from 84-157 Hz for men, and from 158-219 Hz for women. Even though differences may arise between the pitch ranges of participants of different studies, most studies do arrive at similar mean pitch values. Simpson [80] states the average pitch of English speakers is between 100-120 Hz for men, and between 200-220 Hz for women, which seems to correspond to the values presented by the previously mentioned studies. These frequency values are often used as targets for pitch shifting, when the aim is to change the gender of the voice. When developing an ambiguous voice, the aim is often to shift the pitch to a small overlap between the masculine and feminine pitch ranges. This ’neutral zone’ is claimed to be around 145 - 175 Hz, by the creator of ‘genderless’ voice ‘Q’ [4] [81], though again specific frequency values may differ between papers. Formant frequencies are much more difficult to relate to specific frequency values, as there are multiple levels of formants for every vowel in the phonetic alphabet. As a result, the manipulation of formant frequencies is mainly calculated through ratios, instead of specific frequency values.

[3] https://www.aldebaran.com/en/pepper

[4] https://www.genderlessvoice.com/

Similarly to what was discussed with social robots, several factors outside the vocal characteristics of the voice may also influence gender perception. Like the robots, this includes the name and the pronouns used to describe the voice, but it also includes the so-called 'perceived personality gender' [82]. This personality gender mainly follows gender stereotypes, such as the conversational style and word choice of the voice agent [12][15], the context of the task/role of the voice agent [12][83], and the physical location of the interaction [12]. While these factors may have no importance when the gender of the voice is clear through the pitch and formant frequencies, they can be very influential when the voice in question is aimed to be gender ambiguous. Accordingly, creating a true gender ambiguous computer voice remains very difficult, especially when all contextual factors have to be considered.

Interestingly, some researchers disapprove completely of the use of human-like, gendered computer voices. Instead, they advocate for the deliberate design and use of non-human-like voices [84]–[86]. They argue that the use of human-like voices creates expectations that the voice agent has human-like intelligence, leading to overestimations of the agent's real intelligence, a so-called 'habitability gap' [84][85]. As an aside, a similar argument has been made regarding the human-like embodiment of social robots as well [87]–[89]. Moore [84][85] argues for 'vocal appropriateness,' with which the true capabilities of the system are reflected in the voice (i.e. giving a robot a more robotic voice). Furthermore, Aylett, Cowan, and Clark [86] state there is an obsession with naturalness. They raise ethical concerns regarding the mimicry of human voices, and express that one reason for pursuing mimicry is ultimately to deceive the listener.

2.5 Auditory vs. Visual Stimuli

While much is known about factors that may influence gender perception of robots and voices (see Chapters 2.3 and 2.4), very little is known about their relative importance compared to each other. However, the relative importance of auditory and visual gender cues (also called cross-modal gender cues) has been researched more thoroughly in the fields of psychology and neuroscience. Here, one clear conclusion seems to recur in most studies, that is, dominance of visual information over auditory information, when these represent incongruent (mismatching) genders (this only includes binary genders) [8][90]–[92]. The stimuli in these cases were real human faces (either an image or a short video) overlapped with real human voices.

One hypothesis given for the dominance of visual stimuli is the information reliability hypothesis [8]. This hypothesis suggests that the dominant modality is the one that is more appropriate and efficient for the completion of the task at hand [93]. In the instance of Latinus, VanRullen & Taylor [8], they posit that the visual stimuli are dominant as human faces provide easily and immediately extractable information required for gender categorisation, whereas the auditory stimuli are dynamic and therefore need to be heard several times to allow for gender categorisation. From this, it can be inferred that the 'appropriateness' and 'efficiency' as mentioned in the information reliability hypothesis, relate to how clear, and how easily and quickly extractable the provided gender information is.

Interestingly, this hypothesis also describes what happens when one of the modalities is gender ambiguous, as shown by Smith, Grabowecky & Suzuki [94]. In their research, the application of tones with frequencies in the masculine and feminine frequency ranges (see Chapter 2.4) to androgynous faces, had a significant impact on the gender perception of the faces. Correspondingly, while the dominance of auditory stimuli may not match the results from the previously mentioned research, it does comply with the hypothesis, as ambiguous gender information is expected to be less easily extractable than binary gender information due to the predominant conceptualisation of gender as binary (see Chapter 2.1.2). This also corresponds with research conducted on robots and computer voices, where ambiguous stimuli (whether auditory or visual) have been shown to take the gender of the binary gendered stimulus applied to them, numerous times [12][37][40]–[49][73][95].

Somewhat surprisingly, in one study focusing on the perception of uncanniness in social robots by children, they seemed to rely more on the presented auditory cues than the visual cues, to judge the gender of the robot [96]. Following the given hypothesis, this would suggest that the auditory stimuli provided clearer gender information. Additionally, a study using animated audio-visual clips found that the auditory information was relied upon more than the visual information, to determine the emotional context, when its demonstrated emotion was incongruent with the visually displayed emotion [97].

Important to note about the research mentioned here, is the absence of measures regarding gender ambiguity. All barring one, only allowed participants to provide their gender perception by assigning either 'male' or 'female.' Only Paetzel et al. [96] also reported 'neutrality'.

2.6 Measuring Gender Perception

As mentioned by Sutton [12] (see Chapter 2.1.2), the method used to measure gender perception, can greatly impact the results. Still, no one method has been consistently used across different studies. Many of the studies mentioned in this research, across both voices and robots, have used different methods to measure gender perception. What follows here, is an overview and brief comparison of the multitude of methods that have been used in previous work.

2.6.1 Multiple-Choice

The researchers who use this method simply ask participants to tick a box regarding the perceived gender. Typically, there is a male option, a female option, and a third option which differs between studies. This third option has been called 'neutral' [98], 'other' [99], 'unsure' [83], and 'neither male nor female' [48]. When aiming for gender ambiguity, a majority of participants choosing the third option would generally be preferred, though an even split in male and female answers is also seen as a sign of ambiguity.

However, as an even split in male and female answers only represents gender ambiguity across a population, but not on an individual level, it could be argued that it does not illustrate true gender ambiguity. Additionally, as mentioned before (see Chapter 2.1.2), the third option in these cases is rarely selected [14]. Furthermore, Mullennix et al. [14] also state that gender perception is not categorical, thus this measurement method, which is inherently categorical, seems inappropriate for this context.

2.6.2 Single Scale

Possibly the most used method is a single scale. A scale with the opposite ends of the binary gender continuum as the extremes. Some researchers use 'male' and 'female' as the extremes [6][37][43][100], while others use 'masculine' and 'feminine' [49][74]. Generally, a 7-point scale is used, though a 5-point scale [6] and a 100-point scale [37] have also been reported. When using a single scale, the mid-point refers to gender ambiguity both when looking at single responses, as well as averages of many different responses.

Using a scale is preferred over multiple-choice, as it allows study participants to provide more nuance in their responses. Consequently, it has proven to be more sensitive for picking up gender ambiguity [14]. Interestingly, Danielescu et al. [15] made much different use of a scale. They asked participants to rate on a 5-point scale, if a voice sounded non-binary to them, and if they would be comfortable with that voice representing non-binary individuals. While it is unclear why they chose exactly these questions to measure perceived gender ambiguity, it should be noted that this study is the only one to explicitly seek responses from the LGBTQ community, possibly prompting the use of different questions.

2.6.3 Multiple Scales

In several studies with robots [41][46], and one with computer voices [53], researchers used two scales, one scale for masculinity and one scale for femininity. This allows for even more nuance when compared to responses from a single scale. These were not used for any kind of gender ambiguity measurement however, instead, they were mainly used for manipulation checks in studies that only involved binary genders.

In the recent creation of the ROBO-GAP database (see Section 2.3), a similar method was used [7]. Perugia et al. [7] refer to Bem's [101] gender schema theory, she showed that ratings for masculinity and femininity are independent of each other, and thus should be independently assessed, directly contradicting the use of a single scale for gender measurements. In accordance, participants were asked to convey their perception of masculinity, femininity, and gender neutrality in robots on separate 7-point scales, for the creation of the ROBO-GAP database. Following Bem's reasoning, equal endorsement of masculine and feminine attributes represents androgyny.

The method used by Perugia et al. for the creation of the ROBO-GAP database has already been found to allow for more nuance in gender perception measurements. When Roesler, Heuring & Onnasch [37] compared their results from a single gender continuum scale, to the results presented in the ROBO-GAP database, they found that the single scale may have suppressed ambiguities that were present in the database.

2.6.4 The Naming Technique

The last method, which is very dissimilar from the previous, asks study participants to name robots (as in, give robots a name) [37][102]. Participants are completely free in this process, which also results in the application of names that do not imply gender. For example, two studies that applied this technique reported most of the names to be functional, e.g. 'industrial helper,' or 'liftbot' [37][102]. In total, approximately 3/4 of all applied names were classified as nicknames or functional names, leaving only 1/4 as masculine or feminine names.

Despite the relatively low amount of gendered data gathered with this technique, Roesler et al. [37] do report that it revealed results that were not evident from the responses to a single scale gender continuum. Accordingly, they advocate for the use of more ‘subtle’ perceived gender measurement methods, like the naming technique, though given the fact that only 1/4 of the names were gendered, this should likely only be used in conjunction with a less subtle method mentioned before.

Chapter 3

Hypotheses

Based on the previous works discussed, hypotheses can be developed for the research questions at the centre of this study. To recap, this study aims to answer the following questions:

RQ1 What is the influence of *a)* voice and *b)* embodiment on the gender perception of speaking social robots?

With the sub-questions: *(a)* how does voice affect the gender perception of the embodiment of a social robot? And *(b)* how does embodiment affect the gender perception of a synthetic voice? And:

RQ2 What combination of gendered embodiment and gendered voice yields the most promising approach for the creation of ambiguous speaking social robots?

Combinations of binary gendered and gender ambiguous stimuli

H1a In any combination of a binary gendered stimulus and a gender ambiguous stimulus, the ambiguous stimulus is expected to have no effect on the gender scores of the robot-voice combination.

H1b Any combination of binary gendered and ambiguously gendered stimuli is expected to be perceived as the gender of the binary gendered stimulus.

These hypotheses follow from the information reliability hypothesis introduced in Chapter 2.5. This hypothesis states that the modality most appropriate and most efficient for the completion of the task (gender perception in this case) will be dominant [93]. This is expected to be the binary gendered modality, as the predominant conceptualisation of gender is binary [12]. Results in accordance with this hypothesis have already been found in research using images of androgynous faces as visual stimuli [94]. Additionally, previous research involving robots and computer voices has also shown that any (minor) suggestion of (binary) gender in technology may trigger stereotypic responses [54], and that ambiguous technology can be "drawn into" another gender category [12][74].

Combinations of masculine and feminine stimuli

H2a In any combination where the robot and the voice have incongruent genders, and neither is gender ambiguous (i.e. combinations consisting of a masculine and a feminine stimulus), the robot is expected to have a greater effect on the gender scores of the robot-voice combination than the voice.

H2b In any combination consisting of a masculine and a feminine stimulus, the robot is expected to have a greater effect on the gender scores of the voice than vice versa.

This follows from previous research in the fields of psychology and neuroscience, which consistently found dominance of visual stimuli over auditory stimuli with regards to the gender perception of faces and voices [8][90]–[92]. While some research involving less human-like stimuli has reported auditory dominance instead [96][97], only one study specifically addresses gender perception [96], and thus, these results are disregarded in favour of the larger body of literature reporting visual dominance.

The most ambiguous robot-voice combinations

H3 Robot-voice combinations where both the robot and the voice are gender ambiguous are expected to score higher on gender ambiguity measures than any other robot-voice combinations.

This assumption follows from the previous two hypotheses. As per **H1**, any combination of which only one of the stimuli is ambiguous will adopt the gender of the binary gendered stimulus, resulting in lower scores for gender ambiguity measures. Following **H2**, any combination of a masculine and a feminine stimulus will see a greater effect on the gender scores from the robot than the voice, leading to high scores for masculinity or femininity measures (depending on the gender of the robot) rather than gender ambiguity measures. On the other hand, any combination of two stimuli that have the same gender, is expected to stay that gender and score high in the related gender score (i.e. a combination of a feminine robot and voice is expected to score high on femininity). As such, robot-voice combinations made up of two stimuli that individually score high on gender ambiguity (i.e. two gender ambiguous stimuli) are expected to score higher on gender ambiguity measures than any other robot-voice combinations.

Chapter 4

Pre-test

A pre-test was conducted in the form of an online survey, to select stimuli for use in the main experiment. The main goal of the pre-test was to select six robots, with a balanced distribution of masculine, feminine, and ambiguous, and six computer voices, also with a balanced distribution of masculine, feminine, and ambiguous, that clearly present and are clearly perceived as their intended gender, and in the case of the robots, are controlled for their level of anthropomorphism.

4.1 Methods & Materials

4.1.1 Participants

A total of 20 participants took part in the pre-test. Incomplete responses were discarded. Eight women and twelve men completed the survey, most of which were Dutch, though two Germans and a Belgian also completed it. Most participants were 23 or 24 years old, while three participants were in or just before their sixties. Mean age was 29 years old (standard deviation of 13.4), median age was 24. Participants were found through convenience sampling, mainly at the Creative Technology Bachelor's and Interaction Technology Master's at the University of Twente. They had to be 16 years or older, and be able to properly see and hear the stimuli (determining what constitutes "properly see and hear" was at the participant's discretion).

4.1.2 Robots

A total of 18 robots (images) were used in the pre-test, with a balanced distribution of masculine, feminine, and ambiguous robots. These robots were selected through the ROBO-GAP[1] [7] and ABOT databases[2] [66], which hold masculinity, femininity and gender neutrality ratings, and human-likeness ratings respectively, of the same 251 robots. During this selection procedure prior to the pre-test, the goal was to select the six most masculine, feminine, and gender ambiguous robots, while controlling for the level of human-likeness as this could impact gender perception (see Chapter 2.2). These were then used during the pre-test. The selection procedure consisted of several steps.

Step 1: The data from the ROBO-GAP database was downloaded, this includes masculinity, femininity, and gender neutrality ratings of 251 robots on a scale from 1 to 7.

Step 2: For every robot, the absolute difference between the masculinity and femininity ratings was calculated (called AbsDiff from hereon). This step follows from Bem's original gender schema theory [101] that masculinity, femininity, and

[1] https://robo-gap.unisi.it/

[2] http://www.abotdatabase.info/

gender neutrality are independent of each other, meaning that a high masculinity rating does not ensure a low femininity rating. Accordingly, the difference between masculinity and femininity values (the calculated AbsDiff value) can give an insight into the (non-)ambiguity of a robot's gender.

Step 3: A list was then compiled of all robots with an AbsDiff smaller than or equal to 0.3. This list contained a total of 22 robots, which were then classified as ambiguous. As a side note, the neutrality scores were not used for the selection of ambiguous robots, as ambiguity merely suggests that the levels of masculinity and femininity should be equal or similar, which a high neutrality rating does not inherently indicate (see Step 2 and Chapter 2.1.2). Indeed, there are robots in the ROBO-GAP database with a high neutrality rating, that still present a large difference between masculinity and femininity ratings.

Step 4: Robots that were not included in the list of ambiguous robots, were separated into lists for masculine and feminine robots based on their masculinity and femininity ratings (this was already done in the original ROBO-GAP database). Here, the AbsDiff was added to their corresponding gender rating (i.e. for masculine robots, the AbsDiff was added to the masculinity rating), the result of this summation was called the Final Gender Score (FGS). This calculation was performed as a high FGS indicates that both the intended gender score had a high rating and the opposite gender score had a low rating (i.e. if a masculine robot had a very high FGS, it meant that both the masculinity rating was very high and the femininity rating was very low). The higher this score, the more clearly the robot displays its intended gender.

Step 5: The data from the ABOT database was downloaded, this includes human-likeness scores of the same 251 robots as the ROBO-GAP database, on a 100-point scale. This data was then added to the ambiguous, masculine, and feminine lists to allow for comparisons. Interestingly, the mean human-likeness score of the robots on the ambiguous list was 15, while the means for robots on the masculine and feminine lists were 44 and 47 respectively. The median score for the ambiguous list was only 10, while those for the masculine and feminine lists were 43 and 39 respectively. This seems to concur with previous research on anthropomorphism mentioned in Chapters 2.2.2 and 2.2.3, which implies that humanization is linked with the clear presence of gender.

Step 6: To control for the level of human-likeness in the selected robots, a human-likeness score range was defined. All selected robots must be from within this range. However, as became evident in the previous step, the human-likeness scores of the robots on the ambiguous list were very different from those on the masculine or feminine lists. The median human-likeness score of 10 for the ambiguous robots, was lower than the lowest human-likeness scores for the masculine and feminine robots, showing that the overlap in human-likeness scores between the ambiguous robots and the masculine and feminine robots was minimal. Therefore, to find a range of human-likeness scores that worked for all three gender categories, the six ambiguous robots (from the list established in Step 3) with the highest human-likeness scores were selected for the pre-test. The human-likeness scores of these robots were then used to define the range of human-likeness scores from which the masculine and feminine robots were selected. This resulted in a human-likeness score range between 18 and 44.

Step 7: The defined human-likeness score range was applied to the masculine and feminine lists, where every robot outside this range was disregarded. Within each respective list, the six robots with the highest FGS (calculated in step 4) were selected for the pre-test.

Step 8: A total of three (1 masculine, 2 feminine) out of the 18 selected robots, included a human face avatar on a screen. These robots were removed from the selection and replaced by those with the next highest FGS, as the inclusion of such features was not intended for this research due to the (presumed) impact on gender perception ratings. The mean human-likeness scores of the 18 selected robots were 30.9 for both the ambiguous and the feminine robots, and 35.2 for the masculine robots.

This selection procedure led to the selection of the following robots. Meka M1 Mobile Manipulator, Kibo, Topo, Moxi, 3e-a18, and Sanbot Max were the selected ambiguous robots (for images, see Appendix A). Mobiserv, Aryan, Murata Girl, Aila, Robina, and Sanbot were the selected feminine robots (for images, see Appendix B). Lego Mindstorms Nxt 2.0, Hiro, E3, Topio Dio, Rollin Justin, and Aero Drc were the selected masculine robots (for images, see Appendix C).

The images of the robots used in the pre-test were taken directly from the ROBOGAP and ABOT databases (these used the same images). These images have been standardized by the creators of the ABOT database, to show all robots in front of a light-coloured (or white) background. Additionally, at the initial creation of the ABOT database, whenever possible, the chosen images show the robot in a standing, neutral, forward-facing posture, with a neutral or slightly positive facial expression. [66]

4.1.3 Computer Voices

Unlike the robots, the computer voices had to be manually generated. Voice clips for 22 different voices were created for use in the pre-test, one clip per voice, guided by the methodology that was described and validated by Mooshammer & Etzrodt [100]. This methodology uses one computer voice as a root, and shifts the fundamental and formant frequencies to create other voices (male, female, and neutral).

Mooshammer & Etzrodt [100] used a voice from Google WaveNet, which is perceived as one of the most advanced and natural-sounding text-to-speech systems [103]. Specifically, they used the highest-pitched German male voice offered to record one voice clip, and pitch-shifted it to 156 Hz, which they determined to be the middle point between the average male and female voice pitch [100]. The resulting voice clip then served as the root, through which all other voice clips were created with fundamental and formant frequency shifts. In total, they created three male, three female, and nine neutral voices (one of which was the root). The exact fundamental and formant frequency values used can be found in Appendix D.

For this pre-test, two different root voices were used. The first root voice was a male voice from Google Studio (en-US-Studio-M), pitch-shifted to 156 Hz which should be ambiguous (or 'neutral') as stated in the previous paragraph. The aim was to select a voice from Google WaveNet, following the methodology of Mooshammer & Etzrodt [100]. However, all the male, American English voices from Google WaveNet were too high-pitched, as their pitch was already within the pitch range for the to-be-created neutral voices (see Appendix D). Therefore, other Google voice types were also considered, leading to the selection of a Google Studio voice with a lower pitch. Studio is considered a premium Google voice type just like WaveNet, and is specifically designed for use with long texts, like narration or news reading [104].

The second root voice was a voice called Sam, a 'non-binary' voice created by CereProc and Accenture Labs [15]. This voice was selected as it is one of the few commercially available gender ambiguous voices, making it both accessible for use in this study, and indicating that it may be used, or already be in use, for consumer goods. As this voice is already supposed to be gender ambiguous, it was only pitch and formant shifted to create three male and three female voices (again to the fundamental and formant frequency values provided in Appendix D).

Due to the methodology used for the creation of the voice clips, where one voice clip is recorded for the root voice and is then continuously pitch and formant shifted to create different sounding voice clips, all voice clips say the same text with the same accent as the root. To ensure this effect persisted even when using two different root voice clips, the selected Google Studio voice speaks American English as this is the only language in which Sam is available. Additionally, both root voice clips (and thereby all derived voice clips) say the same sentence.

The chosen sentence had to be carefully selected to avoid any gender association, as the goal for the voices was to be solely gendered based on their sound, not the content of their speech. This resulted in the selection of the grammatically correct but semantically nonsensical sentence “colourless green ideas sleep furiously” by Noam Chomsky [105]. In addition, the use of this single, short sentence limits time consumption during experiments compared to longer multi-sentence texts, while still providing enough information to form an impression [51].

Once the sentence was selected, recordings could be made for the root voices. The Sam root voice was recorded through CerePrompt, the text-to-speech engine of CereProc. The Google Studio voice was recorded through a Google webpage featuring a demo space [106], this voice was then pitch-shifted to 156 Hz to create the root voice. This pitch shift and all other pitch and formant shifts needed to create the voice clips, as well as early frequency analyses to find a suitable Google WaveNet or Google Studio voice, were performed through the Praat program, created by Paul Boersma and David Weenink[3] [107].

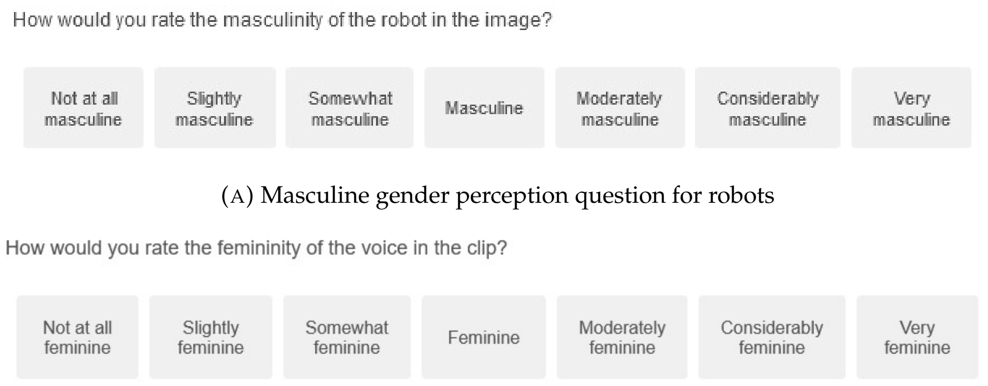
A total of 22 voice clips were created. Fifteen from the Google Studio root; 9 neutral (including the root), 3 male, and 3 female. Seven from the Sam root; 1 neutral (the root), 3 male, and 3 female. After completion, one clear difference was noted between the voices from the different roots, despite them speaking with the same accent and saying the same sentence. Namely, the monotony of the Sam-derived voices compared to the Google Studio-derived voices.

4.1.4 Measures

Gender perception questions for all the voices and robots made up the main body of the survey. Each stimulus was accompanied by three gender perception questions: “how would you rate the [*masculinity* / *femininity* / *gender neutrality*] of the [*voice in the clip* / *robot in the image*]?” With a text-based 7-point scale to answer each question. The scale points were (in order): not at all masculine, slightly masculine, somewhat masculine, masculine, moderately masculine, considerably masculine, and very masculine (see Figure 4.1a). In this 7-point scale, ‘masculine’ was replaced by ‘feminine’ or ‘gender neutral’ depending on the question the scale related to, as seen in Figure 4.1b (see Appendix E and F for full excerpts of the survey). This method for measuring gender perception, using multiple scales, was replicated from the original study creating the ROBO-GAP database [7] (see Chapter 2.6.3), though changes were made to the wording of the questions and scales for increased clarity. For reference, the original wording of the question was: “how would you describe the robot in the image?” Accompanied by three 7-point scales (called *feminine*,

[3] www.praat.org

masculine, and *gender neutral*) ranging from (1) completely disagree to (7) completely agree [7].



How would you rate the masculinity of the robot in the image?

(A) Masculine gender perception question for robots

How would you rate the femininity of the voice in the clip?

(B) Feminine gender perception question for voices

FIGURE 4.1: Gender perception questions for robots and voices

After the gender perception questions, participants were presented with 13 statements regarding social roles, and asked to convey to what extent they agreed with these statements on a scale from "*0% — strongly disagree*" to "*100% — strongly agree*." This was a direct copy of the Social Roles Questionnaire (SRQ) as proposed by Baber & Tucker [108]. They describe it as a method to capture how people think about gender. Correspondingly, it was included in the research as gender attitudes may have an impact on the gender perception results. For the same reason, the participants' gender (using the gender-sensitive method described by Spiel et al. [109]), age, and nationality were also collected. While these were not considered during the eventual selection of robots and voices for the main experiment, they may provide valuable insights nonetheless.

4.1.5 Procedure

The survey was conducted online using Qualtrics. The link to the survey brought participants to a web page with an information brochure and a consent form, where participants were fully informed on the contents of the survey, as well as the goal of the experiment and the research project as a whole. Once consent was established, they could start. The survey was split into several stages. First, all (22) voices were presented in randomized order, one voice at a time. The participants could replay all voice clips as much as they desired, and were asked to provide their perception of the gender of each voice (see Appendix E). Second, after a short intermission, the participants were presented with all (18) robots, also in randomized order, one robot at a time, and asked to provide their perception of the gender of each robot (see Appendix F). In the final stage, participants were asked to fill in the SRQ, as described in the previous section, and provide demographic data. Afterwards, the participants were forwarded to a web page with a small debrief and a final consent check after which the survey was finished.

In total, the survey was expected to take around 15–20 minutes to complete, though no time limit was implemented. The mean of the eventual measured time to completion was 44 minutes, with a median of 24 minutes. It was conducted completely remotely, participants could take part when and where it was convenient for

TABLE 4.1: An overview of the robots and their placeholder names. *AR* stands for 'ambiguous robot,' *MR* stands for 'masculine robot,' and *FR* stands for 'feminine robot.'

3e-a18	= AR_1	Aero Drc	= MR_1	Aila	= FR_1
Kibo	= AR_2	E3	= MR_2	Aryan	= FR_2
Meka M1	= AR_3	Hiro	= MR_3	Mobiserv	= FR_3
Moxi	= AR_4	Lego Mindstorms	= MR_4	Murata Girl	= FR_4
Sanbot Max	= AR_5	Rollin Justin	= MR_5	Robina	= FR_5
Topo	= AR_6	Topio Dio	= MR_6	Sanbot	= FR_6

them. They were only asked to ensure they could properly see and hear all the stimuli. Participants were not able to go back and change their answers to any previously answered question.

4.1.6 Analysis

Analysis of the results used similar calculations as the initial robot selection procedure. First, mean masculinity, femininity, and neutrality scores were calculated for each robot and voice. Then, to select the most gender ambiguous robots and voices, the AbsDiff of each robot and voice was calculated (see Step 2 of Chapter 4.1.2). For the selection of the most masculine and feminine robots and voices, the Final Gender Scores (FGS) were calculated (see Step 4 of Chapter 4.1.2).

4.2 Results

4.2.1 Robots

To increase clarity, the robot names are replaced by placeholders. These placeholders provide immediate information on the intended gender of the robot; the gender for which the robot was initially selected (see Chapter 4.1.2). See Table 4.1 for an overview of the robots and their placeholder names (for all robot images, see Appendix A, B, and C).

As described in the methodology, the mean masculinity (masc), femininity (fem), and gender neutrality (neut) scores were calculated for every robot, as well as the AbsDiff and FGSs originally established in Chapter 4.1.2. Figure 4.2 gives an overview of the masculine and feminine FGSs, as well as the AbsDiffs of all the robots. All exact results can be found in Appendix G.

None of the robots deviated from the gender category for which they were selected. Of the six ambiguous robots, only AR_2 had an AbsDiff greater than 1, at 1.05, while four of the six ambiguous robots had an AbsDiff smaller than 0.5. Of the non-ambiguous robots, the lowest AbsDiff was 1.9 for FR_2, while the next lowest was 2.55 for MR_1. The mean neutrality scores of the ambiguous robots were higher than those of all other robots except one, as the neutrality score of MR_3 was greater than that of AR_5. The highest masculine FGS was 10.4 for MR_6, while the next highest was 7.8 for MR_4. Four out of six masculine robots had a masculine FGS above 7, with MR_3 and MR_1 slightly lower. The mean masculinity scores of all masculine robots were higher than the masculinity scores of all other robots. FR_1 had the highest feminine FGS at 10.8, with the next highest being 9.65 and 9.55 for FR_6 and FR_5 respectively. Of the feminine robots, FR_2 had the lowest feminine

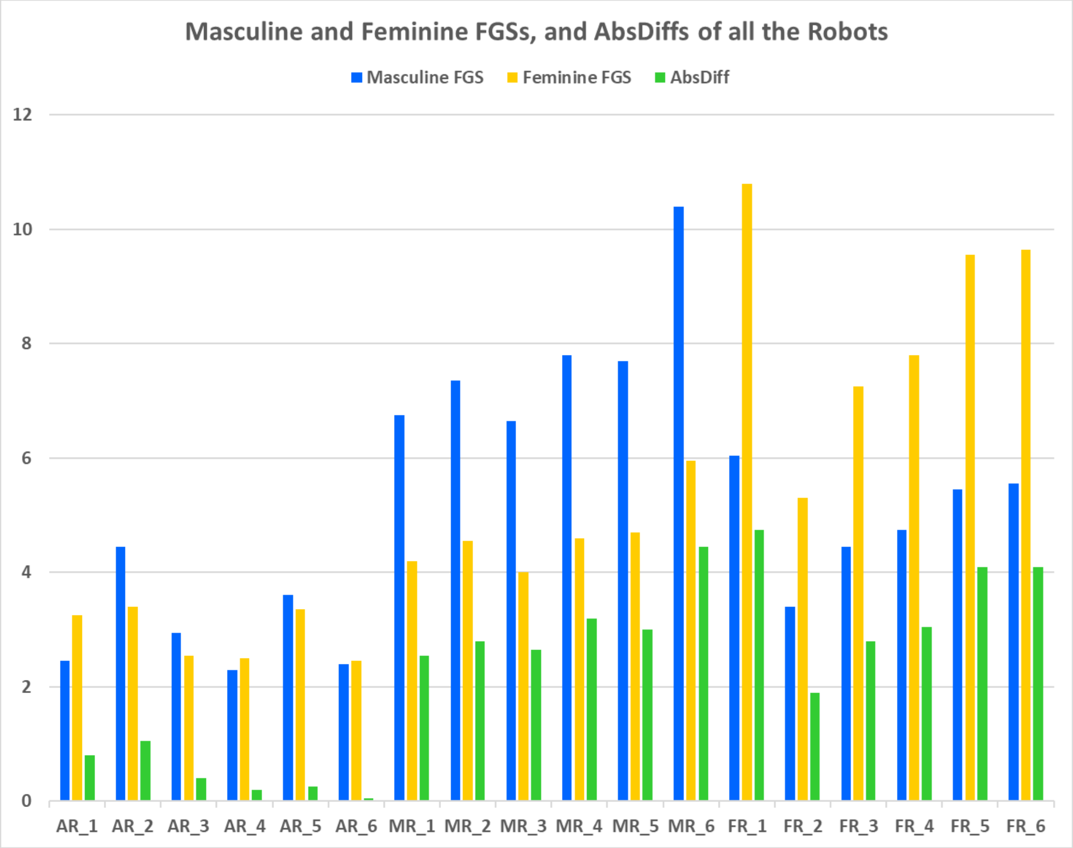


FIGURE 4.2: The masculine and feminine FGSs, and AbsDiffs of all the robots

FGS at 5.3, almost 2 lower than the next lowest, FR_3. Similar to the mean masculinity score of the masculine robots, the mean femininity scores of the feminine robots were higher than those of all other robots. FR_1 is the only robot with a gender score higher than 6, with a mean femininity score of 6.05.

4.2.2 Voices

An overview of the masculine and feminine FGSs, and the AbsDiffs of all the voices can be found in Figure 4.3. The exact results can be found in Appendix H. Naming conventions of the voices are copied from the table in Appendix D. The *F* voices were intended to be feminine, the *M* voices were intended to be masculine, and the *N* voices were intended to be neutral/ambiguous. The voices starting with 'sam' were created from the Sam root voice (sam_n1; which should be ambiguous), all others were based on the Google Studio root voice.

None of the voices had an AbsDiff lower than 1. The lowest AbsDiff was 1.05 for F1, followed by 1.65 for N8 and N9. All remaining voices had an AbsDiff of 2 or higher, including sam_n1 with an AbsDiff of 2.5. The highest masculine FGS was 11.15 for M2, followed closely by M1 at 10.85. The *M* voices from the Sam base scored much lower, between 8.85 and 6.45, which is on par with, or lower than, many of the *N* voices created from the Google Studio root which were meant to be ambiguous. The opposite was the case for the feminine FGSs. The highest feminine FGS was 9.9 for sam_f1, followed by sam_f2 at 9.85, and sam_f3 at 9.05. The *F* voices based on the Google Studio root voice had much lower scores ranging from 7.35 to 4.85. Sam_n1,

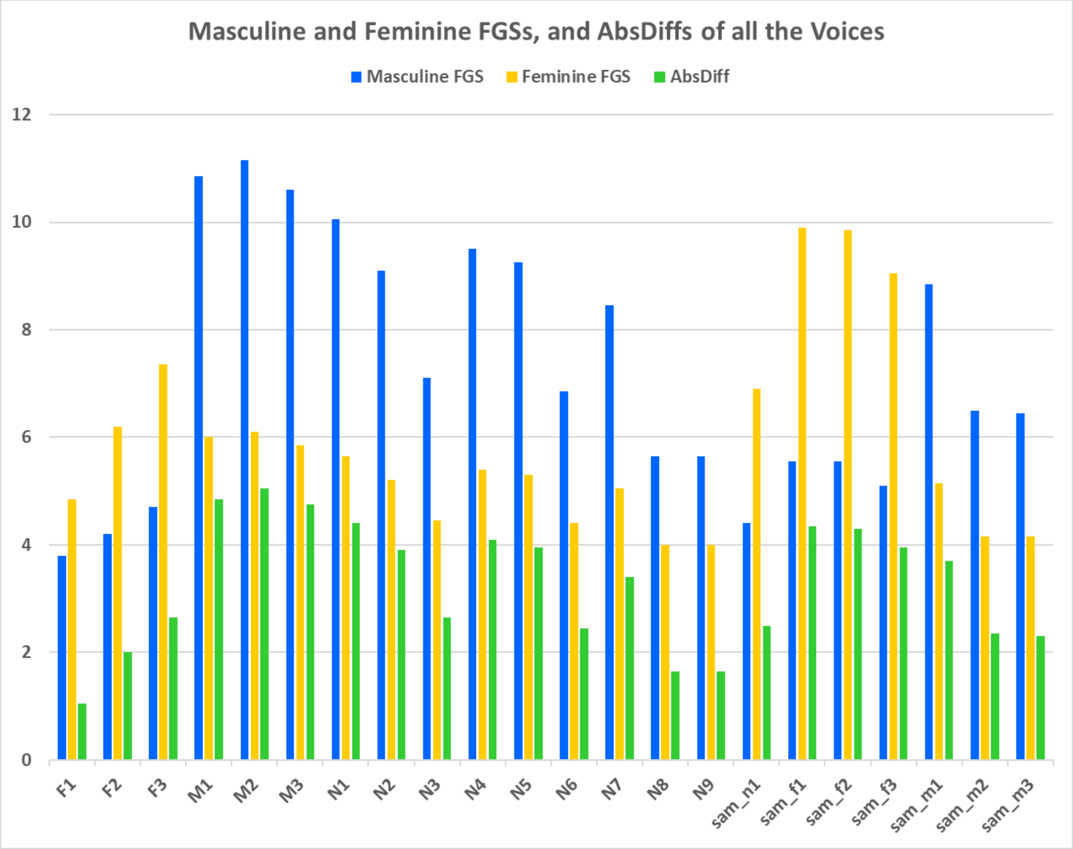


FIGURE 4.3: The masculine and feminine FGSs, and AbsDiffs of all the voices

the original Sam voice created by CereProc and Accenture Labs to be 'non-binary,' had a feminine FGS of 6.9, higher than two of the three *F* voices created from the Google Studio root.

4.3 Discussion

4.3.1 Robots

None of the robots diverged from their perceived gender based on the data in the ROBO-GAP database, though some differences are present. For example (as shown in Table 4.2), the feminine FGSs of the feminine robots and the masculine FGSs of the masculine robots based on the results of this pre-test, were considerably lower than those based on the data from the ROBO-GAP database. Additionally, more so for the masculine robots than the feminine robots, the order of the robots from highest to lowest FGS differs between the pre-test results and the ROBO-GAP database. Most prominently, of the six masculine robots that were included in the pre-test, MR_4 had the lowest masculine FGS based on the ROBO-GAP database, but ranked second highest based on the pre-test results. Though with only a 0.1 difference in FGS between MR_4 and MR_5, as well as between FR_5 and FR_6, it is unlikely there is any statistically significant difference between them.

Similar to the pre-selection of the robots from the ROBO-GAP database, the robots with the highest masculine or feminine FGS are selected for their respective gender categories (see Step 7 in Chapter 4.1.2). As the goal of this pre-test is to select two robots per gender category for the main experiment, the selected masculine robots are MR_6 and MR_4, while the selected feminine robots are FR_1 and FR_6.

TABLE 4.2: Comparison between pre-test results and ROBO-GAP database for masculine and feminine robots

(A) Masculine FGS of Masculine Robots

Robot	Pre-test	ROBO-GAP
MR_6	10.4	11.323
MR_4	7.8	9.223
MR_5	7.7	9.8
MR_2	7.35	9.581
MR_1	6.75	9.742
MR_3	6.65	9.387

(B) Feminine FGS of Feminine Robots

Robot	Pre-test	ROBO-GAP
FR_1	10.8	11.903
FR_6	9.65	10.233
FR_5	9.55	11.129
FR_4	7.8	9.484
FR_3	7.25	9
FR_2	5.3	8.233

Where the masculine and feminine FGSs were all considerably lower in the pretest compared to the ROBO-GAP database, such an effect was not observed in the AbsDiff nor the neutrality scores of the ambiguous robots (see Table 4.3). A comparison between AbsDiff and neutrality scores (as neutrality scores have been more traditionally used to measure non-binary gender) shows that two robots are in both the top three lowest AbsDiff scores, and the top three highest neutrality scores (remember that for AbsDiff a low score signifies ambiguity, while for neutrality score a high score signifies ambiguity). These are AR_6 and AR_4. The two robots that occur in only one of these top threes, AR_5 for lowest AbsDiff (3rd) and AR_1 for highest neutrality score (2nd), score surprisingly badly in the other category. AR_5 has by far the lowest neutrality score of the six ambiguous robots, with a gap of 1.05 to the next lowest score. While AR_1 has the second highest AbsDiff, only surpassed by AR_2.

The selection of ambiguous robots for the main experiment is based on the AbsDiff, similar to the pre-selection from the ROBO-GAP database (see Step 3 in Chapter 4.1.2). As such, the two robots with the lowest AbsDiffs, AR_6 and AR_4, are selected.

TABLE 4.3: Comparison between pre-test results and ROBO-GAP database for ambiguous robots

(A) AbsDiff of Ambiguous Robots

Robot	Pre-test	ROBO-GAP
AR_6	0.05	0
AR_4	0.2	0.167
AR_5	0.25	0.3
AR_3	0.4	0.167
AR_1	0.8	0.167
AR_2	1.05	0.3

(B) Neutrality Score of Ambiguous Robots

Robot	Pre-test	ROBO-GAP
AR_4	5.1	4.933
AR_1	5	5.167
AR_6	4.95	5.333
AR_3	4.4	4.8
AR_2	4.15	3.5
AR_5	3.35	3.567

4.3.2 Voices

The voices diverged considerably from their intended gender, as well as showing considerable differences between the two root voices (as alluded to in the results; see Chapter 4.2.2). As shown in Figure 4.3, most of the voices based on the Google Studio voice scored high on masculinity, including many *N* voices which were meant to be ambiguous. Suggesting that the masculinity of the original recording, from which the root was created, may still be discernible. Several of these *N* voices even had higher masculine FGSs than masculine intended voices based on the Sam voice, which had relatively low masculine FGSs for voices that are meant to be masculine. Conversely, the highest feminine FGSs were all for voices based on the Sam voice, while those of the feminine intended Google Studio-based voices were much lower. Two of the three feminine intended voices from the Google Studio root even had a lower feminine FGS than the Sam base voice (sam_n1), which is meant to be ambiguous. The results highlight a difference between the seemingly more feminine Sam root voice and the seemingly more masculine Google Studio root voice.

As the voices with the highest masculine or feminine FGS are selected for their respective gender categories for the main experiment, similarly to the robots, M1 and M2 are selected as masculine voices, while sam_f1 and sam_f2 are selected as feminine voices.

The relatively high feminine FGS of the Sam root voice also highlights the difficulty of creating an ambiguous voice. While Sam was not explicitly made to be ambiguous, but rather non-binary to be specific [15], given its intricate and advanced creation method relative to the ambiguous intended voices of the Google Studio root used in this research, as well as the producers being big corporations with experience in creating computer-generated voices, it is surprising that seven other voices have a lower AbsDiff, and two other voices have higher neutrality scores, including masculine intended voices based on the original Sam voice (see Figure 4.3 and Appendix H).

Just as with the robots, the ambiguous voices are selected based on their AbsDiff. The first selected voice is F1, as it has the lowest AbsDiff. However, the second lowest AbsDiff belongs to both N8 and N9, at 1.65. To select a second ambiguous robot from these two, their neutrality scores are considered. As N8 has a higher neutrality score than N9, it is selected for the main experiment.

TABLE 4.4: The top 5 lowest AbsDiff and highest neutrality scores of the voices

(A) Top 5 lowest AbsDiff scores of the voices

Voice	AbsDiff
F1	1.05
N8 & N9	1.65
F2	2
sam_m3	2.3
sam_m2	2.35

(B) Top 5 highest neutrality scores of the voices

Voice	Neutrality Score
F2	3.7
sam_m3	3.25
sam_n1	3.1
N8	3.05
F1 & N9	3

A comparison to the results from Mooshammer & Etzrodt [100], whose methodology was followed as exact as possible for the creation of the voices from the Google Studio root (see Chapter 4.1.3), suggests that they were more successful in creating feminine perceived voices. Additionally, their ambiguous intended voices seemed to lean more towards femininity, while the opposite was true here. This difference may be attributed to the use of a different voice for the creation of the root, though it should also be noted that Mooshammer & Etzrodt used a different method to measure gender perception (a single 7-point scale), hindering a true comparison of the results.

4.4 Conclusion

In summary, the following robots were selected for use in the main experiment: MR_6 (masc), MR_4 (masc), FR_1 (fem), FR_6 (fem), AR_6 (amb), and AR_4 (amb). The following voices have also been selected: M1 (masc), M2 (masc), sam_f1 (fem), sam_f2 (fem), F1 (amb), and N8 (amb). It may be argued that the selected robots and voices cannot be concluded to be the two most masculine, feminine, or ambiguous due to small differences between selected and non-selected robots and voices. The goal of this pre-test has still been reached, as the selected robots and voices are found to clearly present and be clearly perceived as the measured gender category.

Chapter 5

Main Experiment

Following the main aim of this research project presented in the introduction of this report (Chapter 1), that is, to find the influence embodiment and voice have on the gender perception of speaking social robots and each other, and to find the most promising approach for the creation of gender ambiguous speaking social robots, a survey is conducted in which the perceived gender of robots and voices is tested both in isolation and in robot-voice combinations.

5.1 Methods & Materials

5.1.1 Participants

In total 38 participants took part in the main experiment, of which 12 were men, 22 were women, one was non-binary, and three preferred not to say their gender. They had a mean age of 38 (standard deviation of 17), and a median age of 31. The most represented nationality was Dutch, with 13 participants coming from the Netherlands, while four participants were British, and three were German. The rest came from a variety of countries throughout Europe, South(-East) Asia, and North America. Participants were partially recruited through communal chat groups related to the Creative Technology Bachelor's and Interaction Technology Master's at the University of Twente. During recruitment, it was stressed to potential participants to not respond if they had already participated in the pre-test. The remaining participants were recruited through research platform SurveyCircle.com, where researchers can share their online surveys and find participants. Equal to the pre-test, participants had to be 16 years or older, and be able to properly see and hear the stimuli (determining what constitutes "properly see and hear" was at the participant's discretion). Additionally, as mentioned, participants of the pre-test were excluded.

5.1.2 Robots & Voices

The six selected robots and voices were presented to the participants in the same manner (with the same images and voice clips) as they were in the pre-test. However, a small addition was made to the robot images, namely an indication of the height of the robot as the original images might not be able to fully communicate this (see Appendix I for an example of the height indication).

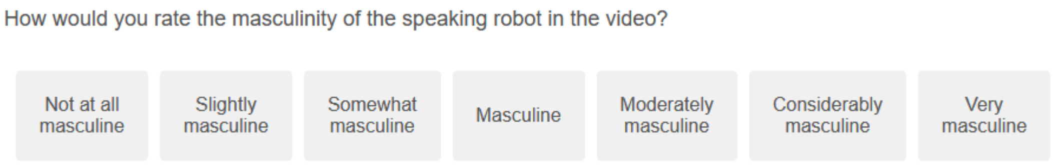
To recap, the selected robots are MR_4 (masc), MR_6 (masc), FR_1 (fem), FR_6 (fem), AR_4 (amb), and AR_6 (amb). The selected voices are M1 (masc), M2 (masc), sam_f1 (fem), sam_f2 (fem), F1 (amb), and N8 (amb).

5.1.3 Robot-Voice Combinations

As this research project specifically concerns speaking social robots, videos had to be made showing the selected robots speaking with the selected voices. To achieve this, a strategy used previously in research concerning the relative importance of auditory and visual stimuli, in the academic fields of psychology and neuroscience, discussed in Chapter 2.5, was replicated. In this strategy, a static image of the visual stimulus is presented together with a sound clip of the auditory stimulus [8][94][110]–[112]. To this end, short videos were created presenting the selected robots and voices simultaneously, using the same still images and voice clips as in the pre-test (the robot images used for the videos did contain the height indication added after the pre-test). All possible robot-voice combinations were created, resulting in a total of 36 (6 x 6) combinations. The duration of the incorporated voice clip determined the duration of the video, resulting in videos with durations of around three seconds.

5.1.4 Measures

The measures used in the survey were the same as those in the pre-test, described in Chapter 4.1.4. All robots, voices, and robot-voice combinations were accompanied by the same three gender perception questions as in the pre-test, regarding masculinity, femininity, and gender neutrality. Figure 5.1 provides an example regarding the masculinity of a robot-voice combination. The questions regarding femininity and gender neutrality were the same, apart from 'feminine' and 'gender neutral' replacing the word 'masculine' in the question (full survey excerpts for questions regarding voices, robots, and robot-voice combinations can be found in Appendix E, F, and J respectively). Participants were, again, also asked to fill in the Social Roles Questionnaire (SRQ) [108], as well as demographic questions regarding gender, age, and nationality. During analysis, no correlation was found between participants' answers to the gender perception questions and their gender, age, nationality, or SRQ answers.



How would you rate the masculinity of the speaking robot in the video?

FIGURE 5.1: Masculine gender perception question for robot-voice combinations

5.1.5 Procedure

Like the pre-test, this survey was conducted online using Qualtrics. When opening the link to the survey, participants first saw an information brochure and a consent form, where they were fully briefed on the contents of the survey and the goal of the experiment. Once participants had consented, they were ready to begin. The survey was split into multiple blocks. First, the participants were asked to give their gender perception of the six voices. Just like in the pre-test, the voice clips were presented in randomized order, and participants could listen to the voice clips as many times as they wanted. Second, they were asked to give their gender perception of the six robots. These were also presented in randomized order. The third stage made up the main body of the survey. This included all the (36) possible robot-voice combinations. Again, these were presented in randomized order. At the end of each block, participants got the opportunity to provide remarks on the foregone stage. In the final stage, participants were asked to fill in the SRQ and provide demographic data, after which they received a small debrief and a final consent check to finalise the survey.

Another small block was presented between the robot-voice combinations and the SRQ. This block specifically focused on one robot called Harmony (which was not in the ROBO-GAP and ABOT databases), as part of a different, separate research project. Within this block, the robot was presented individually and in combination with the voices, in the same way as the rest of the survey, with the same gender perception questions as well as other questions not relevant to this research. The results from this block are not analysed here. Further information about the robot, and the survey block dedicated to it, can be found in Appendix K.

The survey was expected to take around 30–35 minutes to complete, though no time limit was implemented, like the pre-test. The mean measured time to completion was 69 minutes, with a median of 27 minutes (one participant recorded almost 23 hours between starting and finishing the survey, exclusion of this participant would have given a mean duration of 34 minutes). Also, similar to the pre-test, the survey was conducted completely remotely, and participants were not able to go back and change their answers to any previously answered questions.

5.1.6 Analysis

First, to create an overview of the results, gender perception scores for all robots, voices, and robot-voice combinations were calculated, similar to the pre-test (see Chapter 4.1.6), based upon which they were subsequently categorised by gender. These gender categorisations were used to address **RQ2**, concerning which robot-voice combinations are the most promising for the creation of gender ambiguous speaking social robots. Incidentally, they also allowed for a first impression of the possible effects voice and embodiment have on the gender perception of speaking social robots, related to **RQ1**. Though, the main analysis to address **RQ1** was performed through the creation of linear mixed models, to analyse the exact effects the robots and voices had on each other's gender perception scores.

Additional preprocessing steps were taken to prepare the data for use in the linear mixed models, including the creation of extra variables regarding the change to the gender scores of robots and voices once they were combined to make robot-voice combinations. To find exactly how the perceived gender of a robot changed when a specific voice was added to it, the difference between the gender ratings of the individual robot and the robot-voice combination was calculated for all respondents individually. In this case, the robot was referred to as the *base*, while the voice was referred to as the *addition*. This was done for all three gender ratings, as well as the AbsDiff, of every robot-voice combination. The same was also done the other way around, where the difference was calculated between the gender ratings of an individual voice and the robot-voice combination, to find how the perceived gender of a voice changed when a specific robot was added to it. In this case, the voice was referred to as the *base*, while the robot was referred to as the *addition*.

These differences (or *deltas* as they were called) were then used as dependent variables to create the linear mixed models. The independent variables were: the type of the base (robot or voice), the gender of the base (masculine, feminine, or ambiguous), the type of the addition (robot or voice), and the gender of the addition (masculine, feminine, or ambiguous). The variable for the base type was discarded due to redundancy, as the type of the addition already provides this information (if the addition is a voice, the base must be a robot, and vice versa). To avoid possible three-way interactions, and thereby limit the complexity of the models, different models were created for every base gender (masc, fem, and amb), leaving only the type of the addition (*addtype*; robot or voice) and the gender of the addition (*addgen*; masculine, feminine, or ambiguous) as independent variables. The random effects incorporated in the model were the individual participants, the specific robot or voice that was the base (base_id), and the specific robot or voice that was the addition (add_id).

A total of 12 models were created (3 different base genders x 4 different gender scores). The model equations looked like the following (remember that the data used in the models differed between the different base genders):

$$\{GenderScore\}\_delta = \beta_0 + \beta_1 addtype * \beta_2 addgen + b_{participant} + b_{base\_id} + b_{add\_id} + \epsilon$$

Where:

$\{GenderScore\}\_delta$ = the difference between the gender score of the base and the robot-voice combination. Where {*GenderScore*} can be the masculinity score, femininity score, neutrality score, or the AbsDiff.

β_0 = the fixed intercept

$\beta_1 addtype$ = the fixed effect associated with the type of the addition

$\beta_2 addgen$ = the fixed effect associated with the gender of the addition

$b_{participant}$ = the random effect associated with the individual survey participants

b_{base_id} = the random effect associated with the identity of the base (the specific base used)

b_{add_id} = the random effect associated with the identity of the addition (the specific addition used)

ϵ = the residual error term

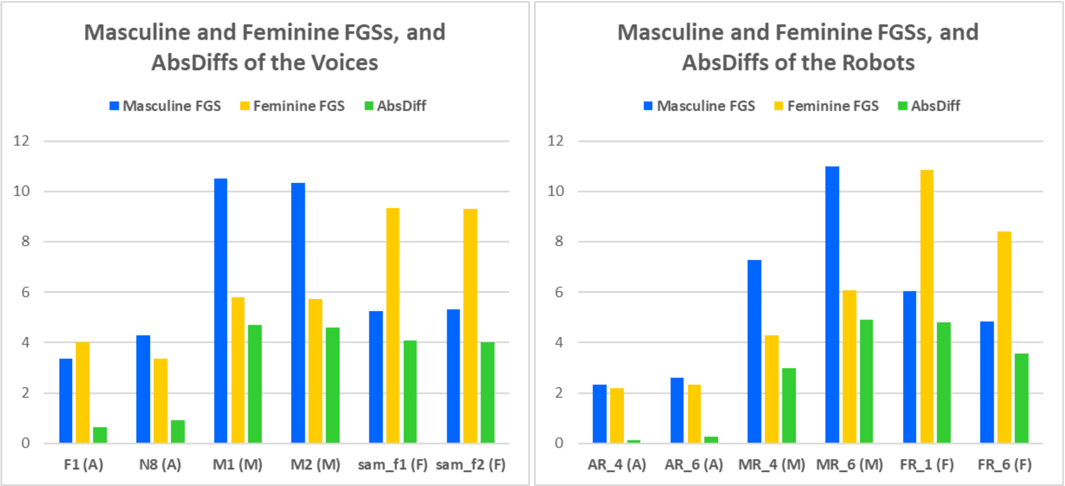
As both independent variables are categorical, dummy variables had to be created. This was done automatically by *R* [113], the software used for this analysis. When using dummy variables, one of the variable values is omitted and used as reference, to be incorporated in the fixed intercept of the model. In all models, the reference value of *addtype* was Robot. The reference value of *addgen* differs per base gender. In the masculine base gender models, the reference value of *addgen* is masculine. In the feminine base gender models, the reference value is feminine, and in the ambiguous base gender models, the reference value is ambiguous.

The linear mixed models were created using the *lme4* [114] and *lmerTest* [115] packages. The *r2_nakagawa* function from the *performance* package [116] was used to calculate the R^2 (model fit) of each model. Interaction plots were created with the *cat_plot* function from the *interactions* package [117], in combination with the *ggplot2* package [118].

5.2 Results

5.2.1 General Overview

The gender perception results of the robots and voices in isolation (see Figure 5.2), show that all were perceived as intended following the results and selection from the



(A) The masculine and feminine FGSs, and AbsDiffs of the voices

(B) The masculine and feminine FGSs, and AbsDiffs of the robots

FIGURE 5.2: The gender perception results of the voices (A) and robots (B) when presented in isolation. The letters after the robot and voice names signify their intended gender (Ambiguous/Masculine/Feminine)

pre-test; all gender scores were similar to those found in the pre-test (a full overview of the isolated gender perception results of the robots and voices can be found in Appendix L).

The gender perception results of the robot-voice combinations can be found in Appendix M. Figures 5.3, 5.4, 5.5, and 5.6 provide an overview of the effect gendered robot or voice additions had on the gender scores of the bases. The graphs give a first idea of what effects may be found through the linear mixed models (described in Chapter 5.1.6), which are presented in Chapters 5.2.3, 5.2.4, and 5.2.5. For instance, the change of the gender scores seems to be slightly greater for robot bases than for voice bases. Additionally, little to no change in the gender scores seems to occur when the base and the addition have the same gender. These results are explored further in the following sections, through statistical analyses of the linear mixed models.

Throughout the survey, the participants were also asked to provide any remarks they had regarding their answers to the gender perception questions. Some remarks stood out, as they were given multiple times. Several participants indicated feeling confused, finding it weird, and struggling to rate the gender scores of combinations with a clear mismatch between voice and robot gender. One participant noted, "[a] mismatch between voice and appearance left me confused." Multiple participants stated they based their gender perception answers more on the voice than on the robot, while others noted basing it on both.

5.2.2 Gender Categorisation

Gender categorisation of the robot-voice combinations was performed based on their AbsDiff, and masculine and feminine FGS (see Figure 5.1; the results upon which this categorisation is based can be found in Appendix M). Here, all combinations with an AbsDiff (absolute difference between the masculinity and femininity score) smaller than 1 are classified as ambiguous. All remaining combinations are classified as either masculine or feminine based on which FGS (Final Gender Score) is higher.

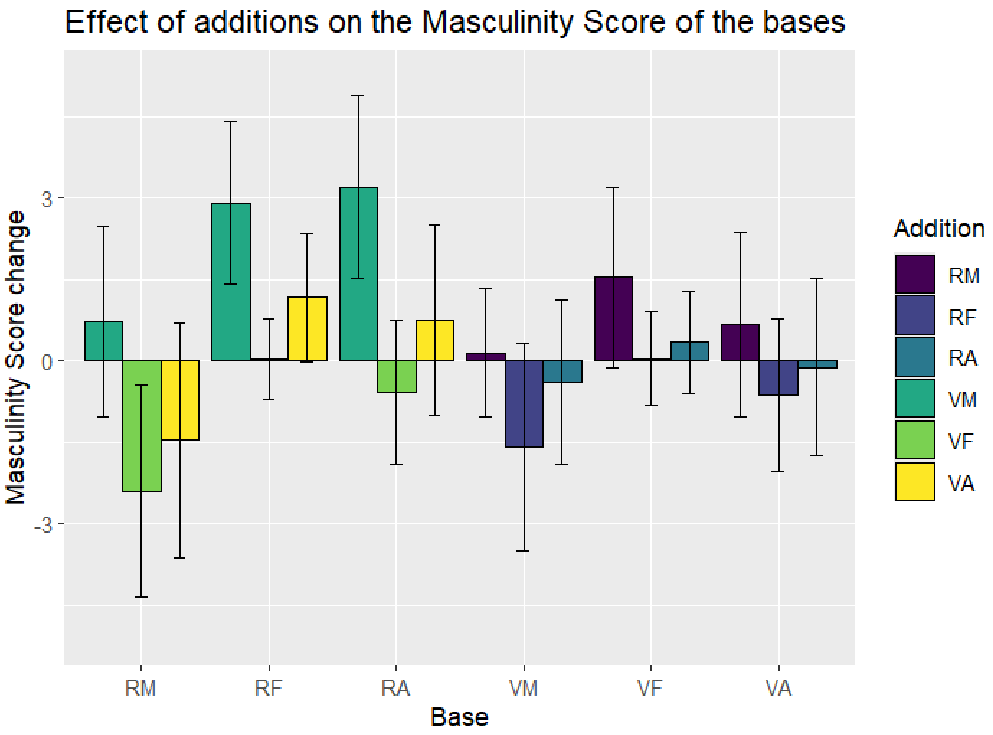


FIGURE 5.3: Bar plot showing the mean change (with standard deviation) of the masculinity score of a base after a robot or voice addition. The bases on the x-axis and the additions in the legend starting with *R* are robots, those starting with *V* are voices. The *M*, *F*, and *A* signify their gender (masculine, feminine, ambiguous).

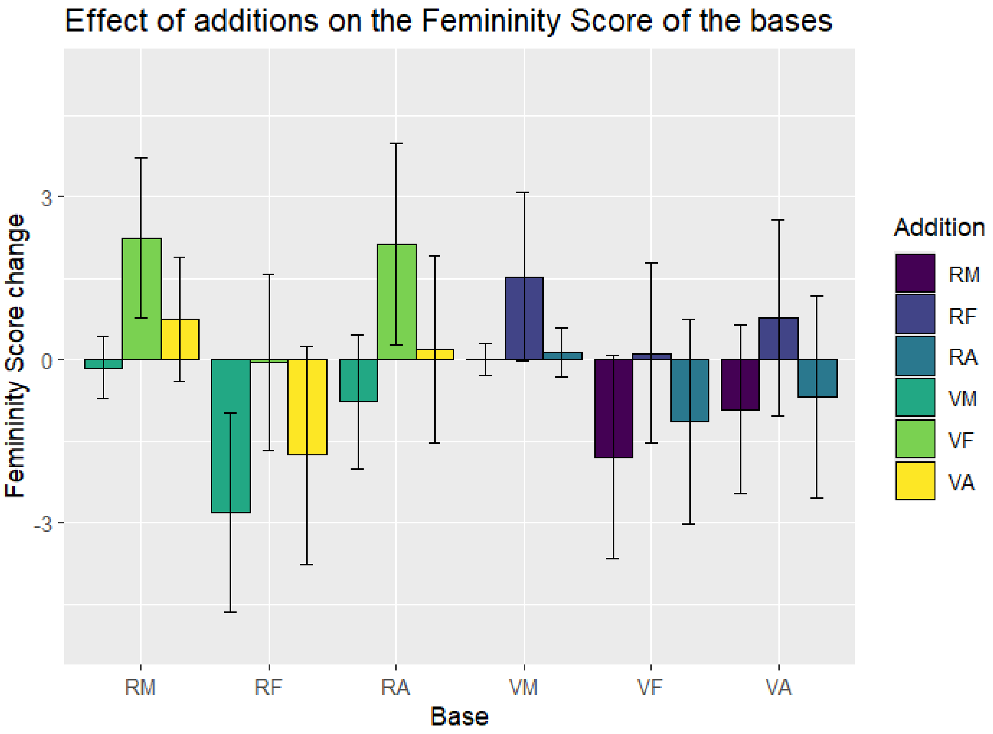


FIGURE 5.4: Bar plot showing the mean change (with standard deviation) of the femininity score of a base after a robot or voice addition.

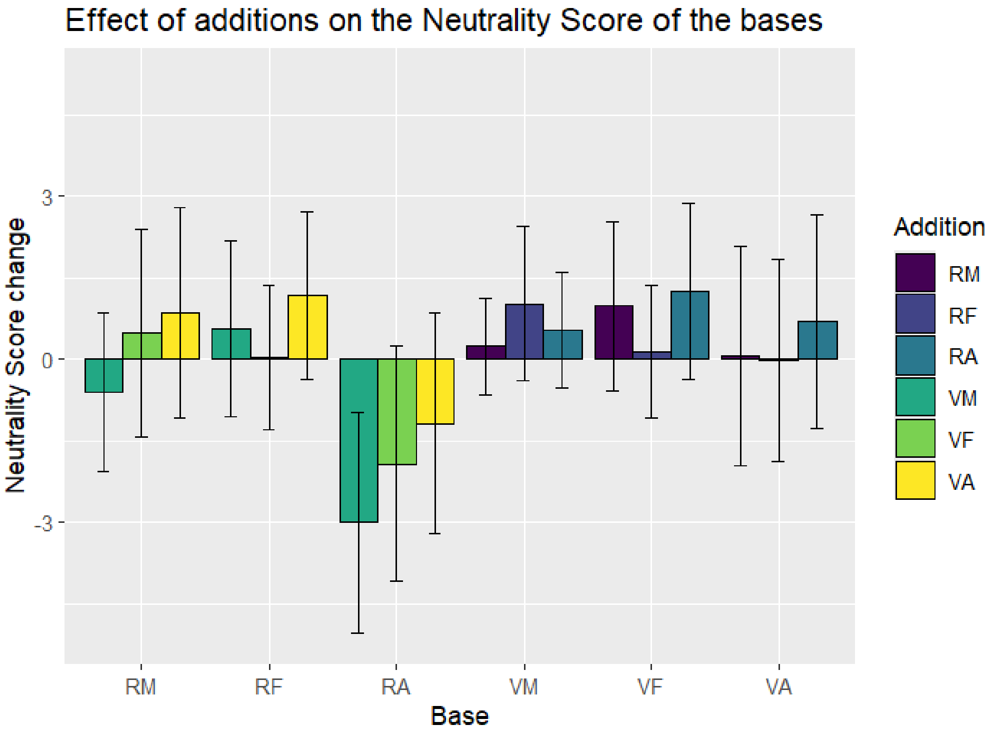


FIGURE 5.5: Bar plot showing the mean change (with standard deviation) of the neutrality score of a base after a robot or voice addition.

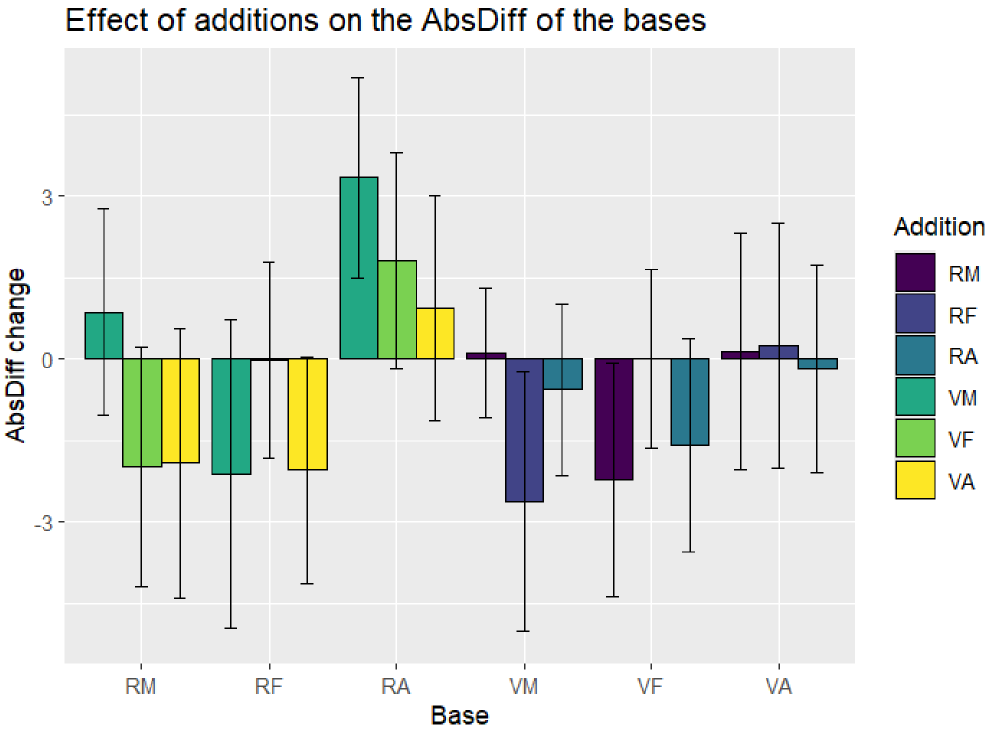


FIGURE 5.6: Bar plot showing the mean change (with standard deviation) of the AbsDiff of a base after a robot or voice addition.

TABLE 5.1: The results from the gender categorisation of the robot-voice combinations. The table shows the robot and voice that make up a combination, coloured corresponding to their individual gender. Followed by the gender of the combination. Blue represents masculinity, pink represents femininity, and green represents ambiguity.

Combo	Robot	Voice	Combo Gender	Combo	Robot	Voice	Combo Gender
a1	MR_6	M1	Masc	d1	MR_6	sam_f2	Amb
a2	MR_4	M1	Masc	d2	MR_4	sam_f2	Fem(Amb)
a3	FR_1	M1	Amb	d3	FR_1	sam_f2	Fem
a4	FR_6	M1	Masc	d4	FR_6	sam_f2	Fem
a5	AR_6	M1	Masc	d5	AR_6	sam_f2	Fem
a6	AR_4	M1	Masc	d6	AR_4	sam_f2	Fem
b1	MR_6	M2	Masc	e1	MR_6	F1	Masc
b2	MR_4	M2	Masc	e2	MR_4	F1	Amb
b3	FR_1	M2	Masc(Amb)	e3	FR_1	F1	Fem
b4	FR_6	M2	Masc	e4	FR_6	F1	Fem
b5	AR_6	M2	Masc	e5	AR_6	F1	Amb
b6	AR_4	M2	Masc	e6	AR_4	F1	Amb
c1	MR_6	sam_f1	Amb	f1	MR_6	N8	Masc
c2	MR_4	sam_f1	Amb	f2	MR_4	N8	Masc
c3	FR_1	sam_f1	Fem	f3	FR_1	N8	Fem(Amb)
c4	FR_6	sam_f1	Fem	f4	FR_6	N8	Amb
c5	AR_6	sam_f1	Fem	f5	AR_6	N8	Masc(Amb)
c6	AR_4	sam_f1	Fem	f6	AR_4	N8	Masc(Amb)

Combinations with an AbsDiff between 1 and 1.5 are still categorised as either masculine or feminine, but with an added ambiguous tag as the AbsDiff is still relatively low.

Of the 36 robot-voice combinations, 16 were categorised as masculine (of which three had an ambiguous tag), 12 were categorised as feminine (of which two had an ambiguous tag), and eight were categorised as ambiguous. When comparing the genders of the robot-voice combinations with the genders of the individual robots and voices, most combinations consisting of a same-gendered robot and voice were also categorised as that same gender. Only two combinations were not, both of which consisted of an ambiguous robot and voice but were categorised as masculine with an ambiguous tag (Combo *f5* and *f6*). Of the eight total combinations that were categorised as ambiguous, four consisted of a masculine and a feminine stimulus, while only two consisted of two ambiguous stimuli, possibly suggesting that combinations consisting of a masculine and a feminine stimulus may also be a viable approach for the creation of ambiguous speaking social robots. However, the reported attitudes from several participants towards such combinations, as mentioned in the previous section, may preclude this.

Disregarding the robot-voice combinations where the robot and the voice had the same gender, ten combinations were categorised as masculine (of which one had an ambiguous tag), eight combinations were categorised as feminine (of which two had an ambiguous tag), and six were categorised as ambiguous. Of these combinations, in fourteen cases the gender was the same as that of the voice (58%), in six cases the gender of the combination was the same as that of the robot (25%), and in four cases the gender of the combination was equal to neither that of the voice nor that of the robot (17%). The cases where the gender of the combination matched neither the voice nor the robot always consisted of one masculine and one feminine stimulus while being categorised as ambiguous; three times this was a feminine voice with a masculine robot (Combo *c1*, *c2*, and *d1*), and one time it was a masculine voice with a feminine robot (Combo *a3*). In the six cases where the gender of the combination matched the gender of the robot, the voice was always ambiguous while the robot was either masculine or feminine. Excluding robot-voice combinations where the robot and the voice had the same gender, only two combinations which contained an ambiguous stimulus were categorised as ambiguous (Combo *e2* and *f4*). One of these combinations consisted of an ambiguous voice and a masculine robot, the other consisted of an ambiguous voice and a feminine robot.

Considering all (8) combinations that consisted of one masculine and one feminine stimulus, three combinations were categorised as masculine (Combo *a4*, *b3*, and *b4*, of which *b4* had an ambiguous tag), these all consisted of a masculine voice and a feminine robot. One combination was categorised as feminine (with an ambiguous tag), this combination consisted of a feminine voice and a masculine robot (Combo *d2).* Four combinations were categorised as ambiguous, of which three consisted of a feminine voice and a masculine robot (Combo *c1*, *c2*, and *d1*), and one consisted of a masculine voice and a feminine robot (Combo *a3*). Of these eight combinations, four had the same gender as neither the voice nor the robot (the ambiguously categorised combinations), the remaining four combinations all had the same gender as the voice.

5.2.3 Masculine Base Models

The results of the linear mixed models (as explained in Chapter 5.1.6) are presented per base gender, with a focus on the gender score most related to the base gender, i.e. for masculine bases the focus is on the model for the masculinity-delta, while for feminine bases the focus is on the model for the femininity-delta.

When examining the interaction plot for the masculinity-delta of masculine bases (see Figure 5.7), both feminine and ambiguous gendered additions are found to lower the masculinity score, feminine additions more so than ambiguous additions. The interaction plot suggests a greater effect from voice additions than from robot additions.

However, when considering the results from the linear mixed model (see Table 5.2), voice additions did not show a significant effect in comparison to robot additions. Though, it should be noted that the significance here is only calculated when the addition gender is at the reference level (masculine in this case). There was a significant interaction effect between the addition type and gender, showing a decrease of the masculinity-delta when feminine voices (β = -1.39, SE = 0.30, p = 0.004) or ambiguous voices (β = -1.68, SE = 0.30, p = 0.001) were added. Aside from that, feminine gendered additions were significant, lowering the masculinity-delta (β = -1.73, SE = 0.21, $p < 0.001$), while ambiguous gendered additions, which also lowered the masculinity-delta, were only marginally significant (β = -0.50, SE = 0.30, p = 0.06). Finally, the model had a conditional R^2 (model fit that takes both random and fixed effects into account) of 0.464.

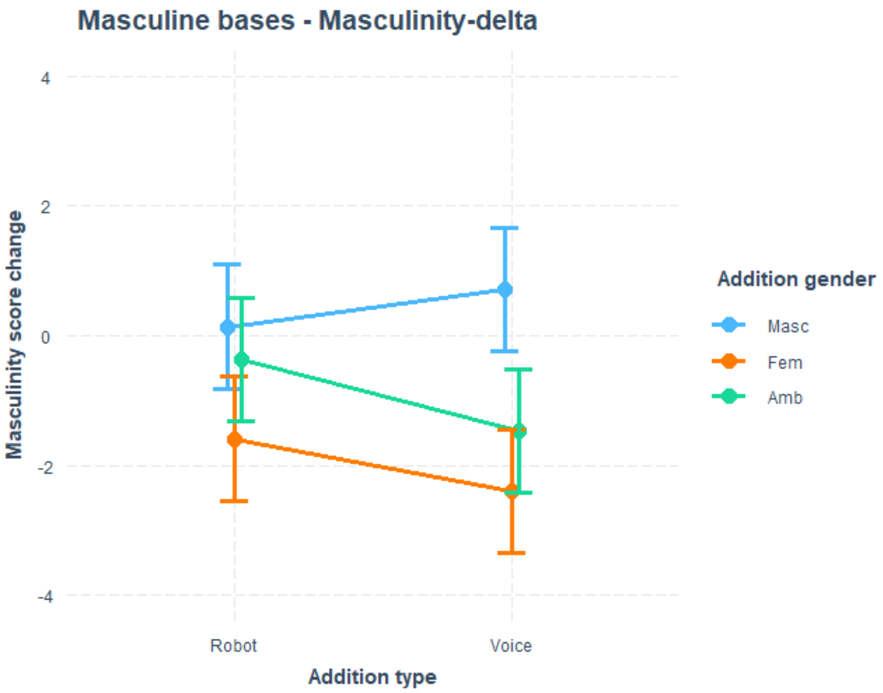


FIGURE 5.7: Interaction plot with 95% confidence intervals of the masculinity-delta model of the masculine bases, showing the effect of the addition type and addition gender on the masculinity-delta

TABLE 5.2: The linear mixed model results for the masculinity-delta of masculine bases

Fixed Effect	Estimate	Standard Error	95% Confidence Interval	p-value
Intercept	0.1382	0.4865	[-0.7344904 , 1.0108100]	0.797384
addtype_V	0.5789	0.6688	[-0.6278076, 1.7857088]	0.466334
addgen_A	-0.5	0.2118	[-0.8603859, -0.1396141]	0.056217
addgen_F	-1.7303	0.2118	[-2.0906491, -1.3698773]	**0.000181**
addtype_V:addgen_A	-1.6842	0.2995	[-2.1938732, -1.1745479]	**0.001352**
addtype_V:addgen_F	-1.3882	0.2995	[-1.8978206, -0.8784953]	**0.00356**

Random Effect	Standard Deviation		
participant	0.7048		
base_id	0.6344		
add_id	0.1151	Conditional R^2	0.464
Residual	1.5499	Marginal R^2	0.261

Examining the models and interaction plots of the other gender scores of the masculine bases (see Appendix N), none of the models recorded a significant effect of voice additions in comparison to robot additions. All models did record one significant interaction effect each. In the femininity-delta model, feminine voice additions caused an increase of the femininity-delta (β = 0.87, SE = 0.34, p = 0.05), while ambiguous voice additions also had a marginally significant effect, increasing the femininity-delta (β = 0.71, SE = 0.34, p = 0.08). In the neutrality-delta model, ambiguous voice additions showed an increase in the neutrality-delta (β = 1.10, SE = 0.21, $p < 0.001$). Lastly, in the AbsDiff-delta model ambiguous voice additions decreased the AbsDiff-delta (β = -2.11, SE = 0.28, $p < 0.001$). Additionally, apart from ambiguous gendered additions in the model for the femininity-delta, all addition genders showed significant effects for the remaining models. Feminine additions increased the femininity-delta (β = 1.52, SE = 0.24, p < 0.001) and neutrality-delta (β = 0.79, SE = 0.15, p < 0.001), while decreasing the AbsDiff-delta (β = -2.74, SE = 0.20, p < 0.001). Similarly, ambiguous additions increased the neutrality-delta (β = 0.37, SE = 0.15, p = 0.01) while decreasing the AbsDiff-delta (β = -0.68, SE = 0.20, p < 0.001). Finally, the femininity and AbsDiff-delta models showed the greatest model fit, with a conditional R^2 of 0.503 and 0.505 respectively.

5.2.4 Feminine Base Models

Similarly to the masculine bases, based on the interaction plot (see Figure 5.8), additions with an incongruent gender from the base lowered the femininity score of feminine bases. Here too, a voice addition seemed to have a greater impact than a robot addition.

The results from the linear mixed model (see Table 5.3) were also similar to those from the model for the masculinity-delta of masculine bases. Finding a significant interaction effect, with the addition of masculine voices showing a decrease of the femininity-delta (β = -0.87, SE = 0.26, p = 0.02). Voice additions did not have a significant main effect, while addition genders did, with both masculine gendered additions (β = -1.91, SE = 0.19, p < 0.001) and ambiguous gendered additions (β = -1.31, SE = 0.19, p < 0.001) lowering the femininity-delta. The conditional R^2 showed a model fit of 0.425.

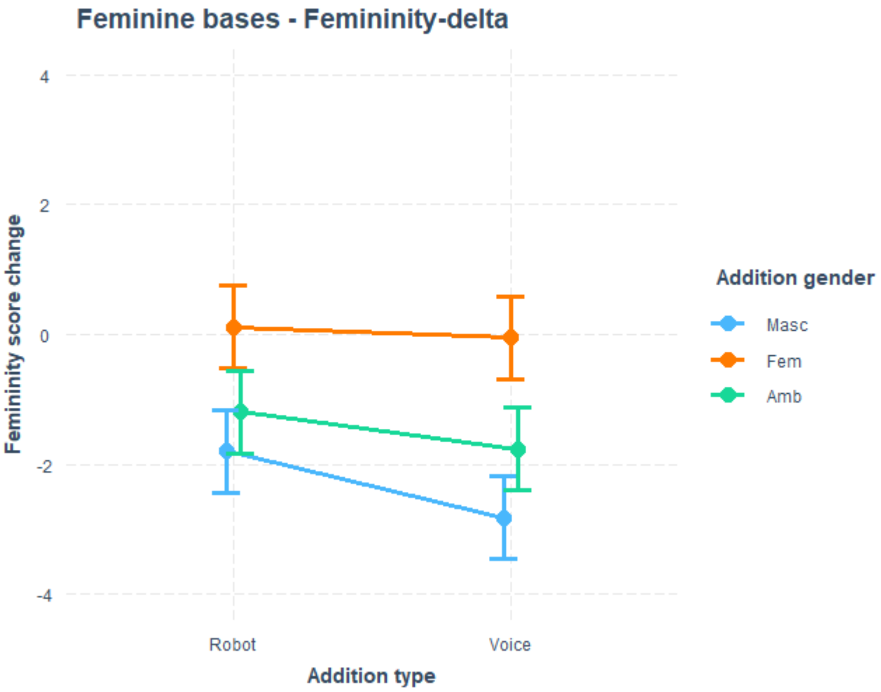


FIGURE 5.8: Interaction plot with 95% confidence intervals of the femininity-delta model of the feminine bases, showing the effect of the addition type and addition gender on the femininity-delta

TABLE 5.3: The linear mixed model results for the femininity-delta of feminine bases

Fixed Effect	Estimate	Standard Error	95% Confidence Interval	p-value
Intercept	0.1118	0.3241	[-0.4859441, 0.7096294]	0.748229
addtype_V	-0.1579	0.4168	[-0.9427645, 0.6269785]	0.733396
addgen_A	-1.3092	0.1855	[-1.6685974, -0.9498238]	**0.000404**
addgen_M	-1.9079	0.1855	[-2.2672816, -1.548508]	**4.93E-05**
addtype_V:addgen_A	-0.4079	0.2623	[-0.9161445, 0.1003549]	0.170898
addtype_V:addgen_M	-0.8684	0.2623	[-1.3766709, -0.3601714]	**0.016183**

Random Effect	Standard Deviation		
participant	0.83175		
base_id	0.37323		
add_id	0.02703	Conditional R^2	0.425
Residual	1.59954	Marginal R^2	0.238

Similar to the masculine base models, none of the models for the feminine bases (see Appendix O) showed a significant effect of voice additions in comparison to robot additions. The neutrality and AbsDiff-delta models did not record a significant interaction effect. The masculinity-delta model did, as masculine voice additions increased the masculinity-delta (β = 1.39, SE = 0.34, p = 0.007), while ambiguous voice additions also saw a marginally significant effect, increasing the masculinity-delta (β = 0.74, SE = 0.34, p = 0.08). Furthermore, all addition genders were significant except for ambiguous additions in the masculinity-delta model. Masculine additions increased the masculinity-delta (β = 1.49, SE = 0.24, p < 0.001) and neutrality-delta (β = 0.85, SE = 0.15, p < 0.001), while decreasing the AbsDiff-delta (β = -2.22, SE = 0.24, p < 0.001). Ambiguous additions also increased the neutrality-delta (β = 1.16, SE = 0.15, p < 0.001) while decreasing the AbsDiff-delta (β = -1.72, SE = 0.21, p < 0.001). The masculinity-delta model saw the greatest conditional R^2 of the feminine base models, at 0.49.

5.2.5 Ambiguous Base Models

While the interaction plots of the neutrality and AbsDiff-delta of ambiguous bases are complete opposites of each other (see Figure 5.9 and 5.10), high gender ambiguity being represented by high neutrality scores and low AbsDiffs, suggests that the plots show very similar results. Both showed little change for robot additions, and much more for voice additions. Additionally, the feminine and ambiguous addition genders were almost parallel in both interaction plots, while the masculine addition gender showed a much steeper angle between robot and voice additions. Interestingly, the addition of a voice seemed to always have a negative effect on ambiguity regardless of how ambiguity is measured, lowering the neutrality score while increasing the AbsDiff.

The linear mixed models for both the neutrality and AbsDiff-delta (see Table 5.4 and 5.5) showed a significant effect of voice additions compared to robot additions, lowering the neutrality-delta (β = -1.67, SE = 0.26, p < 0.001) while increasing the AbsDiff-delta (β = 1.03, SE = 0.25, p < 0.001). Both models also recorded a significant interaction effect when masculine voices were added, amplifying the decrease of the neutrality-delta (β = -1.11, SE = 0.36, p = 0.02), and the increase of the AbsDiff-delta (β = 2.11, SE = 0.34, p < 0.001). In the neutrality-delta model, feminine gendered additions recorded a significant effect lowering the neutrality-delta (β = -0.69, SE = 0.25, p = 0.03), while masculine gendered additions only recorded a marginally significant effect, lowering the neutrality-delta (β = -0.61, SE = 0.25, p = 0.05). Contrarily, the

AbsDiff-delta model showed no significant effects from the addition genders. The conditional R^2 showed a model fit of 0.438 and 0.34 for the neutrality-delta model and the AbsDiff-delta model respectively.

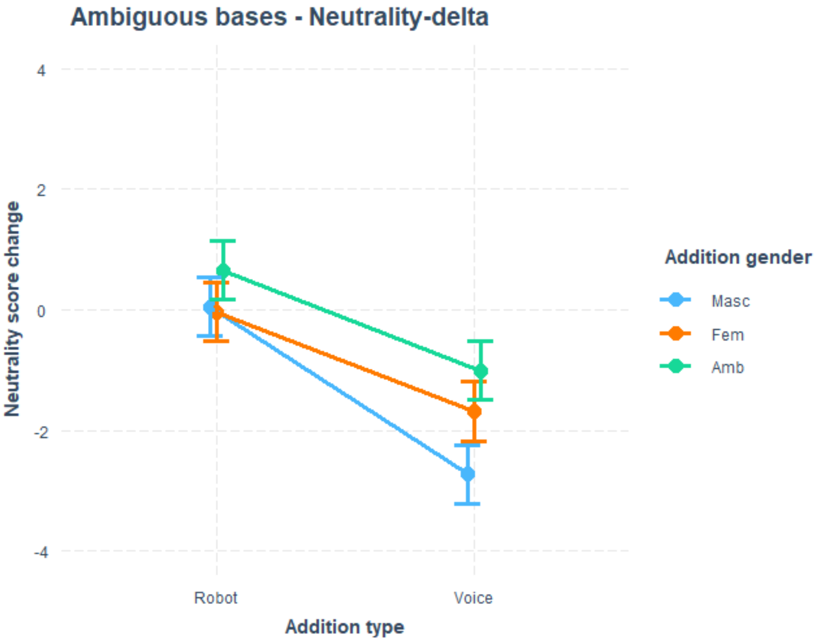


FIGURE 5.9: Interaction plot with 95% confidence intervals of the neutrality-delta model of the ambiguous bases, showing the effect of the addition type and addition gender on the neutrality-delta

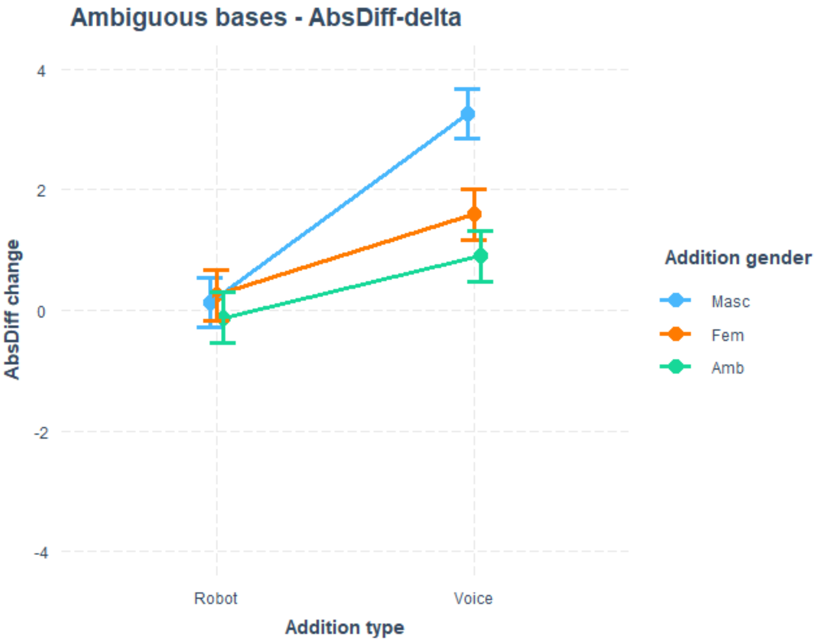


FIGURE 5.10: Interaction plot with 95% confidence intervals of the AbsDiff-delta model of the ambiguous bases, showing the effect of the addition type and addition gender on the AbsDiff-delta

TABLE 5.4: The linear mixed model results for the neutrality-delta of ambiguous bases

Fixed Effect	Estimate	Standard Error	95% Confidence Interval	p-value
Intercept	0.65789	0.24903	[0.2350550, 1.0807179]	**0.017828**
addtype_V	-1.67105	0.26402	[-2.0651886, -1.2769167]	**0.000944**
addgen_F	-0.69079	0.25273	[-1.0849254, -0.2966535]	**0.034034**
addgen_M	-0.61184	0.25273	[-1.0059781, -0.2177061]	0.051794
addtype_V:addgen_F	0.01974	0.35741	[-0.5376556 , 0.5771293]	0.957754
addtype_V:addgen_M	-1.10526	0.35741	[-1.6626556, -0.5478707]	**0.021321**

Random Effect	Standard Deviation		
participant	1.01591		
base_id	0.07641		
add_id	0.15326	Conditional R^2	0.438
Residual	1.75189	Marginal R^2	0.244

TABLE 5.5: The linear mixed model results for the AbsDiff-delta of ambiguous bases

Fixed Effect	Estimate	Standard Error	95% Confidence Interval	p-value
Intercept	-0.125	0.2135	[-0.50399429, 0.2539943]	0.57135
addtype_V	1.0263	0.2523	[0.59647801, 1.4561536]	**0.009433**
addgen_F	0.375	0.2379	[-0.05483778, 0.8048378]	0.165993
addgen_M	0.2566	0.2379	[-0.17325883, 0.6864167]	0.322195
addtype_V:addgen_F	0.3158	0.3364	[-0.29209295, 0.9236719]	0.384111
addtype_V:addgen_M	2.1053	0.3364	[1.49738074, 2.7131456]	**0.000772**

Random Effect	Standard Deviation		
participant	0.72342		
base_id	0.08412		
add_id	0.09176	Conditional R^2	0.34
Residual	1.91324	Marginal R^2	0.243

The results from the masculinity and femininity-delta models were opposites of each other (see Appendix P). Considering the interaction plots, masculine additions increased the masculinity score while decreasing the femininity score, while feminine additions increased the femininity score while decreasing the masculinity score. Interestingly, the decreasing effect masculine and feminine additions had on the femininity and masculinity scores showed almost no difference between the robot and voice addition types, while in both cases the increasing effect was much greater from voice additions than from robot additions. Additionally, masculine additions caused both greater increases and greater decreases of masculinity and femininity scores, than feminine additions. Lastly, the addition of ambiguous voices caused greater change in the masculinity and femininity score than the addition of ambiguous robots.

Regarding the linear mixed models, voice additions recorded a significant effect increasing both the masculinity-delta (β = 0.79, SE = 0.29, p = 0.03) and the femininity-delta (β = 0.84, SE = 0.27, p = 0.04). Both the masculinity and femininity-delta models also showed a significant interaction effect, as masculine voice additions amplified the increase in the masculinity-delta (β = 1.68, SE = 0.41, p = 0.006) and the decrease in the femininity-delta (β = -0.71, SE = 0.27, p = 0.04). Masculine additions were found to be significant in the masculinity-delta model, having a positive effect (β = 0.75, SE = 0.29, p = 0.04). Feminine additions were found to be significant in the femininity-delta model, also having a positive effect (β = 1.45, SE =

0.19, $p < 0.001$). The masculinity-delta model had the highest conditional R^2 of the ambiguous base model, at 0.491.

Chapter 6

Discussion

6.1 RQ1 — The influence of embodiment and voice on the gender perception of speaking social robots, and each other

RQ1 concerned how robot embodiments and voices would influence each other's perceived gender and the perceived gender of a speaking social robot. Two hypotheses related to this question were constructed. First, ambiguous stimuli were expected to have no effect on the gender perception (scores) of robot-voice combinations, if the other stimulus was binary gendered (**H1**). Second, if the robot and the voice have incongruent genders and neither is ambiguous, the robot was expected to have a greater effect on the gender scores of the robot-voice combination, than the voice. In the same scenario, the robot was expected to have a greater effect on the gender scores of the voice, than the voice would have on the robot (**H2**).

When comparing the impact voices and robots had on each other's gender scores, whether the gender scores increased or decreased was very much dependent on the specific gender score measured, the gender of the base, and the gender of the addition. However, the rate of the increase or decrease seems to be related to the type of the addition, whether it is a robot or a voice. This follows from the interaction plots of the linear mixed models, which show that in most cases the change of the gender scores caused by voice additions is further removed from zero than their robot addition counterparts. Additionally, the results from the gender categorisation of the robot-voice combinations showed that 58% of combinations (excluding combinations consisting of same-gendered robots and voices) had the same gender as the voice, while only 25% had the same gender as the robot. When only considering combinations consisting of one masculine and one feminine stimulus this changed to 50% for voices and 0% for robots. Thus, pointing towards a greater impact from voices than from robots, on gender scores.

However, when considering the results of the mixed linear models, voices did not always have a greater impact on the gender scores, than robots did. Voices on their own only showed a reliable difference from robots in the models for the ambiguous bases, though this may have been caused by individual differences in ambiguity between the ambiguous robots and the ambiguous voices. In the models for masculine and feminine bases, voices only showed a reliable difference from robots in combination with a specific gender. Such a reliable difference was, for example, found when considering combinations consisting of one masculine and one feminine stimulus. Consequently, hypothesis **H2**, suggesting dominance of visual stimuli for the gender perception of combinations consisting of one masculine and one feminine stimulus, is rejected. Additionally, it should be noted that auditory dominance also seems to extend into differently gendered combinations, though this is dependent on the gender of both the voice and the robot.

As a whole, these results match with those from Paetzel et al. [96], who found dominance of auditory stimuli over visual stimuli for the gender perception of speaking robots by children. It should be noted that they only used one masculine and one feminine visual stimulus, created through the projection of a masculine or a feminine face onto the Furhat robot. The human-likeness of these visual stimuli was not considered and may therefore differ from the visual stimuli used in the current study.

The rejection of the hypothesis may be related to the difference in stimuli used in the current study, compared to the studies at the base of the hypothesis, as the stimuli used there were human faces and voices instead of robots and computer voices [8][90]–[92]. This represents a major difference in the anthropomorphism of the used stimuli, which may have caused the differing results. Considering the information reliability hypothesis [93], these results suggest that visual stimuli of relatively low anthropomorphism are less appropriate and efficient than computer voices for gender perception.

Considering the effect of ambiguous stimuli when paired with binary gendered stimuli. Two out of sixteen combinations of an ambiguous stimulus and a binary gendered stimulus were categorised as ambiguous. Several models for masculine and feminine bases, mainly those for the AbsDiff and Neutrality Score, also showed a reliable effect from ambiguous additions. Three out of four models for the masculine bases also recorded a reliable effect in combination with voice additions. While these effects are small in comparison to the effects of masculine and feminine stimuli, they are significant. Therefore, the findings in this study diverge slightly from earlier studies, which found or inferred dominance of the binary gendered stimulus [12][37][40]–[49][73][94][95]. However, it should be noted that the method used for measuring gender perception in the current study was more nuanced than what was used previously, possibly allowing for more nuanced analysis which may have led to differing findings. Based on these results, hypothesis **H1**, suggesting no impact from ambiguous stimuli on gender perception results, when combined with binary gendered stimuli, is rejected.

These results provide new insights into the interplay between auditory and visual stimuli regarding the gender perception of speaking social robots. It builds on the findings of Paetzel et al. [96], concurring that auditory stimuli are more important than visual stimuli, while adding that their importance relative to each other is also, at least partially, related to their individually perceived genders. Additionally, it adds to current research through the inclusion of ambiguous stimuli, while also showing that ambiguous stimuli have a measurable effect on the gender scores of binary gendered stimuli, though this effect is generally smaller compared to that of masculine and feminine stimuli.

6.2 RQ2 — The most promising robot-voice combination for the creation of gender ambiguous speaking social robots

RQ2 concerned which kinds of gendered robot-voice combinations would be most promising for the creation of gender ambiguous speaking social robots. It was expected that combinations consisting of two ambiguous stimuli would be the best, as they would score higher on gender ambiguity measures than other combinations (**H3**).

A total of eight out of 36 combinations were categorised as ambiguous. Two of these consisted of an ambiguous robot and voice, while four were combinations of a masculine and a feminine stimulus (three of which were a feminine voice combined with a masculine robot). The last two both had ambiguous voices, one in combination with a masculine robot, the other with a feminine robot. Based on these results the most promising combinations to create ambiguous speaking social robots may be those consisting of a masculine and a feminine stimulus, specifically a feminine voice and a masculine robot. However, the highest neutrality score and the lowest AbsDiff are both found on combinations consisting of two ambiguous stimuli, though not the same combination.

As a result, the hypothesis can be accepted on the basis that the combinations that scored the highest on gender ambiguity measures consisted of two ambiguous stimuli. However, seeing as more ambiguous combinations consist of a masculine and a feminine stimulus, it can be argued that these combinations, especially those consisting of a feminine voice and a masculine robot, are the most promising for the creation of gender ambiguous speaking social robots. Still, based on the remarks from some survey participants, who reported confusion, or finding it weird when they were presented with combinations of stimuli with clearly mismatched genders, extra research should be done to study people's attitudes towards speaking social robots that present a clear mismatch between embodiment and voice gender.

As an aside, when comparing the results of combinations consisting of a feminine robot and a masculine voice, with combinations consisting of a masculine robot and a feminine voice. Three out of four of the combinations with a masculine voice are categorised as masculine, while three out of four of the combinations with a feminine voice are categorised as ambiguous. This aligns with findings from Paetzel et al. [96], who found that almost twice as many kids categorised combinations of a feminine auditory and a masculine visual stimulus as neutral, than the inverse, while also being rated slightly more difficult to assign a gender. However, it should be noted that these combinations were only categorised as neutral in less than 40% of cases. From these results, as Paetzel et al. [96] noted as well, the dominance of auditory stimuli over visual stimuli seems more apparent in combinations where the auditory stimulus is masculine. This, in turn, may suggest a greater impact from masculine stimuli than feminine stimuli, which is somewhat consistent with earlier research suggesting robots are perceived as masculine by default [7][37][64][65].

These results show that aside from combinations of ambiguous voices and embodiments, combinations of a masculine and a feminine stimulus may also be an option for the creation of gender ambiguous speaking social robots. However, neither of these methods is currently fully reliable for the design of gender ambiguous speaking social robots. While two combinations consisting of ambiguous voices and embodiments were categorised as ambiguous, two others were categorised as masculine (possibly caused by individual differences between the used voices). Highlighting the fine line between ambiguity and non-ambiguity, the necessity of using stimuli that are clearly ambiguous, and the difficulty in creating them. Additionally, while three out of four combinations of feminine voices and masculine embodiments were categorised as ambiguous, participants noted feeling confused when presented with combinations of clearly incongruent genders. As such, it is unclear how people would treat and interact with such a speaking social robot. Emphasising the necessity for research concerning people's attitudes towards speaking social robots consisting of a masculine and a feminine stimulus.

6.3 Limitations

The current study investigated the effects of voice and embodiment on the gender perception of speaking social robots, and each other, doing so through the presentation of voice clips and robot images individually, and in pairs. As such, the gender perception results are based on the physical appearance of the robots, and the sound of the voices. While, as noted in Chapter 2.3 and 2.4, other factors such as the perceived personality of, and the tasks performed by the robot or voice may also influence the perceived gender, these factors were outside the scope of this study and therefore neutralised during the experiments (by having the voices say a nonsensical sentence, and presenting the robots in a neutral position against a white background). As a result, the findings presented here do not account for these factors, and thus may not fully capture how people perceive the gender of speaking social robots in more natural (real world) contexts.

Additionally, the selected robots were controlled for their level of anthropomorphism, where robots of comparatively low anthropomorphism were selected to be used in the experiments. As anthropomorphism and gender seem to be linked (see Chapter 2.2), the use of robots of a higher level of anthropomorphism may have effects on the perceived gender that were not captured in the current study.

The reliability and generalisability of the results are also impacted by the relatively small amount of robots and voices used, as only two stimuli were used per gender per type in the main experiment. While this was done to limit the size of the survey for the main experiment (as it was deemed preferable to have all participants evaluate all robot-voice combinations, instead of splitting up the survey and only showing each participant a subset), it may have allowed for characteristics of individual stimuli to influence the results.

Additionally, especially due to the low number of stimuli per condition (2), stimuli of the same gender category in the main experiment would have preferably had similar gender scores, which was not always the case. Differences were present between the ambiguity of the ambiguous robots and voices, which may have impacted the results. Though this likely should have been expected due to the known difficulty of creating gender ambiguous voices (see Chapter 2.4). Similarly, the difference between the masculine FGSs of the two masculine robots used (10.9 vs. 7.2), as well as between the feminine FGSs of the two feminine robots used (10.8 vs. 8.4), may have affected the results.

The participant samples used in the pre-test and the main-experiment may not have been optimally representative of the target population due to the use of convenience sampling for both surveys, as this led to a relatively large portion of the responses coming from students. Similarly, while the use of SurveyCircle to recruit participants for the main experiment allowed for increased geographical diversity, many of the respondents were likely academics due to the nature of the website.

The creation of separate models for each base gender may have limited the scope of the analysis. However, not using separate models and instead adding the base gender to the model as a third independent variable resulted in errors in both R and SPSS, which prohibited analysis. As such, making separate models per base gender was not only done to avoid three-way interactions, but also out of necessity.

6.4 Design Recommendations

Based on the auditory dominance (when the voice is binary gendered) and the relatively low impact of ambiguous stimuli on gender perception found in this study, designers of speaking social robots may consider prioritising the design, development, or selection of the voice, in combination with a broader adoption of ambiguously gendered embodiments when pursuing specifically gendered design goals. Following the presented findings, exclusively using ambiguously gendered embodiments while using only the voice to provide the desired gender cues, is sufficient to reach the desired gender. Simultaneously, this would reduce the complexity of any coordination between voice and embodiment design concerning the desired gender outcome, due to the low impact of gender ambiguous embodiments on gender perception. Furthermore, this would increase the adaptability of the social robot, as simply changing the voice would change the perceived gender of the social robot in its entirety. Moreover, if these embodiments or their designs are reused, which may be more likely due to the increased adaptability, it would reduce the necessary resources for the design of the embodiment, and instead allow for greater focus on the design of the voice.

Chapter 7

Conclusion

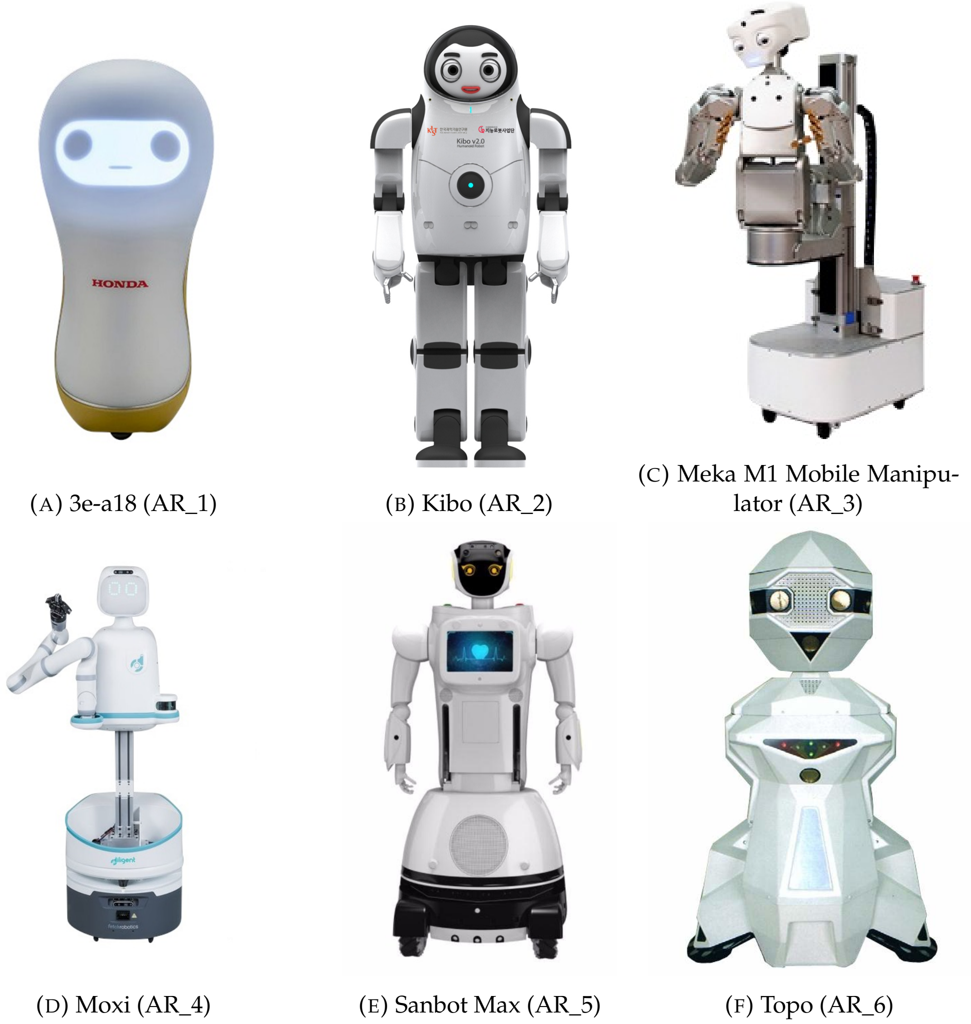
This research shows that auditory stimuli are leading for the gender perception of speaking social robots, when the voice and the embodiment have incongruent genders and the voice is not gender ambiguous. Additionally, it finds that speaking social robots consisting of one binary gendered and one ambiguously gendered stimulus are generally categorised the same as the binary gendered stimulus. However, it also shows that ambiguous stimuli do have a measurable effect on the gender scores of binary gendered stimuli, typically affecting the scores related to gender ambiguity such that they suggest an increase in ambiguity.

The findings demonstrate the delicate interplay between voice and embodiment for the gender perception of speaking social robots and highlight the necessity for careful consideration of both embodiment and voice gender in the design of speaking social robots, especially when gender is one of the design criteria. Furthermore, it highlights the difficulty of creating an ambiguous speaking social robot, before even considering the possible impact of additional personality- or task-based gender cues. However, based on the gender perception results, it does imply that creating robot-voice combinations of which one stimulus is masculine and the other is feminine, may be an additional viable method for creating speaking social robots that are perceived as ambiguous, though further research studying people's attitudes towards such combinations is required.

In future work, this research could be expanded by conducting experiments that include real or simulated interactions between a person and a social robot, to investigate additional factors that may influence the perceived gender, such as the perceived personality or the job of the social robot. Furthermore, future research is needed to determine if these results can be generalised to contexts that include robot embodiments of a higher level of human-likeness. Finally, future studies should address people's attitudes towards speaking social robots where the individual genders of the robot embodiment and the voice are incongruent, to further understand the possible implications of this research.

Appendix A

Ambiguous Robot Images



(A) 3e-a18 (AR_1)

(B) Kibo (AR_2)

(C) Meka M1 Mobile Manipulator (AR_3)

(D) Moxi (AR_4)

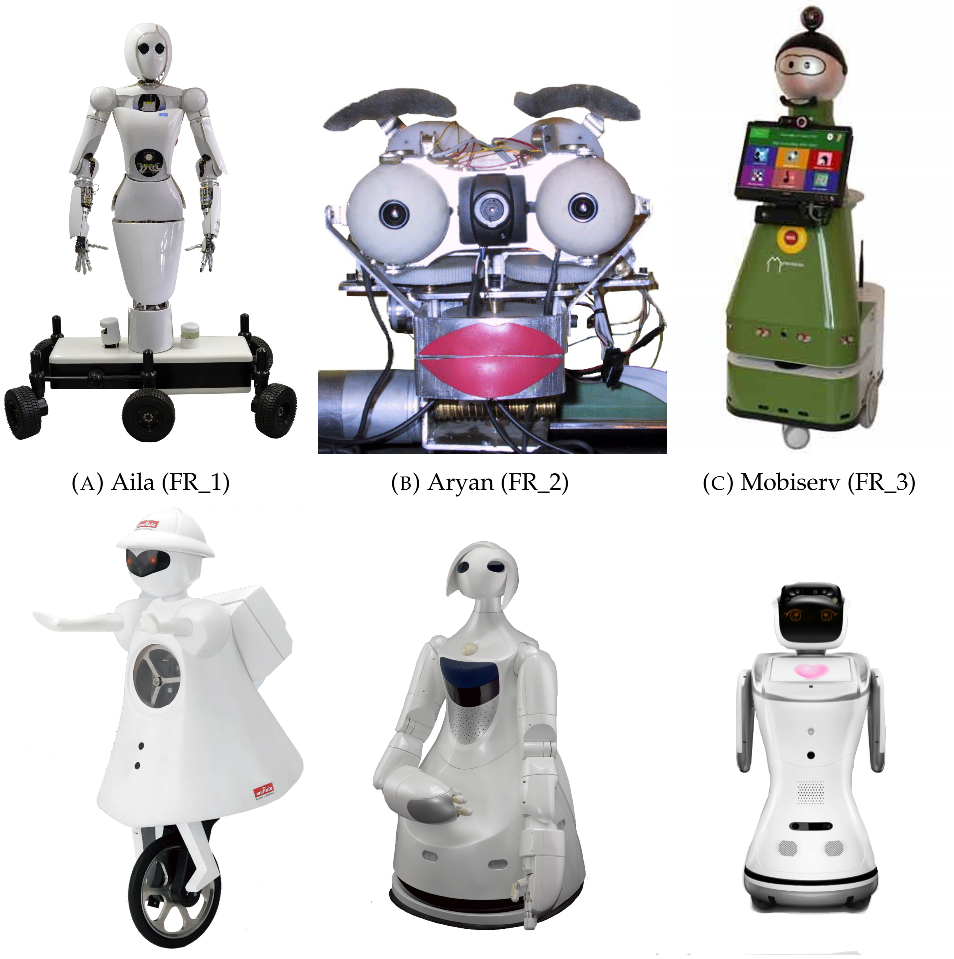
(E) Sanbot Max (AR_5)

(F) Topo (AR_6)

FIGURE A.1: The ambiguous robots selected for the pre-test

Appendix B

Feminine Robot Images



(A) Aila (FR_1)

(B) Aryan (FR_2)

(C) Mobiserv (FR_3)

(D) Murata Girl (FR_4)

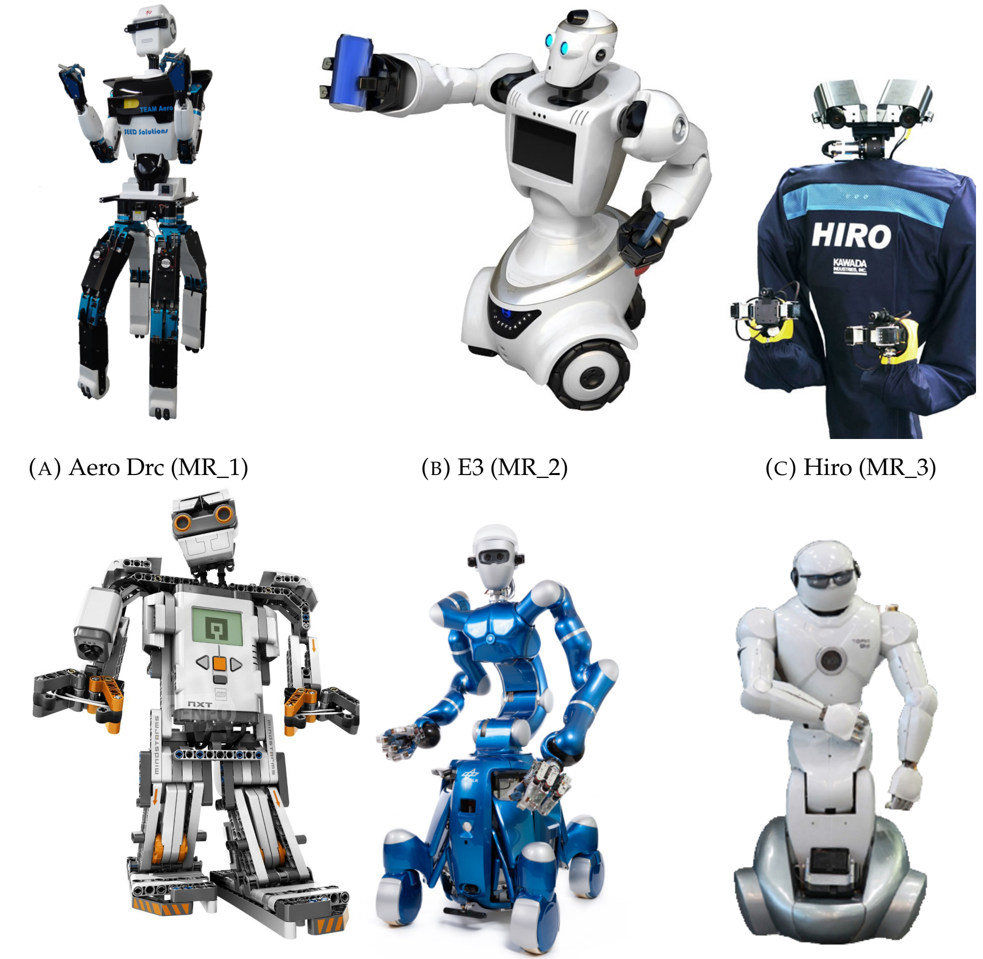
(E) Robina (FR_5)

(F) Sanbot (FR_6)

FIGURE B.1: The feminine robots selected for the pre-test

Appendix C

Masculine Robot Images



(A) Aero Drc (MR_1)

(B) E3 (MR_2)

(C) Hiro (MR_3)

(D) Lego Mindstorms Nxt 2.0 (MR_4)

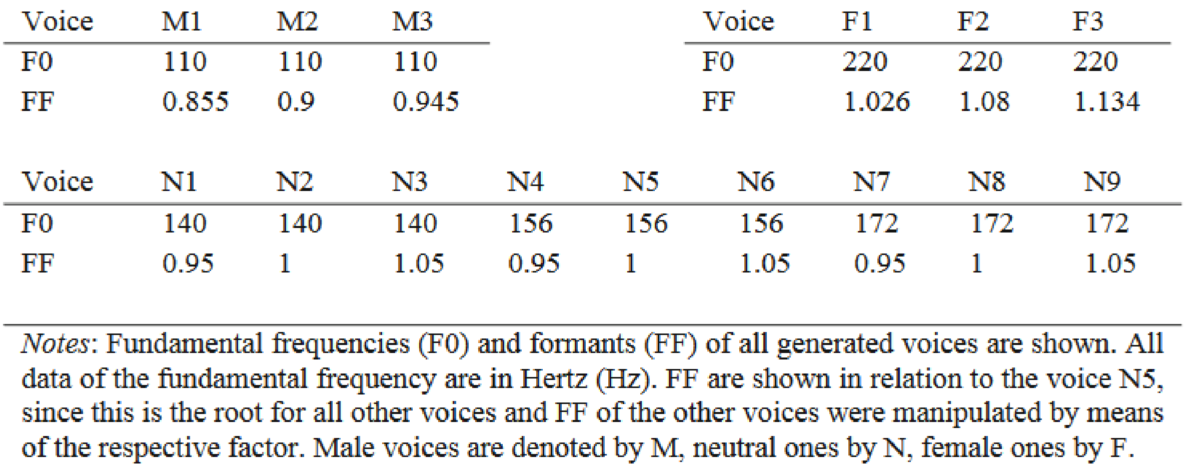
(E) Rollin Justin (MR_5)

(F) Topio Dio (MR_6)

FIGURE C.1: The masculine robots selected for the pre-test

Appendix D

Voice Fundamental & Formant Frequencies



Voice	M1	M2	M3
F0	110	110	110
FF	0.855	0.9	0.945

Voice	F1	F2	F3
F0	220	220	220
FF	1.026	1.08	1.134

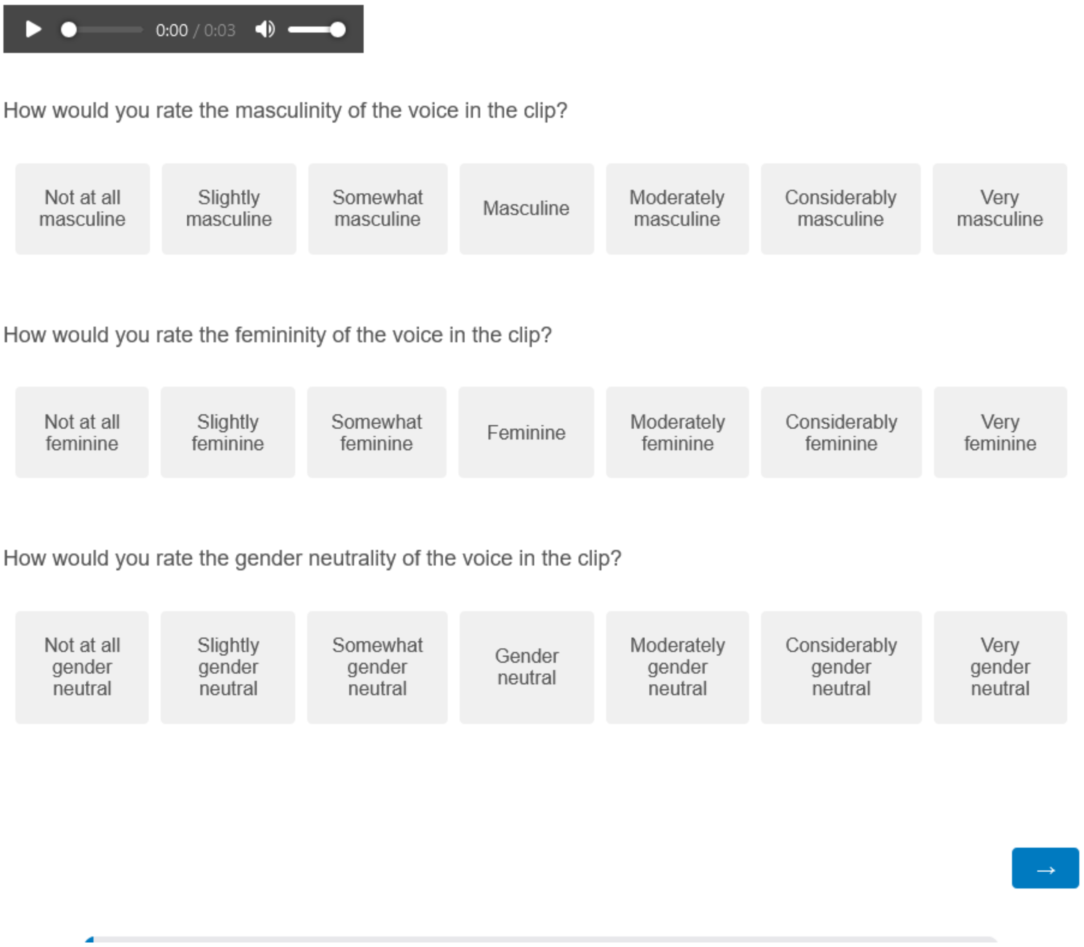
Voice	N1	N2	N3	N4	N5	N6	N7	N8	N9
F0	140	140	140	156	156	156	172	172	172
FF	0.95	1	1.05	0.95	1	1.05	0.95	1	1.05

Notes: Fundamental frequencies (F0) and formants (FF) of all generated voices are shown. All data of the fundamental frequency are in Hertz (Hz). FF are shown in relation to the voice N5, since this is the root for all other voices and FF of the other voices were manipulated by means of the respective factor. Male voices are denoted by M, neutral ones by N, female ones by F.

FIGURE D.1: The fundamental and formant frequency parameters calculated by Mooshammer & Etzrodt [100]

Appendix E

Pre-Test Survey: Voice Excerpt



How would you rate the masculinity of the voice in the clip?

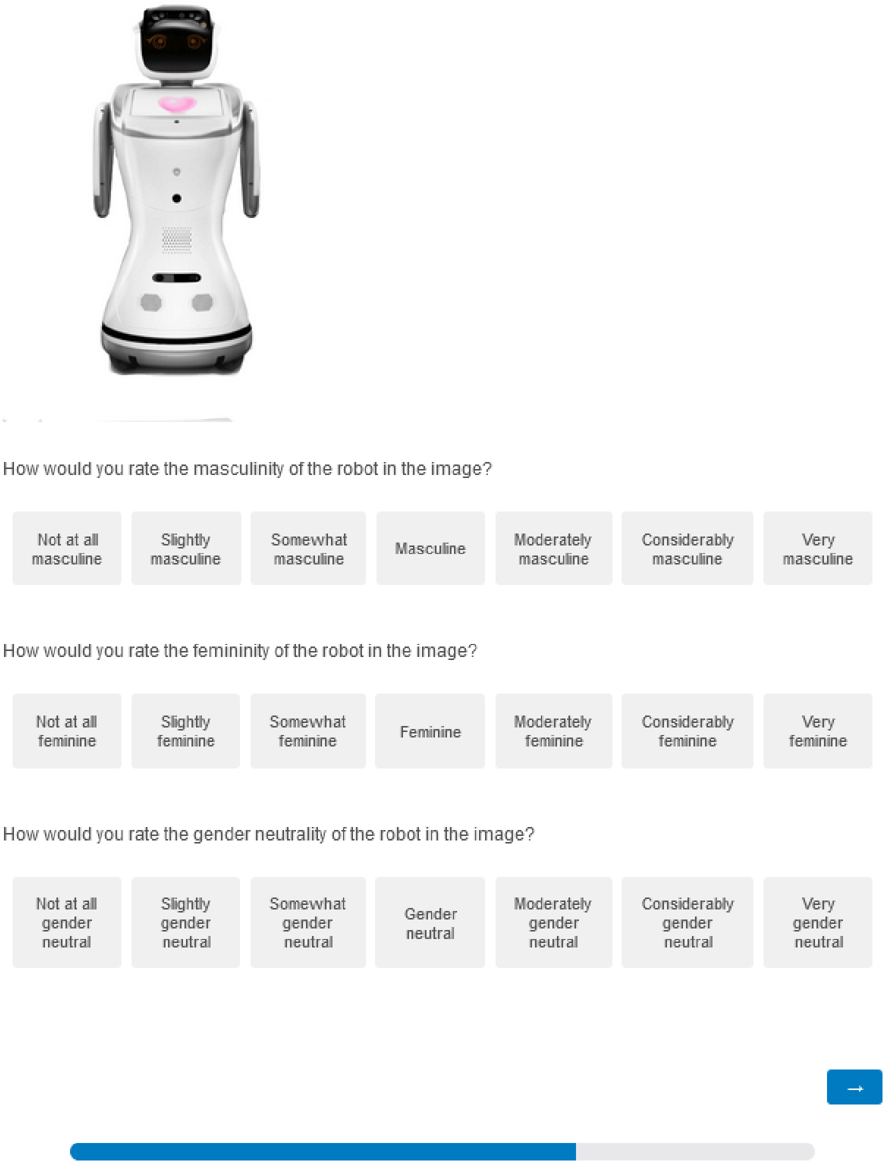
How would you rate the femininity of the voice in the clip?

How would you rate the gender neutrality of the voice in the clip?

FIGURE E.1: An excerpt from the pre-test survey, showing the media player with a voice clip at the top accompanied by the three gender perception questions.

Appendix F

Pre-Test Survey: Robot Excerpt



How would you rate the masculinity of the robot in the image?

Not at all masculine	Slightly masculine	Somewhat masculine	Masculine	Moderately masculine	Considerably masculine	Very masculine

How would you rate the femininity of the robot in the image?

Not at all feminine	Slightly feminine	Somewhat feminine	Feminine	Moderately feminine	Considerably feminine	Very feminine

How would you rate the gender neutrality of the robot in the image?

Not at all gender neutral	Slightly gender neutral	Somewhat gender neutral	Gender neutral	Moderately gender neutral	Considerably gender neutral	Very gender neutral

FIGURE F.1: An excerpt from the pre-test survey, showing the robot image at the top accompanied by the three gender perception questions.

Appendix G

Pre-Test: Robot Gender Perception Results

TABLE G.1: The gender perception results of the robots in the pre-test

					Final Gender Score	
Robot	Gender	Average	SD	AbsDiff	Masc	Fem
AR_1	Masc	1.65	0.933302	0.8	2.45	3.25
	Fem	2.45	1.877148			
	Neut	5	1.91943			
AR_2	Masc	3.4	1.500877	1.05	4.45	3.4
	Fem	2.35	1.348488			
	Neut	4.15	1.7252			
AR_3	Masc	2.55	1.276302	0.4	2.95	2.55
	Fem	2.15	1.598519			
	Neut	4.4	1.984148			
AR_4	Masc	2.1	1.209611	0.2	2.3	2.5
	Fem	2.3	1.592747			
	Neut	5.1	1.97084			
AR_5	Masc	3.35	1.631112	0.25	3.6	3.35
	Fem	3.1	1.48324			
	Neut	3.35	1.424411			
AR_6	Masc	2.35	1.386969	0.05	2.4	2.45
	Fem	2.4	1.698296			
	Neut	4.95	1.700619			
MR_1	Masc	4.2	1.704483	2.55	6.75	4.2
	Fem	1.65	1.03999			
	Neut	2.9	1.860956			
MR_2	Masc	4.55	1.700619	2.8	7.35	4.55
	Fem	1.75	1.371707			
	Neut	2.55	1.637553			
MR_3	Masc	4	1.863782	2.65	6.65	4
	Fem	1.35	0.587143			
	Neut	3.6	1.875044			
MR_4	Masc	4.6	1.846761	3.2	7.8	4.6
	Fem	1.4	0.502625			
	Neut	2.9	1.803505			
MR_5	Masc	4.7	1.260743	3	7.7	4.7
	Fem	1.7	1.128576			
	Neut	2.75	1.681947			

Table G.1 continued from previous page

					Final Gender Score	
Robot	**Gender**	**Average**	**SD**	**AbsDiff**	**Masc**	**Fem**
	Masc	5.95	1.234376			
MR_6	Fem	1.5	1.192079	4.45	10.4	5.95
	Neut	1.65	0.933302			
	Masc	1.3	0.656947			
FR_1	Fem	6.05	1.050063	4.75	6.05	10.8
	Neut	1.35	0.587143			
	Masc	1.5	0.888523			
FR_2	Fem	3.4	1.984148	1.9	3.4	5.3
	Neut	2.9	1.68273			
	Masc	1.65	0.875094			
FR_3	Fem	4.45	1.877148	2.8	4.45	7.25
	Neut	2.85	1.663066			
	Masc	1.7	0.801315			
FR_4	Fem	4.75	1.371707	3.05	4.75	7.8
	Neut	2.55	1.145931			
	Masc	1.35	0.812728			
FR_5	Fem	5.45	1.468081	4.1	5.45	9.55
	Neut	1.65	0.74516			
	Masc	1.45	0.604805			
FR_6	Fem	5.55	1.468081	4.1	5.55	9.65
	Neut	2.1	1.209611			

Appendix H

Pre-Test: Voice Gender Perception Results

TABLE H.1: The gender perception results of the voices in the pre-test

					Final Gender Score	
Voice	**Gender**	**Mean**	**SD**	**AbsDiff**	**Masc**	**Fem**
F1	Masc	2.75	1.164158	1.05	3.8	4.85
	Fem	3.8	1.823819			
	Neut	3	1.337712			
F2	Masc	2.2	1.105013	2	4.2	6.2
	Fem	4.2	1.43637			
	Neut	3.7	2.226633			
F3	Masc	2.05	1.145931	2.65	4.7	7.35
	Fem	4.7	1.525226			
	Neut	2.45	1.503505			
M1	Masc	6	1.256562	4.85	10.85	6
	Fem	1.15	0.48936			
	Neut	1.25	0.786398			
M2	Masc	6.1	1.071153	5.05	11.15	6.1
	Fem	1.05	0.223607			
	Neut	1.25	0.638666			
M3	Masc	5.85	1.089423	4.75	10.6	5.85
	Fem	1.1	0.447214			
	Neut	1.55	1.145931			
N1	Masc	5.65	1.03999	4.4	10.05	5.65
	Fem	1.25	0.444262			
	Neut	1.55	0.998683			
N2	Masc	5.2	1.196486	3.9	9.1	5.2
	Fem	1.3	0.571241			
	Neut	1.8	1.105013			
N3	Masc	4.45	1.356272	2.65	7.1	4.45
	Fem	1.8	0.615587			
	Neut	2.15	1.03999			
N4	Masc	5.4	1.095445	4.1	9.5	5.4
	Fem	1.3	0.470162			
	Neut	1.6	0.940325			
N5	Masc	5.3	0.978721	3.95	9.25	5.3
	Fem	1.35	0.48936			
	Neut	1.65	0.875094			

Table H.1 continued from previous page

					Final Gender Score	
Voice	**Gender**	**Mean**	**SD**	**AbsDiff**	**Masc**	**Fem**
	Masc	4.4	1.273206			
N6	Fem	1.95	0.759155	2.45	6.85	4.4
	Neut	2.6	1.313893			
	Masc	5.05	1.145931			
N7	Fem	1.65	0.74516	3.4	8.45	5.05
	Neut	2.4	1.391705			
	Masc	4	1.256562			
N8	Fem	2.35	1.136708	1.65	5.65	4
	Neut	3.05	1.316894			
	Masc	4	1.337712			
N9	Fem	2.35	0.812728	1.65	5.65	4
	Neut	3	1.450953			
	Masc	1.9	0.718185			
sam_n1	Fem	4.4	1.46539	2.5	4.4	6.9
	Neut	3.1	1.552587			
	Masc	1.2	0.410391			
sam_f1	Fem	5.55	1.276302	4.35	5.55	9.9
	Neut	2.3	0.864505			
	Masc	1.25	0.444262			
sam_f2	Fem	5.55	1.356272	4.3	5.55	9.85
	Neut	1.9	1.020836			
	Masc	1.15	0.366348			
sam_f3	Fem	5.1	1.586124	3.95	5.1	9.05
	Neut	2.05	1.316894			
	Masc	5.15	1.598519			
sam_m1	Fem	1.45	0.825578	3.7	8.85	5.15
	Neut	2.4	1.46539			
	Masc	4.15	1.694418			
sam_m2	Fem	1.8	1.151658	2.35	6.5	4.15
	Neut	2.7	1.525226			
	Masc	4.15	1.598519			
sam_m3	Fem	1.85	1.089423	2.3	6.45	4.15
	Neut	3.25	1.743409			

Appendix I

Robot Height Indication

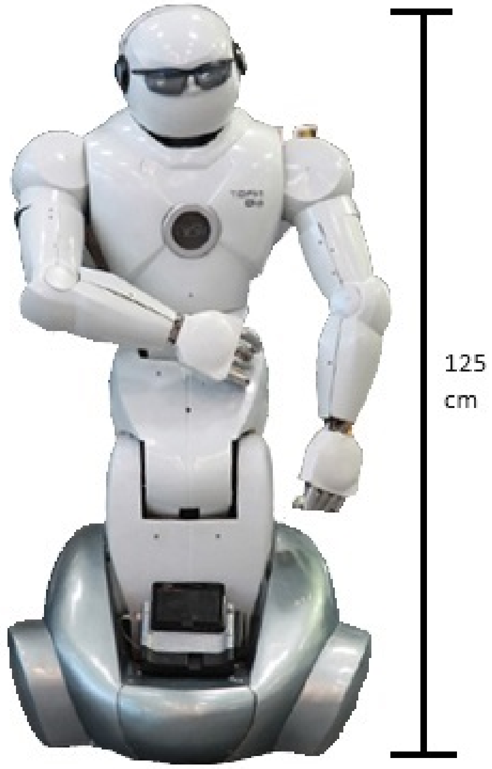
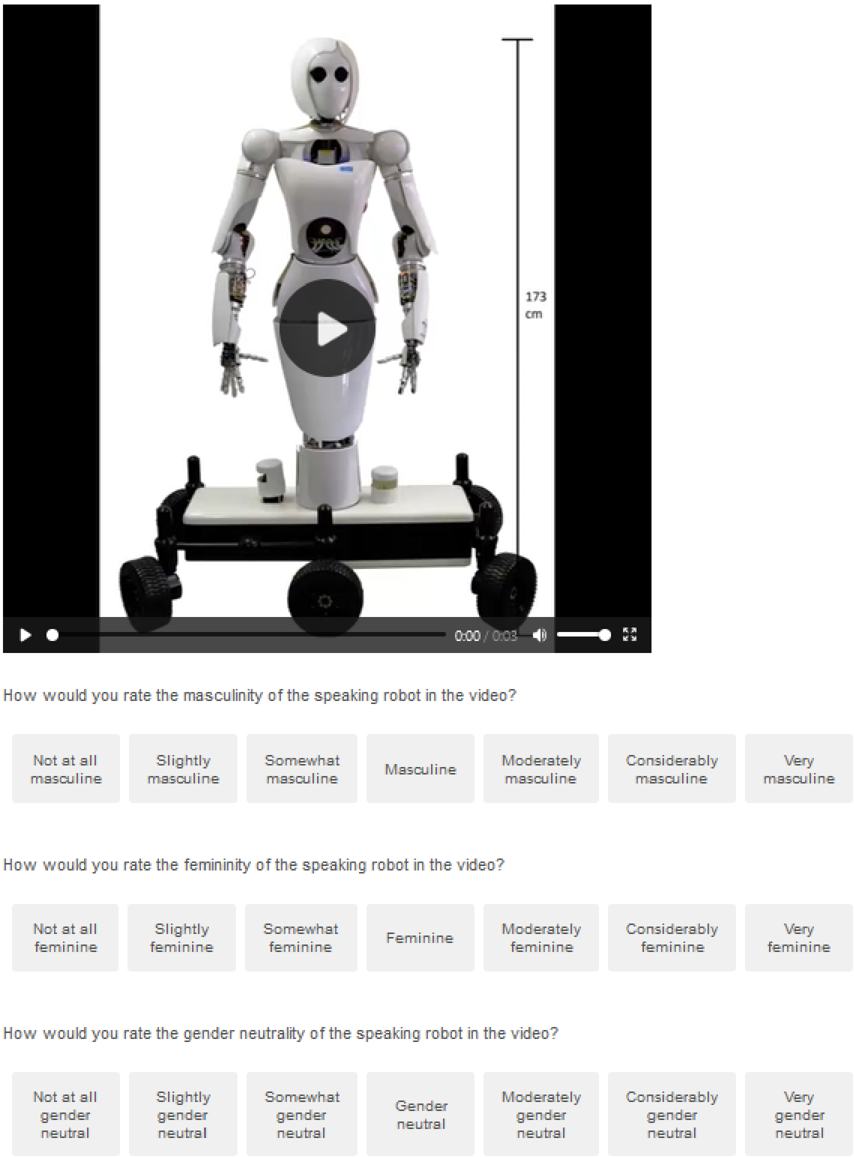


FIGURE I.1: The image of robot MR_6 (Topio Dio) with the height indication added for the main experiment. For all other robots used in the main experiment, a height indication was added in the exact same manner.

Appendix J

Main Experiment Survey: Combination Excerpt



How would you rate the masculinity of the speaking robot in the video?

Not at all masculine	Slightly masculine	Somewhat masculine	Masculine	Moderately masculine	Considerably masculine	Very masculine

How would you rate the femininity of the speaking robot in the video?

Not at all feminine	Slightly feminine	Somewhat feminine	Feminine	Moderately feminine	Considerably feminine	Very feminine

How would you rate the gender neutrality of the speaking robot in the video?

Not at all gender neutral	Slightly gender neutral	Somewhat gender neutral	Gender neutral	Moderately gender neutral	Considerably gender neutral	Very gender neutral

FIGURE J.1: An excerpt from the main experiment survey, showing the robot-voice combination video at the top accompanied by the three gender perception questions.

Appendix K

The Harmony Robot

The Harmony robot (see Figure K.1) was added to the main experiment survey as part of a different, separate research project. The robot was made to transport biomedical samples between labs in a hospital environment; survey participants were constantly reminded of this fact with every question. Participants were asked for their perception of the gender of the robot and the robot-voice combinations, in the same manner as for the rest of the survey. Differing from the main survey body, participants were also asked to rate the compatibility of the Harmony robot and the voice, on a 7-point scale for every robot-voice combination (see Figure K.2). Furthermore, once all the stimuli had been presented, the participants were also asked to provide the reasoning behind their answers to the gender perception questions, as well as to the compatibility question. None of the results from these questions are analysed here, as they are part of separate research entirely. The gender perception results are also not included in this research, even though it is formatted in the same way, as the Harmony robot was not selected from the ROBO-GAP database and through the pre-test, which means it was not controlled for human-likeness which the other robots were.

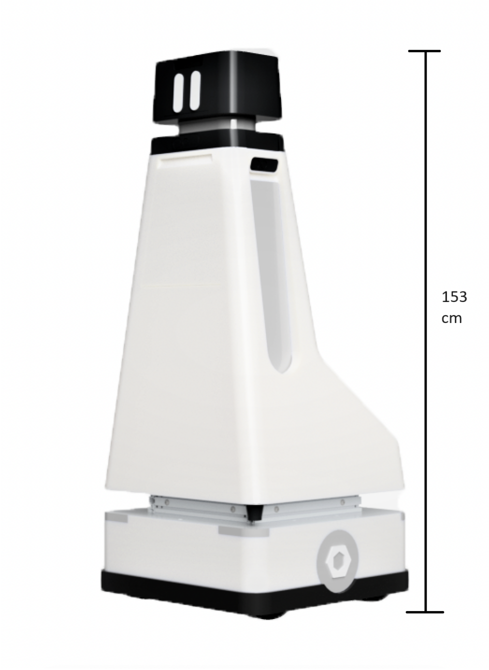
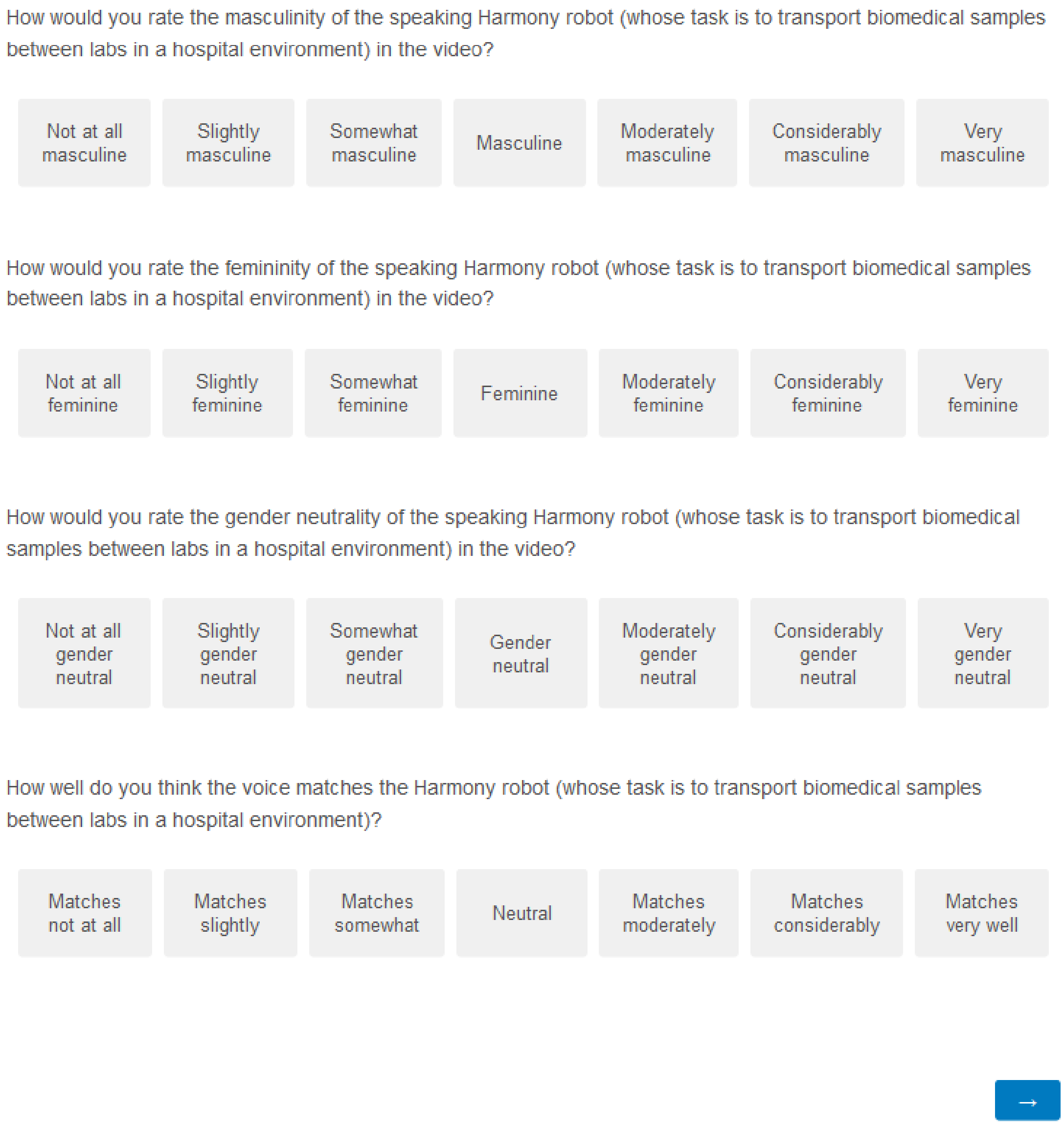


FIGURE K.1: The Harmony robot as presented in the survey



How would you rate the masculinity of the speaking Harmony robot (whose task is to transport biomedical samples between labs in a hospital environment) in the video?

Not at all masculine	Slightly masculine	Somewhat masculine	Masculine	Moderately masculine	Considerably masculine	Very masculine

How would you rate the femininity of the speaking Harmony robot (whose task is to transport biomedical samples between labs in a hospital environment) in the video?

Not at all feminine	Slightly feminine	Somewhat feminine	Feminine	Moderately feminine	Considerably feminine	Very feminine

How would you rate the gender neutrality of the speaking Harmony robot (whose task is to transport biomedical samples between labs in a hospital environment) in the video?

Not at all gender neutral	Slightly gender neutral	Somewhat gender neutral	Gender neutral	Moderately gender neutral	Considerably gender neutral	Very gender neutral

How well do you think the voice matches the Harmony robot (whose task is to transport biomedical samples between labs in a hospital environment)?

Matches not at all	Matches slightly	Matches somewhat	Neutral	Matches moderately	Matches considerably	Matches very well

FIGURE K.2: The questions asked below a video of a Harmony robot-voice combination

Appendix L

Main Experiment: Isolated Gender Perception Results

TABLE L.1: The isolated gender perception results of the voices in the main-experiment

					Final Gender Score	
Voice	Gender	Mean	SD	AbsDiff	Masc	Fem
F1 (*A*)	Masc	2.736842	1.329178	0.631579	3.368421	4
	Fem	3.368421	1.53202			
	Neut	3.184211	1.690278			
N8 (*A*)	Masc	3.368421	1.364037	0.921053	4.289474	3.368421
	Fem	2.447368	1.288145			
	Neut	2.789474	1.358813			
M1 (*M*)	Masc	5.789474	1.211608	4.710526	10.5	5.789474
	Fem	1.078947	0.358795			
	Neut	1.315789	0.739074			
M2 (*M*)	Masc	5.736842	1.031509	4.605263	10.34211	5.736842
	Fem	1.131579	0.34257			
	Neut	1.342105	0.708112			
sam_f1 (*F*)	Masc	1.184211	0.4565	4.078947	5.263158	9.342105
	Fem	5.263158	1.308686			
	Neut	1.736842	1.031509			
sam_f2 (*F*)	Masc	1.315789	0.873182	4	5.315789	9.315789
	Fem	5.315789	1.453891			
	Neut	1.631579	0.997864			

TABLE L.2: The isolated gender perception results of the robots in the main-experiment

					Final Gender Score	
Robot	**Gender**	**Mean**	**SD**	**AbsDiff**	**Masc**	**Fem**
AR_4 (*A*)	Masc	2.184211	1.159106	0.1316	2.3158	2.1842
	Fem	2.052632	1.29338			
	Neut	4.763158	1.69951			
AR_6 (*A*)	Masc	2.342105	1.457067	0.2632	2.6053	2.3421
	Fem	2.078947	1.421487			
	Neut	4.552632	2.036125			
MR_4 (*M*)	Masc	4.289474	1.784436	2.9737	7.2632	4.2895
	Fem	1.315789	0.619732			
	Neut	2.842105	1.636184			
MR_6 (*M*)	Masc	6.078947	1.075063	4.8947	10.974	6.0789
	Fem	1.184211	0.4565			
	Neut	1.5	1.006734			
FR_1 (*F*)	Masc	1.263158	0.554306	4.7895	6.0526	10.842
	Fem	6.052632	1.272312			
	Neut	1.421053	0.792927			
FR_6 (*F*)	Masc	1.263158	0.50319	3.5789	4.8421	8.4211
	Fem	4.842105	1.619581			
	Neut	2.157895	1.197437			

Appendix M

Main Experiment: Combination Gender Perception Results

TABLE M.1: The gender perception results of the robot-voice combinations. The colours of the robots and the voices represent their individually perceived gender. The colours in the last column represent the perceived gender of the robot-voice combinations. Blue represents masculinity, pink represents femininity, green represents ambiguity.

							Final Gender Score		
Combo	**Robot**	**Voice**	**Gender**	**Average**	**SD**	**AbsDiff**	**Masc**	**Fem**	**Gender Category**
a1	MR_6	M1	Masc	6.052632	1.161252	4.9474	11	6.0526	Masc
			Fem	1.105263	0.388307				
			Neut	1.394737	0.789782				
a2	MR_4	M1	Masc	5.868421	1.277054	4.8158	10.684	5.8684	Masc
			Fem	1.052632	0.226294				
			Neut	1.552632	1.107648				
a3	FR_1	M1	Masc	3.789474	1.613422	0.6842	4.4737	3.7895	Amb
			Fem	3.105263	1.942268				
			Neut	2.421053	1.445058				
a4	FR_6	M1	Masc	4.552632	1.639007	2.2105	6.7632	4.5526	Masc
			Fem	2.342105	1.236305				
			Neut	2.342105	1.438398				
a5	AR_6	M1	Masc	5.421053	1.286764	4.2368	9.6579	5.4211	Masc
			Fem	1.184211	0.392859				
			Neut	1.763158	1.195356				
a6	AR_4	M1	Masc	5.368421	1.383709	4.0789	9.4474	5.3684	Masc
			Fem	1.289474	0.515065				
			Neut	2.184211	1.522005				
b1	MR_6	M2	Masc	5.973684	1.077706	4.7895	10.763	5.9737	Masc
			Fem	1.184211	0.4565				
			Neut	1.631579	1.30324				
b2	MR_4	M2	Masc	5.710526	1.206019	4.6316	10.342	5.7105	Masc
			Fem	1.078947	0.273276				
			Neut	1.657895	1.168883				
b3	FR_1	M2	Masc	4.026316	1.635532	1.2105	5.2368	4.0263	Masc(Amb)
			Fem	2.815789	1.768421				
			Neut	2.263158	1.464613				
b4	FR_6	M2	Masc	4.315789	1.23256	2.0789	6.3947	4.3158	Masc
			Fem	2.236842	1.050977				
			Neut	2.368421	1.422238				
b5	AR_6	M2	Masc	5.552632	1.26699	4.2105	9.7632	5.5526	Masc
			Fem	1.342105	0.627148				
			Neut	1.894737	1.203362				

Table M.1 continued from previous page

Combo	Robot	Voice	Gender	Average	SD	AbsDiff	Final Gender Score		Gender Category
							Masc	Fem	
b6	AR_4	M2	Masc	5.263158	1.464613	3.9474	9.2105	5.2632	Masc
			Fem	1.315789	0.574469				
			Neut	1.868421	1.234001				
c1	MR_6	sam_f1	Masc	3.026316	1.762781	0.3684	3.3947	3.7632	Amb
			Fem	3.394737	1.263617				
			Neut	2.421053	1.407155				
c2	MR_4	sam_f1	Masc	2.5	1.502251	0.8947	3.3947	4.2895	Amb
			Fem	3.394737	1.284828				
			Neut	3.078947	1.666548				
c3	FR_1	sam_f1	Masc	1.210526	0.474079	4.2632	5.4737	9.7368	Fem
			Fem	5.473684	1.428226				
			Neut	1.815789	1.204839				
c4	FR_6	sam_f1	Masc	1.421053	0.948158	3.7105	5.1316	8.8421	Fem
			Fem	5.131579	1.509807				
			Neut	1.842105	1.000711				
c5	AR_6	sam_f1	Masc	1.815789	0.925765	2.4211	4.2368	6.6579	Fem
			Fem	4.236842	1.364298				
			Neut	2.763158	1.69951				
c6	AR_4	sam_f1	Masc	1.789474	0.963044	2.1579	3.9474	6.1053	Fem
			Fem	3.947368	1.659491				
			Neut	3.157895	1.938602				
d1	MR_6	sam_f2	Masc	3.105263	1.885786	0.5263	3.6316	4.1579	Amb
			Fem	3.631579	1.459749				
			Neut	2.605263	1.498458				
d2	MR_4	sam_f2	Masc	2.5	1.502251	1.0526	3.5526	4.6053	Fem(Amb)
			Fem	3.552632	1.349622				
			Neut	2.526316	1.428226				
d3	FR_1	sam_f2	Masc	1.289474	0.767865	4.2632	5.5526	9.8158	Fem
			Fem	5.552632	1.408418				
			Neut	1.789474	1.37856				
d4	FR_6	sam_f2	Masc	1.236842	0.489578	4.2105	5.4474	9.6579	Fem
			Fem	5.447368	1.408418				
			Neut	1.789474	1.069425				
d5	AR_6	sam_f2	Masc	1.684211	0.77478	2.3421	4.0263	6.3684	Fem
			Fem	4.026316	1.497509				
			Neut	3.052632	1.52364				
d6	AR_4	sam_f2	Masc	1.447368	0.724004	2.7105	4.1579	6.8684	Fem
			Fem	4.157895	1.405132				
			Neut	2.921053	1.791596				
e1	MR_6	F1	Masc	3.789474	1.742286	1.6842	5.4737	3.7895	Masc
			Fem	2.105263	1.007793				
			Neut	2.710526	1.505089				
e2	MR_4	F1	Masc	3.236842	1.683532	0.8947	4.1316	3.2368	Amb
			Fem	2.342105	1.27928				
			Neut	3.368421	1.822102				
e3	FR_1	F1	Masc	2.105263	1.157571	2.0789	4.1842	6.2632	Fem
			Fem	4.184211	1.783638				
			Neut	2.710526	1.468735				
e4	FR_6	F1	Masc	2.131579	1.255712	1.5789	3.7105	5.2895	Fem
			Fem	3.710526	1.69112				
			Neut	3.131579	1.509807				
e5	AR_6	F1	Masc	2.789474	1.473329	0.4211	3.2105	2.7895	Amb
			Fem	2.368421	1.566905				
			Neut	3.921053	1.745753				
e6	AR_4	F1	Masc	2.578947	1.65434	0.0263	2.6053	2.5789	Amb
			Fem	2.552632	1.605689				
			Neut	3.789474	1.862255				

Table M.1 continued from previous page

							Final Gender Score		
Combo	**Robot**	**Voice**	**Gender**	**Average**	**SD**	**AbsDiff**	**Masc**	**Fem**	**Gender Category**
f1	MR_6	N8	Masc	4.368421	1.667188	2.6579	7.0263	4.3684	Masc
			Fem	1.710526	0.9273				
			Neut	2.657895	1.529464				
f2	MR_4	N8	Masc	3.473684	1.447026	1.6579	5.1316	3.4737	Masc
			Fem	1.815789	0.865409				
			Neut	3.394737	1.603029				
f3	FR_1	N8	Masc	2.552632	1.201291	1.0526	3.6053	4.6579	Fem(Amb)
			Fem	3.605263	1.808979				
			Neut	2.921053	1.583386				
f4	FR_6	N8	Masc	2.894737	1.390887	0.3421	3.2368	3.5789	Amb
			Fem	3.236842	1.618044				
			Neut	3.052632	1.558713				
f5	AR_6	N8	Masc	3.394737	1.534107	1.4211	4.8158	3.3947	Masc(Amb)
			Fem	1.973684	1.262491				
			Neut	3.236842	1.851723				
f6	AR_4	N8	Masc	3.105263	1.590334	1.0526	4.1579	3.1053	Masc(Amb)
			Fem	2.052632	1.393952				
			Neut	3.631579	1.86607				

Appendix N

Main Experiment: Masculine Base Results

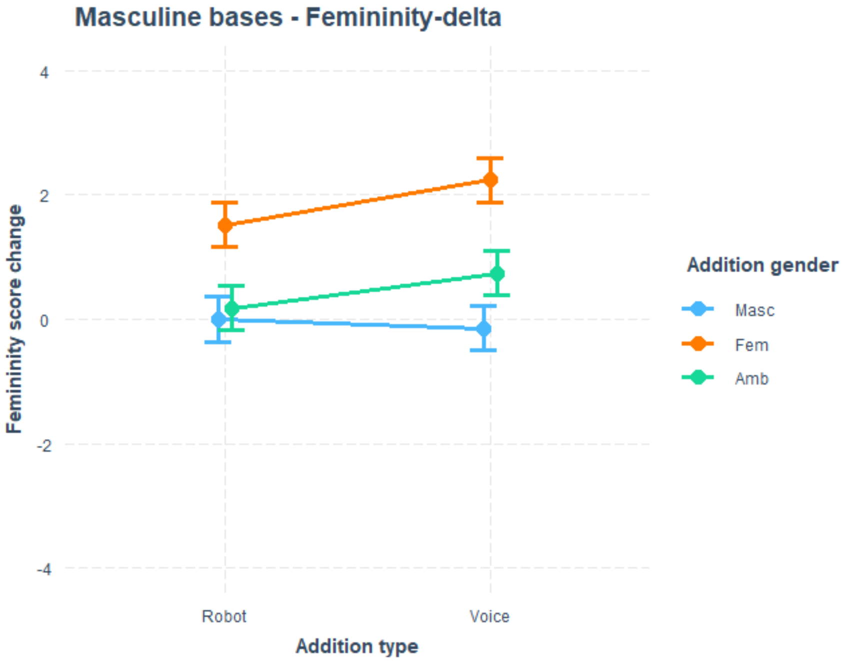


FIGURE N.1: Interaction plot with 95% confidence intervals of the femininity-delta model of the masculine bases, showing the effect of the addition type and addition gender on the femininity-delta

TABLE N.1: The linear mixed model results for the femininity-delta of masculine bases

Fixed Effect	Estimate	Standard Error	95% Confidence Interval	p-value
Intercept	2.40E-15	0.183	[-0.2876987, 0.2876992]	1
addtype_V	-0.1447	0.2444	[-0.5212654, 0.2317924]	0.575429
addgen_A	0.1776	0.2436	[-0.1987583, 0.5540216]	0.493292
addgen_F	1.52	0.2436	[1.1433470, 1.8961269]	**0.000785**
addtype_V:addgen_A	0.7105	0.3445	[0.1782307, 1.2428222]	0.084729
addtype_V:addgen_F	0.8684	0.3445	[0.3361254, 1.4007170]	**0.045214**

Random Effect	Standard Deviation		
participant	0.36983		
base_id	0.02051		
add_id	0.21712	Conditional R^2	0.503
Residual	0.9622	Marginal R^2	0.404

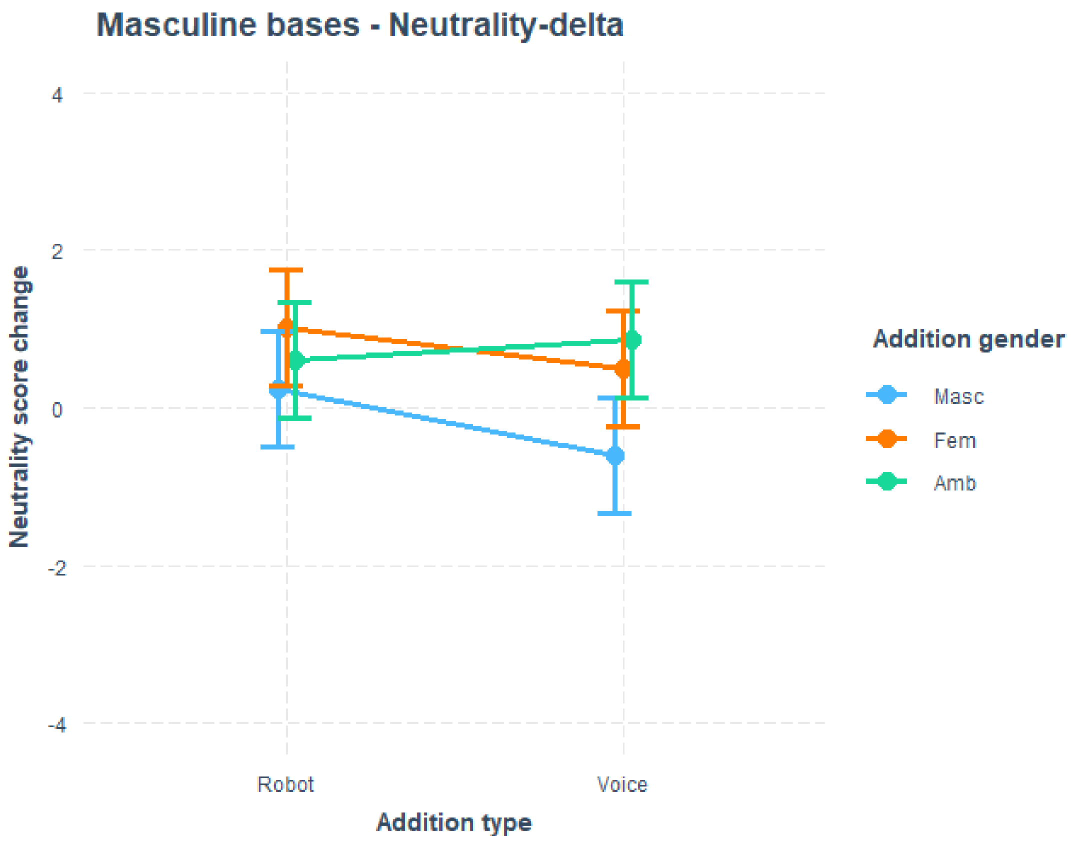


FIGURE N.2: Interaction plot with 95% confidence intervals of the neutrality-delta model of the masculine bases, showing the effect of the addition type and addition gender on the neutrality-delta

TABLE N.2: The linear mixed model results for the neutrality-delta of masculine bases

Fixed Effect	Estimate	Standard Error	95% Confidence Interval	p-value
Intercept	0.2303	0.3741	[-0.45398556, 0.91451369]	0.5864
addtype_V	-0.8421	0.5065	[-1.77818481, 0.09397714]	0.2239
addgen_A	0.3684	0.1504	[0.07389233, 0.66294978]	**0.0145**
addgen_F	0.7895	0.1504	[0.49494496, 1.08400241]	**1.94E-07**
addtype_V:addgen_A	1.1053	0.2128	[0.68873664, 1.52178967]	**1.94E-07**
addtype_V:addgen_F	0.3092	0.2128	[-0.10731599, 0.72573704]	0.1465

Random Effect	Standard Deviation		
participant	0.6656		
base_id	0.4836		
add_id	-	Conditional R^2	0.358
Residual	1.3116	Marginal R^2	0.105

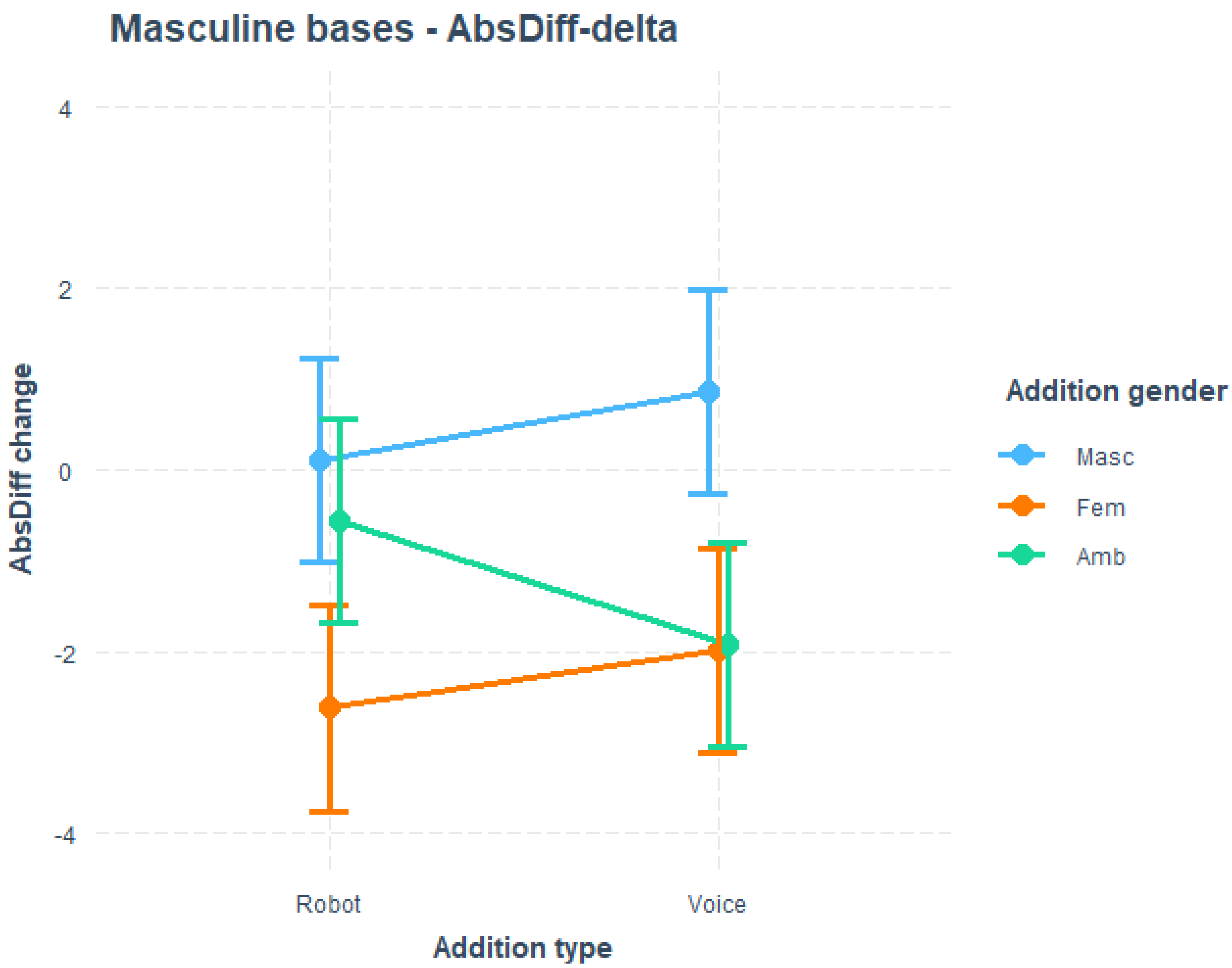


FIGURE N.3: Interaction plot with 95% confidence intervals of the AbsDiff-delta model of the masculine bases, showing the effect of the addition type and addition gender on the AbsDiff-delta

TABLE N.3: The linear mixed model results for the AbsDiff-delta of masculine bases

Fixed Effect	Estimate	Standard Error	95% Confidence Interval	p-value
Intercept	0.1118	0.5753	[-0.9378506 , 1.1615379]	0.860593
addtype_V	0.75	0.7864	[-0.6967098, 2.1967143]	0.433845
addgen_A	-0.6776	0.1956	[-1.0605227, -0.2947405]	**0.000557**
addgen_F	-2.7368	0.1956	[-3.1197332, -2.353951]	**< 2E-16**
addtype_V:addgen_A	-2.1053	0.2766	[-2.6467529, -1.5637734]	**7.08E-14**
addtype_V:addgen_F	-0.1118	0.2766	[-0.6533319, 0.4296477]	0.68606

Random Effect	Standard Deviation			
participant	0.9083			
base_id	0.7617			
add_id	-		Conditional R^2	0.505
Residual	1.7051		Marginal R^2	0.266

Appendix O

Main Experiment: Feminine Base Results

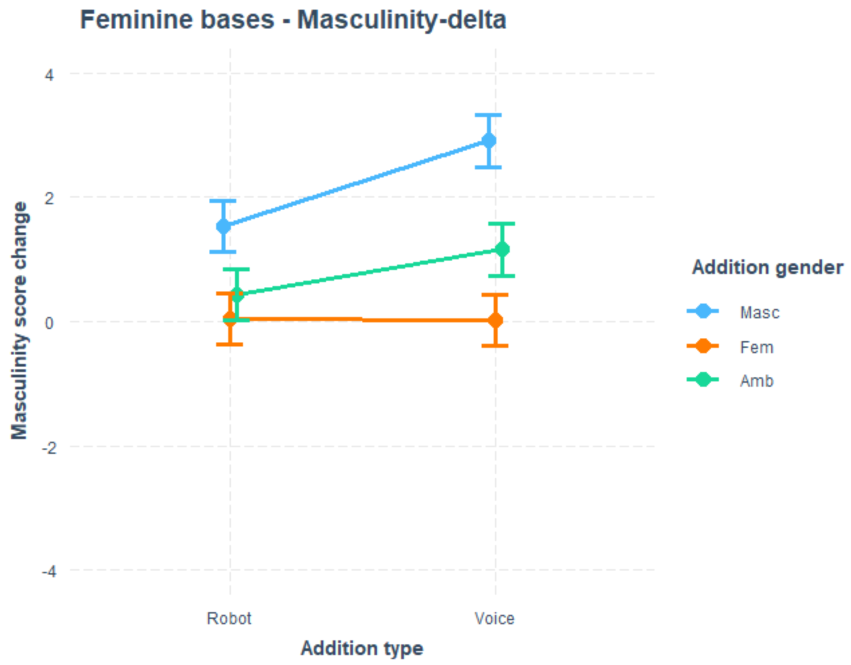


FIGURE O.1: Interaction plot with 95% confidence intervals of the masculinity-delta model of the feminine bases, showing the effect of the addition type and addition gender on the masculinity-delta

TABLE O.1: The linear mixed model results for the masculinity-delta of feminine bases

Fixed Effect	Estimate	Standard Error	95% Confidence Interval	p-value
Intercept	0.03947	0.21209	[-0.30273007, 0.3816789]	0.856968
addtype_V	-0.0132	0.28748	[-0.47203186, 0.4457173]	0.964809
addgen_A	0.39474	0.24371	[0.00533412, 0.7841396]	0.156428
addgen_M	1.49342	0.24371	[1.10401833, 1.8828238]	**0.000863**
addtype_V:addgen_A	0.73684	0.34466	[0.18614349, 1.2875407]	0.076375
addtype_V:addgen_M	1.38816	0.34466	[0.83745928, 1.9388565]	**0.006899**

Random Effect	Standard Deviation		
participant	0.3731		
base_id	0.1525		
add_id	0.2063	Conditional R^2	0.49
Residual	1.1316	Marginal R^2	0.408

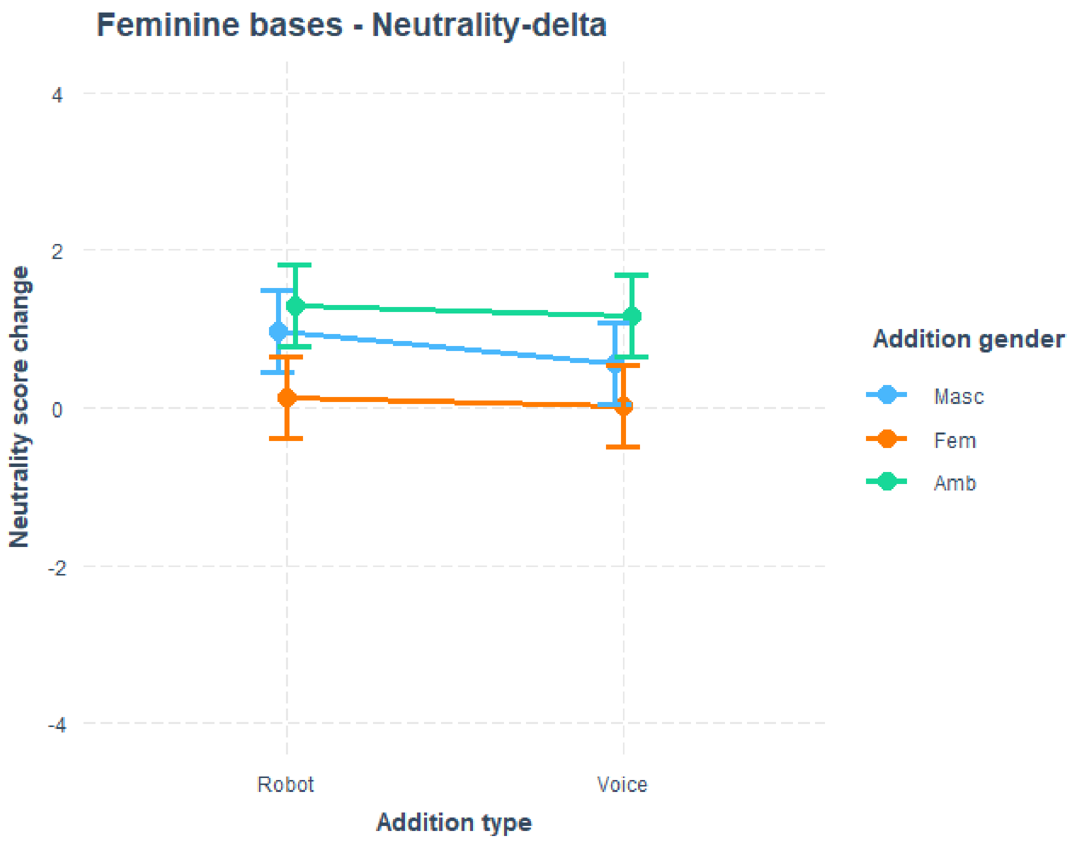


FIGURE O.2: Interaction plot with 95% confidence intervals of the neutrality-delta model of the feminine bases, showing the effect of the addition type and addition gender on the neutrality-delta

TABLE O.2: The linear mixed model results for the neutrality-delta of feminine bases

Fixed Effect	Estimate	Standard Error	95% Confidence Interval	p-value
Intercept	0.125	0.26446	[-0.3631092, 0.6131097]	0.662
addtype_V	-0.1053	0.34154	[-0.7484091, 0.5378846]	0.781
addgen_A	1.16447	0.15073	[0.8693949, 1.4595524]	**3.07E-14**
addgen_M	0.84868	0.15073	[0.5536055, 1.1437630]	**2.43E-08**
addtype_V:addgen_A	-0.0197	0.21316	[0.4370412, 0.3975675]	0.926
addtype_V:addgen_M	-0.3092	0.21316	[-0.7265149, 0.1080938]	0.147

Random Effect	Standard Deviation			
participant	0.6643			
base_id	0.3065			
add_id	-		Conditional R^2	0.31
Residual	1.314		Marginal R^2	0.097

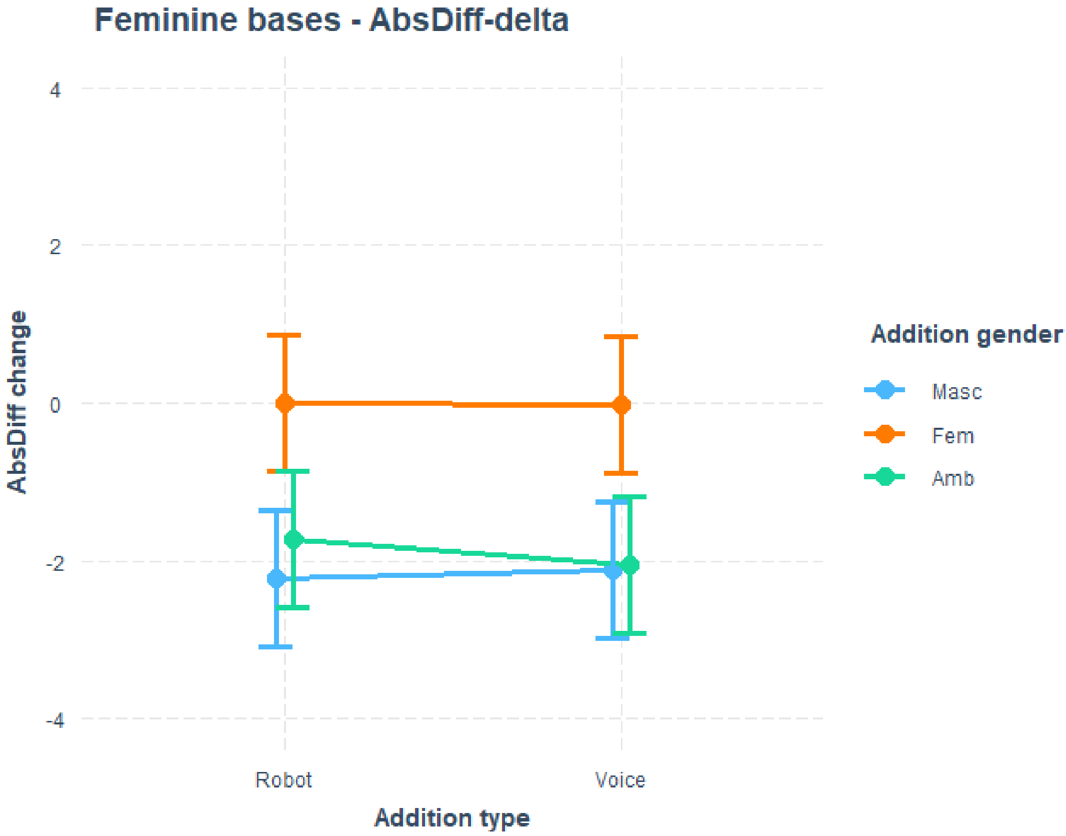


FIGURE O.3: Interaction plot with 95% confidence intervals of the AbsDiff-delta model of the feminine bases, showing the effect of the addition type and addition gender on the AbsDiff-delta

TABLE O.3: The linear mixed model results for the AbsDiff-delta of feminine bases

Fixed Effect	Estimate	Standard Error	95% Confidence Interval	p-value
Intercept	-0.0066	0.441693	[-0.8171712, 0.8040150]	0.989
addtype_V	-0.0132	0.588831	[-1.1085310, 1.0822184]	0.984
addgen_A	-1.7171	0.214718	[-2.1374551, -1.2967554]	**4.06E-15**
addgen_M	-2.2171	0.214718	[-2.6374551, -1.7967554]	**< 2E-16**
addtype_V:addgen_A	-0.3158	0.303657	[-0.9102539 , 0.2786750]	0.299
addtype_V:addgen_M	0.11184	0.303657	[-0.4826224 , 0.7063066]	0.713

Random Effect	Standard Deviation			
participant	0.9087			
base_id	0.5483			
add_id	-		Conditional R^2	0.37
Residual	1.8719		Marginal R^2	0.167

Appendix P

Main Experiment: Ambiguous Base Results

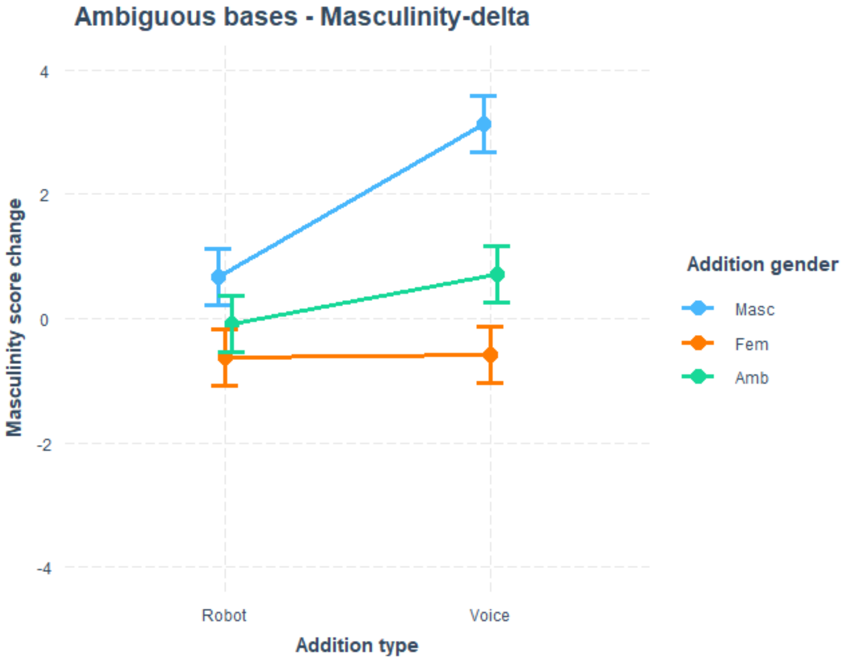


FIGURE P.1: Interaction plot with 95% confidence intervals of the masculinity-delta model of the ambiguous bases, showing the effect of the addition type and addition gender on the masculinity-delta

TABLE P.1: The linear mixed model results for the masculinity-delta of ambiguous bases

Fixed Effect	Estimate	Standard Error	95% Confidence Interval	p-value
Intercept	-0.0855	0.23043	[-0.4599738, 0.28892119]	0.71855
addtype_V	0.78947	0.28817	[0.3398508, 1.23909665]	**0.03376**
addgen_F	-0.5461	0.28817	[-0.9956755, -0.09642966]	0.10693
addgen_M	0.75	0.28817	[0.3003771, 1.19962297]	**0.04052**
addtype_V:addgen_F	-0.7368	0.40754	[-1.3727049, -0.10097921]	0.12061
addtype_V:addgen_M	1.68421	0.40754	[1.0483477, 2.32007342]	**0.00613**

Random Effect	Standard Deviation		
participant	0.6633		
base_id	-		
add_id	0.2324	Conditional R^2	0.491
Residual	1.486	Marginal R^2	0.377

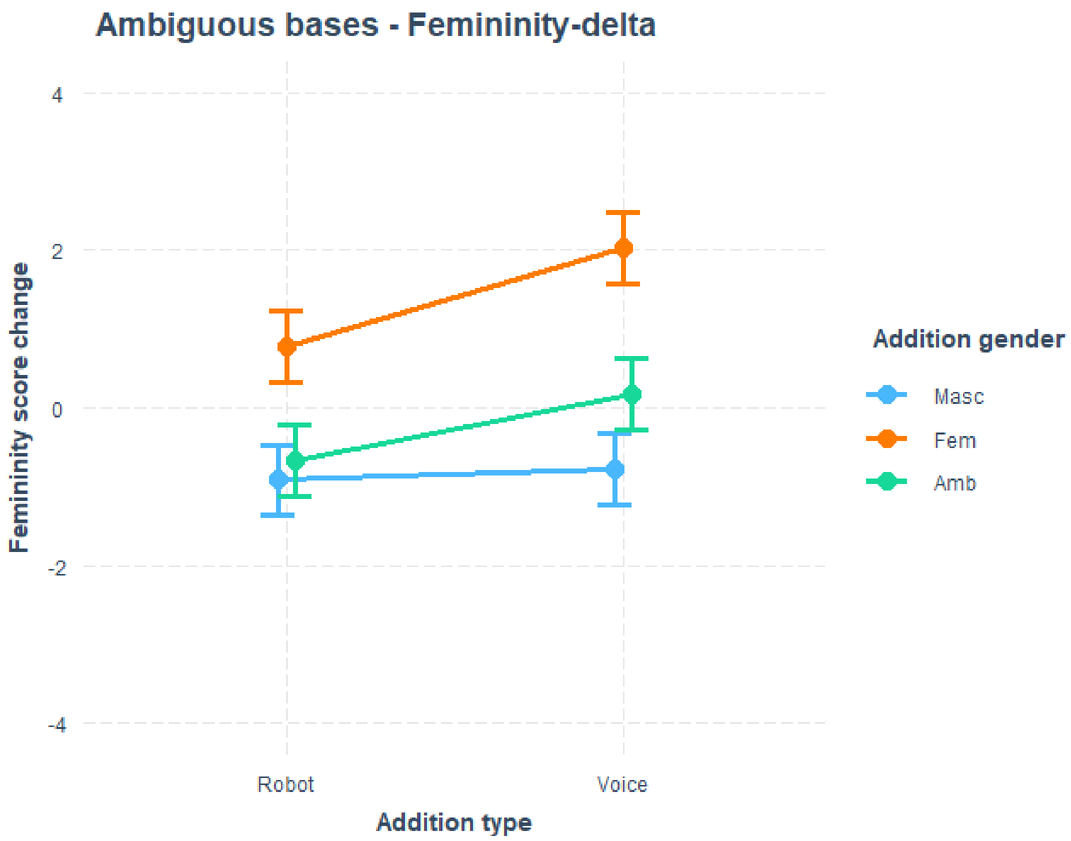


FIGURE P.2: Interaction plot with 95% confidence intervals of the femininity-delta model of the ambiguous bases, showing the effect of the addition type and addition gender on the femininity-delta

TABLE P.2: The linear mixed model results for the femininity-delta of ambiguous bases

Fixed Effect	Estimate	Standard Error	95% Confidence Interval	p-value
Intercept	-0.6711	0.2289	[-1.0877361, -0.2543711]	**0.020802**
addtype_V	0.8421	0.2743	[0.33435400, 1.3498580]	**0.038887**
addgen_F	1.4474	0.1898	[1.09652847 , 1.7982084]	**0.000266**
addgen_M	-0.2434	0.1898	[-0.59426101, 0.1074189]	0.247063
addtype_V:addgen_F	0.4079	0.2685	[-0.08826789, 0.9040574]	0.17949
addtype_V:addgen_M	-0.7105	0.2685	[-1.20668895, -0.2143637]	**0.038208**

Random Effect	Standard Deviation		
participant	0.74938		
base_id	0.19803		
add_id	0.06361	Conditional R^2	0.412
Residual	1.55925	Marginal R^2	0.265

Bibliography

[1] M. West, R. Kraut, and H. E. Chew, "I'd blush if I could: Closing gender divides in digital skills through education," Tech. Rep., Jan. 2019. DOI: 10.54675/rapc9356.

[2] H. Walk, *Amazon Echo is magical. it's also turning my kid into an asshole.* Apr. 2016. [Online]. Available: https://www.linkedin.com/pulse/amazon-echo-magical-its-also-turning-my-kid-asshole-hunter-walk.

[3] F. Eyssel and F. Hegel, "(S)he's got the look: Gender stereotyping of robots," *Journal of Applied Social Psychology*, vol. 42, no. 9, pp. 2213–2230, 2012.

[4] F. Laricchia, *Number of voice assistants in use worldwide 2019-2024*, Apr. 2020. [Online]. Available: https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/.

[5] Mordor Intelligence, *Social robots market size & share analysis*. [Online]. Available: https://www.mordorintelligence.com/industry-reports/social-robots-market.

[6] K. Markopoulos, G. Maniati, G. Vamvoukakis, *et al.*, "Generating gender-ambiguous text-to-speech voices," *arXiv preprint arXiv:2211.00375*, 2022.

[7] G. Perugia, S. Guidi, M. Bicchi, and O. Parlangeli, "The shape of our bias: Perceived age and gender in the humanoid robots of the abot database," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2022, pp. 110–119.

[8] M. Latinus, R. VanRullen, and M. J. Taylor, "Top-down and bottom-up modulation in processing bimodal face/voice stimuli," *BMC neuroscience*, vol. 11, no. 1, pp. 1–13, 2010.

[9] World Health Organization. [Online]. Available: https://www.who.int/health-topics/gender#tab=tab_1.

[10] Merriam-Webster. [Online]. Available: https://www.merriam-webster.com/grammar/sex-vs-gender-how-they2019re-different.

[11] A. Cuncic, *What is gender identity?* Jul. 2021. [Online]. Available: https://www.verywellmind.com/what-is-gender-identity-5187156.

[12] S. J. Sutton, "Gender ambiguous, not genderless: Designing gender in voice user interfaces (VUIs) with sensitivity," in *Proceedings of the 2nd conference on conversational user interfaces*, 2020, pp. 1–8.

[13] Merriam-Webster. [Online]. Available: https://www.merriam-webster.com/dictionary/ambiguous.

[14] J. W. Mullennix, K. A. Johnson, M. Topcu-Durgun, and L. M. Farnsworth, "The perceptual representation of voice gender," *The Journal of the Acoustical Society of America*, vol. 98, no. 6, pp. 3080–3095, 1995.

[15] A. Danielescu, S. A. Horowit-Hendler, A. Pabst, K. M. Stewart, E. M. Gallo, and M. P. Aylett, "Creating inclusive voices for the 21st century: A non-binary text-to-speech for conversational assistants," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–17.

[16] Merriam-Webster. [Online]. Available: https://www.merriam-webster.com/dictionary/androgyny.

[17] I. Leite, C. Martinho, and A. Paiva, "Social robots for long-term interaction: A survey," *International Journal of Social Robotics*, vol. 5, pp. 291–308, 2013.

[18] A. Henschel, G. Laban, and E. S. Cross, "What makes a robot social? a review of social robots from science fiction to a home or hospital near you," *Current Robotics Reports*, vol. 2, pp. 9–19, 2021.

[19] F. Hegel, C. Muhl, B. Wrede, M. Hielscher-Fastabend, and G. Sagerer, "Understanding social robots," in *2009 Second International Conferences on Advances in Computer-Human Interactions*, IEEE, 2009, pp. 169–174.

[20] C. Breazeal, "Toward sociable robots," *Robotics and autonomous systems*, vol. 42, no. 3-4, pp. 167–175, 2003.

[21] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and autonomous systems*, vol. 42, no. 3-4, pp. 143–166, 2003.

[22] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: A three-factor theory of anthropomorphism.," *Psychological review*, vol. 114, no. 4, p. 864, 2007.

[23] R. Kühne and J. Peter, "Anthropomorphism in human–robot interactions: A multidimensional conceptualization," *Communication Theory*, vol. 33, no. 1, pp. 42–52, 2023.

[24] A. E. Martin and M. F. Mason, "Hey Siri, I love you: People feel more attached to gendered technology," *Journal of Experimental Social Psychology*, vol. 104, p. 104 402, 2023.

[25] A. E. Martin and M. F. Mason, "What does it mean to be (seen as) human? The importance of gender in humanization.," *Journal of Personality and Social Psychology*, 2022.

[26] R. Frazer, "Experimental operationalizations of anthropomorphism in HCI contexts: A scoping review," *Communication Reports*, vol. 35, no. 3, pp. 173–189, 2022.

[27] T. Zhang, D. B. Kaber, B. Zhu, M. Swangnetr, P. Mosaly, and L. Hodge, "Service robot feature design effects on user perceptions and emotional responses," *Intelligent service robotics*, vol. 3, pp. 73–88, 2010.

[28] K. L. Nowak and J. Fox, "Avatars and computer-mediated communication: A review of the definitions, uses, and effects of digital representations," *Review of Communication Research*, vol. 6, pp. 30–53, 2018.

[29] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *Journal of experimental social psychology*, vol. 52, pp. 113–117, 2014.

[30] P. Aggarwal and A. L. McGill, "Is that car smiling at me? Schema congruity as a basis for evaluating anthropomorphized products," *Journal of consumer research*, vol. 34, no. 4, pp. 468–479, 2007.

[31] K.-P. Tam, S.-L. Lee, and M. M. Chao, "Saving Mr. Nature: Anthropomorphism enhances connectedness to and protectiveness toward nature," *Journal of Experimental Social Psychology*, vol. 49, no. 3, pp. 514–521, 2013.

[32] C. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '94, Boston, Massachusetts, USA: Association for Computing Machinery, 1994, pp. 72–78, ISBN: 0897916506. DOI: 10.1145/191666.191703.

[33] A. Clodic, E. Pacherie, R. Alami, and R. Chatila, "Key elements for human-robot joint action," in *Robophilosophy*, 2014, pp. 159–177.

[34] B. R. Duffy, "Anthropomorphism and the social robot," *Robotics and autonomous systems*, vol. 42, no. 3-4, pp. 177–190, 2003.

[35] C. McGarty, "Social categorization," *Oxford Research Encyclopedia of Psychology*, Mar. 2018. DOI: 10.1093/acrefore/9780190236557.013.308. [Online]. Available: https://doi.org/10.1093/acrefore/9780190236557.013.308.

[36] F. Eyssel and D. Kuchenbrandt, "Social categorization of social robots: Anthropomorphism as a function of robot group membership," *British Journal of Social Psychology*, vol. 51, no. 4, pp. 724–731, 2012.

[37] E. Roesler, M. Heuring, and L. Onnasch, "(Hu)man-like robots: The impact of anthropomorphism and language on perceived robot gender," *International Journal of Social Robotics*, pp. 1–12, 2023.

[38] E. L. Haines, K. Deaux, and N. Lofaro, "The times they are a-changing... or are they not? A comparison of gender stereotypes, 1983–2014," *Psychology of Women Quarterly*, vol. 40, no. 3, pp. 353–363, 2016.

[39] S. Guidi, L. Boor, L. van der Bij, R. Foppen, O. Rikmenspoel, and G. Perugia, "Ambivalent stereotypes towards gendered robots: The (im)mutability of bias towards female and neutral robots," in *International Conference on Social Robotics*, Springer, 2022, pp. 615–626.

[40] S. I. Behrens, A. K. K. Egsvang, M. Hansen, and A. M. Møllegård-Schroll, "Gendered robot voices and their influence on trust," in *Companion of the 2018 ACM/IEEE international conference on human-robot interaction*, 2018, pp. 63–64.

[41] A. Galatolo, G. I. Melsión, I. Leite, and K. Winkle, "The right (wo)man for the job? Exploring the role of gender when challenging gender stereotypes with a social robot," *International Journal of Social Robotics*, pp. 1–15, 2022.

[42] T. Law, M. Chita-Tegmark, and M. Scheutz, "The interplay between emotional intelligence, trust, and gender in human–robot interaction: A vignette-based study," *International Journal of Social Robotics*, vol. 13, no. 2, pp. 297–309, 2021.

[43] N. Reich-Stiebert and F. Eyssel, "(Ir)relevance of gender? On the influence of gender stereotypes on learning with a robot," in *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, 2017, pp. 166–176.

[44] M. Siegel, C. Breazeal, and M. I. Norton, "Persuasive robotics: The influence of robot gender on human behavior," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2009, pp. 2563–2568.

[45] S. Song, J. Baba, J. Nakanishi, Y. Yoshikawa, and H. Ishiguro, "Mind the voice!: Effect of robot voice pitch, robot voice gender, and user gender on user perception of teleoperated robots," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–8.

[46] B. Tay, Y. Jung, and T. Park, "When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction," *Computers in Human Behavior*, vol. 38, pp. 75–84, 2014.

[47] A. Wong, A. Xu, and G. Dudek, "Investigating trust factors in human-robot shared control: Implicit gender bias around robot voice," in *2019 16th Conference on Computer and Robot Vision (CRV)*, IEEE, 2019, pp. 195–200.

[48] D. Bryant, J. Borenstein, and A. Howard, "Why should we gender? The effect of robot gendering and occupational stereotypes on human trust and perceived competency," in *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, 2020, pp. 13–21.

[49] D. J. Rea, Y. Wang, and J. E. Young, "Check your stereotypes at the door: An analysis of gender typecasts in social human-robot interaction," in *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7*, Springer, 2015, pp. 554–563.

[50] P. Belin, P. E. G. Bestelmeyer, M. Latinus, and R. Watson, "Understanding voice perception," *British Journal of Psychology*, vol. 102, no. 4, pp. 711–725, Jun. 2011. DOI: 10.1111/j.2044-8295.2011.02041.x.

[51] P. McAleer, A. Todorov, and P. Belin, "How do you say 'hello'? Personality impressions from brief novel voices," *PloS one*, vol. 9, no. 3, e90779, 2014.

[52] M. S. Tsantani, P. Belin, H. M. Paterson, and P. McAleer, "Low vocal pitch preference drives first impressions irrespective of context in male voices but not in female voices," *Perception*, vol. 45, no. 8, pp. 946–963, 2016.

[53] E. J. Lee, C. Nass, and S. Brave, "Can computer-generated speech have gender? An experimental test of gender stereotype," in *CHI'00 extended abstracts on Human factors in computing systems*, 2000, pp. 289–290.

[54] C. Nass, Y. Moon, and N. Green, "Are machines gender neutral? Gender-stereotypic responses to computers with voices," *Journal of applied social psychology*, vol. 27, no. 10, pp. 864–876, 1997.

[55] L. Z. McArthur and R. M. Baron, "Toward an ecological theory of social perception.," *Psychological Review*, vol. 90, no. 3, pp. 215–238, Jul. 1983. DOI: 10.1037/0033-295x.90.3.215.

[56] L. A. Zebrowitz and M. A. Collins, "Accurate social perception at zero acquaintance: The affordances of a gibsonian approach," *Personality and Social Psychology Review*, vol. 1, no. 3, pp. 204–223, Aug. 1997. DOI: 10.1207/s15327957pspr0103_2.

[57] J. Rhim, Y. Kim, M.-S. Kim, and D. Y. Yim, "The effect of gender cue alterations of robot to match task attributes on user's acceptance perception," HCIK '15, pp. 51–57, 2014.

[58] G. Hwang, J. Lee, C. Y. Oh, and J. Lee, "It sounds like a woman: Exploring gender stereotypes in South Korean voice assistants," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–6.

[59] J. A. Bargh, "The case against the controllability of automatic stereotype effects," *Dual-process theories in social psychology*, p. 361, 1999.

[60] J. Fink, "Anthropomorphism and human likeness in the design of robots and human-robot interaction," in *Social Robotics*, Springer Berlin Heidelberg, 2012, pp. 199–208. DOI: 10.1007/978-3-642-34103-8_20.

[61] W. B. Mendes, J. Blascovich, S. B. Hunter, B. Lickel, and J. T. Jost, "Threatened by the unexpected: Physiological responses during social interactions with expectancy-violating partners.," *Journal of personality and social psychology*, vol. 92, no. 4, p. 698, 2007.

[62] E. Gustavsson, "Virtual servants: Stereotyping female front-office employees on the internet," *Gender, Work and Organization*, vol. 12, no. 5, pp. 400–419, Sep. 2005. DOI: 10.1111/j.1468-0432.2005.00281.x.

[63] A. E. Martin and M. L. Slepian, "The primacy of gender: Gendered cognition underlies the big two dimensions of social cognition," *Perspectives on Psychological Science*, vol. 16, no. 6, pp. 1143–1158, 2021.

[64] O. Parlangeli, P. Palmitesta, M. Bracci, E. Marchigiani, and S. Guidi, "Gender role stereotypes at work in humanoid robots," *Behaviour & Information Technology*, pp. 1–12, 2022.

[65] M. L. Walters, D. S. Syrdal, K. L. Koay, K. Dautenhahn, and R. Te Boekhorst, "Human approach distances to a mechanical-looking robot with different robot voice styles," in *RO-MAN 2008-The 17th IEEE international symposium on robot and human interactive communication*, IEEE, 2008, pp. 707–712.

[66] E. Phillips, X. Zhao, D. Ullman, and B. F. Malle, "What is human-like? Decomposing robots' human-like appearance using the anthropomorphic robot (ABOT) database," in *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, 2018, pp. 105–113.

[67] G. Trovato, C. Lucho, and R. Paredes, "She's electric—the influence of body proportions on perceived gender of robots across cultures," *Robotics*, vol. 7, no. 3, p. 50, 2018.

[68] B. I. Fagot, M. D. Leinbach, B. E. Hort, and J. Strayer, "Qualities underlying the definitions of gender," *Sex Roles*, vol. 37, pp. 1–18, 1997.

[69] T. Lieven, B. Grohmann, A. Herrmann, J. R. Landwehr, and M. Van Tilburg, "The effect of brand design on brand gender perceptions and brand preference," *European Journal of Marketing*, vol. 49, no. 1/2, pp. 146–169, 2015.

[70] M. van Tilburg, T. Lieven, A. Herrmann, and C. Townsend, "Beyond "pink it and shrink it" perceived product gender, aesthetics, and product evaluation," *Psychology & Marketing*, vol. 32, no. 4, pp. 422–437, 2015.

[71] S. J. Cunningham and C. N. Macrae, "The colour of gender stereotyping," *British Journal of Psychology*, vol. 102, no. 3, pp. 598–614, 2011.

[72] A. C. Hess and V. Melnyk, "Pink or blue? The impact of gender cues on brand perceptions," *European Journal of Marketing*, vol. 50, no. 9/10, pp. 1550–1574, 2016.

[73] S. C. Steinhaeusser, P. Schaper, O. Bediako Akuffo, P. Friedrich, J. Ön, and B. Lugrin, "Anthropomorphize me! Effects of robot gender on listeners' perception of the social robot NAO in a storytelling use case," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 529–534.

[74] G. Aşkın, İ. Saltık, T. E. Boz, and B. A. Urgen, "Gendered actions with a genderless robot: Gender attribution to humanoid robots in action," *International Journal of Social Robotics*, pp. 1–17, 2023.

[75] D. Stanton, M. Shannon, S. Mariooryad, *et al.*, "Speaker generation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7897–7901. DOI: 10.1109/ICASSP43922.2022.9747345.

[76] P. N. Ladefoged, *Vowel formants*, 1998. [Online]. Available: https://www.britannica.com/science/phonetics/Vowel-formants.

[77] G. Reid, *Formant synthesis*, Mar. 2001. [Online]. Available: https://www.soundonsound.com/techniques/formant-synthesis.

[78] D. E. Re, J. J. O'Connor, P. J. Bennett, and D. R. Feinberg, "Preferences for very low and very high voice pitch in humans," *PloS one*, vol. 7, no. 3, e32719, 2012.

[79] M. Biemans, "The effect of biological gender (sex) and social gender (gender identity) on three pitch measures," *Linguistics in the Netherlands*, vol. 15, no. 1, pp. 41–52, 1998.

[80] A. P. Simpson, "Phonetic differences between male and female speech," *Language and linguistics compass*, vol. 3, no. 2, pp. 621–640, 2009.

[81] N. Nørgaard, *How to create a genderless voice*, Dec. 2019. [Online]. Available: https://www.youtube.com/watch?v=qH6KB7MrOPw.

[82] C. I. Nass and S. Brave, *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge, 2005.

[83] S. Tolmeijer, N. Zierau, A. Janson, J. S. Wahdatehagh, J. M. M. Leimeister, and A. Bernstein, "Female by default?–exploring the effect of voice assistant gender and pitch on trait and trust attribution," in *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–7.

[84] R. K. Moore, "Appropriate voices for artefacts: Some key insights," in *1st International workshop on vocal interactivity in-and-between humans, animals and robots*, 2017.

[85] R. K. Moore, "Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction," *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, pp. 281–291, 2017.

[86] M. P. Aylett, B. R. Cowan, and L. Clark, "Siri, echo and performance: You have to suffer darling," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–10.

[87] M. M. de Graaf, S. Ben Allouch, and J. A. Van Dijk, "What makes robots social?: A user's perspective on characteristics for social human-robot interaction," in *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7*, Springer, 2015, pp. 184–193.

[88] D. Dereshev, D. Kirk, K. Matsumura, and T. Maeda, "Long-term value of social robots through the eyes of expert users," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[89] C. Rivoire and A. Lim, "Habit detection within a long-term interaction with a social robot: An exploratory study," in *Proceedings of the International Workshop on Social Learning and Multimodal Interaction for Designing Artificial Agents*, 2016, pp. 1–6.

[90] W. Wannagat, L. Huestegge, E. Landmann, G. Nieding, and S. M. Huestegge, "Development of visual dominance in face-voice integration: Evidence from cross-modal compatibility effects in a gender categorization task," *Cognitive Development*, vol. 64, p. 101 263, 2022.

[91] S. M. Huestegge and T. Raettig, "Crossing gender borders: Bidirectional dynamic interaction between face-based and voice-based gender categorization," *Journal of voice*, vol. 34, no. 3, 487–e1, 2020.

[92] Z. F. Peynircioğlu, W. Brent, J. R. Tatz, and J. Wyatt, "McGurk effect in gender identification: Vision trumps audition in voice judgments," *The Journal of General Psychology*, vol. 144, no. 1, pp. 59–68, 2017.

[93] T. S. Andersen, K. Tiippana, and M. Sams, "Factors influencing audiovisual fission and fusion illusions," *Cognitive Brain Research*, vol. 21, no. 3, pp. 301–308, 2004.

[94] E. L. Smith, M. Grabowecky, and S. Suzuki, "Auditory-visual crossmodal integration in perception of face gender," *Current Biology*, vol. 17, no. 19, pp. 1680–1685, 2007.

[95] B. L. Turner and B. Christenson, "Alexa or Alex or neither? Exploring gender-neutral voices, gender framing and consumer judgments of synthetic voices," *ACR North American Advances*, 2020.

[96] M. Paetzel, C. Peters, I. Nyström, and G. Castellano, "Effects of multimodal cues on children's perception of uncanniness in a social robot," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 297–301.

[97] E. Mower, M. J. Mataric, and S. Narayanan, "Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information," *IEEE Transactions on Multimedia*, vol. 11, no. 5, pp. 843–855, 2009.

[98] L. MacLellan, *Hear what a genderless AI voice sounds like—and consider why it matters*, Mar. 2019. [Online]. Available: https://qz.com/work/1577597/this-ai-voice-is-gender-neutral-unlike-siri-and-alexa.

[99] T. E. Matheus, B. Suomiya, M. Soh, and Y. Toshimasa, "Designing gender ambiguous voice agents," *International Journal of Affective Engineering*, vol. 22, no. 1, pp. 53–62, 2023.

[100] S. Mooshammer and K. Etzrodt, "Social research with gender-neutral voices in chatbots–the generation and evaluation of artificial gender-neutral voices with praat and google wavenet," in *Chatbot Research and Design: 5th International Workshop, CONVERSATIONS 2021, Virtual Event, November 23–24, 2021, Revised Selected Papers*, Springer, 2022, pp. 176–191.

[101] S. L. Bem, "The measurement of psychological androgyny.," *Journal of consulting and clinical psychology*, vol. 42, no. 2, p. 155, 1974.

[102] E. Roesler, L. Naendrup-Poell, D. Manzey, and L. Onnasch, "Why context matters: The influence of application domain on preferred degree of anthropomorphism and gender attribution in human–robot interaction," *International Journal of Social Robotics*, vol. 14, no. 5, pp. 1155–1166, 2022.

[103] A. v. d. Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[104] *Standard, wavenet, neural2, and studio voices*, en. [Online]. Available: https://cloud.google.com/text-to-speech/docs/wavenet#studio_voices_preview.

[105] N. Chomsky, *Syntactic Structure*. Mouton, 1957.

[106] *Text-to-Speech AI: Lifelike Speech Synthesis | Google Cloud*. [Online]. Available: https://cloud.google.com/text-to-speech?hl=en.

[107] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer [Computer program]*, version 6.2.06, 1992-2022. [Online]. Available: www.praat.org.

[108] K. M. Baber and C. J. Tucker, "The social roles questionnaire: A new approach to measuring attitudes toward gender," *Sex Roles*, vol. 54, pp. 459–467, 2006.

[109] K. Spiel, O. L. Haimson, and D. Lottridge, "How to do better with gender on surveys: A guide for hci researchers," *Interactions*, vol. 26, no. 4, pp. 62–65, 2019.

[110] J. B. Freeman and N. Ambady, "When two become one: Temporally dynamic integration of the face and voice," *Journal of experimental social psychology*, vol. 47, no. 1, pp. 259–263, 2011.

[111] B. De Gelder and J. Vroomen, "The perception of emotions by ear and by eye," *Cognition & Emotion*, vol. 14, no. 3, pp. 289–311, 2000.

[112] S. Campanella and P. Belin, "Integrating face and voice in person perception," *Trends in cognitive sciences*, vol. 11, no. 12, pp. 535–543, 2007.

[113] T. R. Foundation, *R: The R Project for Statistical Computing*. [Online]. Available: https://www.r-project.org/.

[114] , "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015. DOI: 10.18637/jss.v067.i01.

[115] , "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017. DOI: 10.18637/jss.v082.i13".

[116] D. Lüdecke, M. S. Ben-Shachar, I. Patil, P. Waggoner, and D. Makowski, "performance: An R package for assessment, comparison and testing of statistical models," *Journal of Open Source Software*, vol. 6, no. 60, p. 3139, 2021. DOI: 10.21105/joss.03139.

[117] J. A. Long, *Interactions: Comprehensive, user-friendly toolkit for probing interactions*, R package version 1.2.0, 2024. DOI: 10.32614/CRAN.package.interactions. [Online]. Available: https://cran.r-project.org/package=interactions.

[118] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016, ISBN: 978-3-319-24277-4. [Online]. Available: https://ggplot2.tidyverse.org.