

Exploring Voice Agent Gender in a Running Coach Application

by

Casey O'Neill

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2025

© Casey O'Neill 2025

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In daily life, people commonly use voice agents such as Siri and Alexa to perform everyday tasks. With rapid technological advancements, voice agents are becoming more humanlike and there is a growing interest to use them in high-stakes tasks such as mental health advice and therapy. However, it is important to understand the potential harms posed by these voice agents of the future. This work focuses on gender-based social justice problems introduced by the widespread use of humanlike gendered voice agents, many of which disproportionately affect women due to the high prevalence of commercial voice agents that are female or “female by default”. To contribute to an understanding of how to design gendered voice agents that reduce these potential harms, we explore reactions to a gendered voice agent running coach. We build a voice agent running coach system for smartphones that includes three voice options (male, female, gender ambiguous) which we validate through an in-person survey study ($n = 30$). We use our system in a field study ($n = 18$) in which participants run with the agent for three weeks and attend two in-person or online sessions with the researcher. We present a statistical analysis of survey data and key themes from a reflexive thematic analysis of interview data from the study. We conclude with a discussion of actions designers can take to create gendered voice agents more responsibly. We also offer recommendations for future research into gender and gender-based stereotyping in voice agents.

Acknowledgements

I would first like to acknowledge the instructors and supervisors who shared their time and knowledge with me. Thanks to my supervisor, Dr. Edith Law, for allowing me the freedom to take on this meaningful work. Thanks to my readers, Dr. Dan Vogel and Dr. Ana Crisan, for your valuable and thoughtful feedback. Thanks to Dr. Scott Bateman for generously offering me a space in your lab when it was needed and to the folks there for making me feel welcome. Thanks to Dr. Daniel Rea for your enthusiasm and for providing me with the wonderful opportunity to share our work with the international research community.

I would also like to thank the friends who supported me through this endeavour. Thanks to my labmates and colleagues for your encouragement and guidance. I hope you enjoyed losing at trivia as much as I did. Thanks to my (boy)friends back home in New Brunswick, who have been there for me through it all. Thanks especially to Catherine and Christopher DeLenardo for your positivity, kindness, and hospitality.

Finally, thank you to my family: Kelly, Hugh, and Anika. Thanks for answering every phone call and for never once giving up on me. Without your love and dedication, this work would not have been possible.

Table of Contents

Author's Declaration	ii
Abstract	iii
Acknowledgements	iv
List of Figures	viii
List of Tables	x
1 Introduction	1
2 Background Literature	3
3 System	7
3.1 Voice Agent	7
3.1.1 Voice Design	8
3.1.2 Dialogue Design	8
3.2 Interface	9
3.2.1 Login and Home Screens	10
3.2.2 Run Configuration Screen	11
3.2.3 In-Progress Run Screen	12
3.2.4 Post-Run Survey Screen	13
3.3 Data Collection	14

4	Pilot Study	15
4.1	Method	15
4.2	Measures	16
4.3	Participants	16
4.4	Results	18
4.4.1	Perceived Gender	19
4.4.2	Trait Ascription	19
4.4.3	Godspeed	22
4.5	Discussion	23
5	Field Study	25
5.1	Method	25
5.2	Measures	27
5.3	Participants	27
5.4	Results	29
5.4.1	Quantitative	30
5.4.2	Qualitative	34
6	Discussion	39
6.1	Key Findings	40
6.2	Limitations and Future Work	41
7	Conclusion	44
	References	45
	APPENDICES	55
A	Pilot Study	56
A.1	Questionnaires	56
A.1.1	Demographics Questionnaire	56
A.1.2	Voice Agent Evaluation Questionnaire	59

B	Field Study	61
B.1	Questionnaires	61
B.1.1	Demographics Questionnaire	61
B.1.2	Post-Session Questionnaire	65
B.1.3	Voice Agent Evaluation Questionnaire	66
B.2	Interview Questions	67

List of Figures

3.1	Login and home screens and key features.	10
3.2	Run configuration screen and key features.	11
3.3	In-progress run screen and key features.	12
3.4	Post-run survey screen and key features.	13
4.1	Gender perception scores in each condition. Scores range from 1 (masculine) to 4 (ambiguous) to 7 (feminine).	19
4.2	Trait ascription scores for stereotypically male traits in each condition. Scores range from 1 (strongly disagree the agent has this trait) to 5 (strongly agree the agent has this trait).	20
4.3	Trait ascription scores for stereotypically female traits in each condition. Scores range from 1 (strongly disagree the agent has this trait) to 5 (strongly agree the agent has this trait).	21
4.4	Averages for Godspeed Questionnaire Series scores in each condition. Error bars indicate standard error.	22
5.1	Gender perception scores in each condition. Scores range from 1 (masculine) to 4 (ambiguous) to 7 (feminine).	30
5.2	Trait ascription scores for stereotypically male traits in each condition. Scores range from 1 (strongly disagree the agent has this trait) to 5 (strongly agree the agent has this trait).	31
5.3	Trait ascription scores for stereotypically female traits in each condition. Scores range from 1 (strongly disagree the agent has this trait) to 5 (strongly agree the agent has this trait).	32

5.4	Averages for Godspeed Questionnaire Series scores in each condition. Error bars indicate standard error.	33
-----	--	----

List of Tables

3.1	Statement categories and example statements used with our voice agent feedback template.	8
4.1	Google Cloud Text-to-Speech API voice configurations used in the pilot study.	15
4.2	Operational definitions of running experience levels.	17
4.3	Pilot study participant demographics.	18
5.1	Field study participant demographics.	29

Chapter 1

Introduction

Voice agents, often referred to as voice assistants, have become ubiquitous. In daily life, people use voice agents like Amazon’s Alexa and Apple’s Siri to perform everyday tasks, such as checking the forecast, controlling smart home devices, and scheduling appointments. With advancements in machine learning and voice-based technologies, voice agents are becoming more humanlike and are being used in increasingly high-stakes tasks that pose greater risks to users than the simple chores voice agents have traditionally been designed to perform. For example, some people are turning to voice chatbots like OpenAI’s ChatGPT for mental health advice and therapy [36, 72]. There also has been a recent trend in news media coverage of people seeking deeper emotional relationships with voice agents, such as friendships [75] and romantic partnerships [31]. As we progress towards these voice agents of the future, it is important to understand the potential harms such systems pose and how interaction designers can mitigate them. In this work, we aim to better understand the gender-based social justice problems posed by gendered voice agents.

Today’s leading commercial voice assistants are predominantly female or “female by default”. Even when male or gender ambiguous voice options are offered, most users continue to use a female voice [5]. In public and academic discourse, scholars have warned that the prevalence of female voice assistants in subservient roles can reinforce harmful gender stereotypes that have historically been used to justify the treatment of women as second-class citizens [81, 82, 58]. These concerns are echoed by research studies which find that gender-based stereotypes are frequently applied to voice agents [17, 18, 49] and that gendered technologies unable to navigate antisocial dynamics can exacerbate gender-based violence and harassment [28]. There is even evidence that some commercial voice agents are intentionally designed to conform to expectations of stereotypical gender roles [34, 25]. However, despite these growing concerns and an interest amongst both private companies

and research communities to expand the use of voice agent technologies, there is no widely accepted solution to the gender-based social justice problems of voice agents.

Given the myriad problems associated with the use of gendered voice agents, it is important to better understand the role of gender and gender stereotyping in interactions with gendered voice agents. Our work aims to defend the following thesis statement:

The widespread use of gendered voice agents, many of which are female or “female by default”, introduces gender-based social justice problems. Previous scholarship has focused on how women in particular are affected. Gendered agents can reinforce problematic stereotypes of women and exacerbate antisocial behaviours towards real women. By exploring the influence of voice agent gender and gender stereotyping in interactions with gendered voice agents across a variety of contexts, we can offer insights for designers to create ethical gendered voice agents with reduced harms to users and the broader society.

To defend this statement, we investigate the following research questions:

- **RQ1:** What gender-based stereotypes are applied to a voice agent running coach?
- **RQ2:** How do human and agent gender influence interactions with a voice agent running coach?

To address these research questions, we built a gendered voice agent running coach with three voice options: male, female, and gender ambiguous. We chose a coaching task because coaching is a male-dominated profession [38, 40] and thus a coaching agent is more likely to elicit gender stereotypical reactions. We conducted an in-person survey study ($n = 30$) to validate the voices would be perceived correctly in terms of gender and used the system in a within-participants field study ($n = 18$). Participants used the system over a 3-week period and attended two in-person or online sessions with the researcher. We collected survey and interview data to evaluate how users perceived and interacted with the voice agent running coach. We used a reflexive thematic analysis approach to uncover a story of three key themes, concluding with our discussion of steps designers can take to create gendered voice agents more responsibly. We also identify limitations of our approach and offer recommendations for future research into gender and gender-based stereotyping of voice agents.

Chapter 2

Background Literature

Voice agents are a type of conversational agent. Conversational agents are broadly defined as technologies that emulate human conversation [37]. Compared to other conversational agents such as robots and chatbots, voice agents use speech as their primary modality. Research on voice agents is interdisciplinary and has been conducted under a variety of domains, including Human-Computer Interaction (HCI), Human-Robot Interaction (HRI), and Artificial Intelligence (AI). In this section, we discuss the literature on gender and gender stereotyping in voice agents and other conversational agents. Finally, we explain our motivation to use a voice agent running coach in our field study.

Gender ascription in voice agents. The tendency to assign gender to technologies is well known [1]. This phenomenon can be attributed to the Computers are Social Actors (CASA) paradigm, which posits that people tend to apply the same social rules and expectations to technologies that they would to humans [53]. In 1977, Nass et al. demonstrated that people ascribe gender onto machines embedded with minimal gender cues [52] and as conversational agents become more humanlike, the tendency to gender them increases [76]. Past work has revealed various gender cues that can be used to manipulate the perceived gender of a conversational agent. Researchers have successfully leveraged verbal gender cues such as voice pitch [77] and speaking style [39], social gender cues such as names [54] and pronouns [64], and embodied gender cues such as robot body shape [4, 79] and chatbot avatar appearance [26]. However, users may not always interpret gender cues as intended. For example, roboticists and laypeople alike have gendered the humanoid robot Pepper in widely different ways, despite manufacturer claims that Pepper is gender neutral [68]. Researchers have also observed a male technology bias in which people ascribe masculine traits onto androgynous technologies that do not possess distinctive gender cues [57, 83].

With increasing visibility for transgender and gender non-binary people, interest in gender ambiguous voice agents is growing [12, 43, 13]. Sutton has recommended the term *gender ambiguous* to refer to androgynous voices that possess a mix of masculine and feminine characteristics [74]. This framing aims to offer more nuance than terminology such as *genderless* or *gender neutral*, which suggest that a voice is devoid of gender. Indeed, some researchers believe it is impossible to create a voice that is truly genderless due to the strong tendency to ascribe gender onto machines [71, 70]. There is a large body of work which explores approaches for generating gender ambiguous synthetic voices [78, 44, 17, 77, 80, 48, 35], such as combining the timbre of a male speaker with a female speaking style [39] or training TTS models on speech samples collected from gender non-binary and transgender individuals [13]. Researchers emphasize the need to work with transgender and gender non-binary users to create technologies that promote inclusion and diversity [12, 65, 73]. This finding is supported by evidence that gender ambiguous voices tend to be rated more negatively by gender non-binary participants than cisgender participants [14]. We contribute to an understanding of gender ambiguous voice agents by including a gender ambiguous voice as a control in our field study in addition to the traditionally studied male and female voices.

Gender stereotyping of voice agents. Gender stereotyping involves overgeneralizing about the attitudes, traits, or behaviour patterns of women or men [21, 42]. In keeping with the CASA paradigm, people tend to apply gender stereotypes to technologies. Anthropomorphizing and assigning gender to technologies increases gender stereotyping [61, 45] and gender stereotypes can be elicited by minimal gender cues [52]. Gender-stereotypical responses can also be elicited by non-gender-based design factors, such as task type and environment. For example, users are more likely to apply gender stereotypes when chatbots operate under gender-stereotypical subject domains like banking or mechanics [46], and an investigation into domestic humanoid robots for home use found that female robots are preferred over more masculine ones [7]. Further, users may be more likely to apply gender stereotypes when a machine’s assigned gender does not match stereotypical expectations about the task it is performing [46, 11].

There is a large body of work suggesting the effects of voice agent gender are driven by gender-based stereotypical responses [41, 49, 69, 29, 44, 51, 63]. For example, Nagar found that responses to a self-service voice assistant were consistent with stereotypes of men and women [51]. Participants expected a female voice assistant to be more sympathetic and helpful, while a male voice was preferred for higher-complexity tasks. Further, participants responded more positively to failure recovery by a male voice assistant. Nagar suggests these responses could be driven by stereotypical views that men are highly competent and

better at problem solving, while women are warmer and more passive. Mahmood and Huang similarly investigated gender biases in responses to voice assistant failure recovery strategies and found that male participants behaved more socially with female voice assistants that apologized for mistakes [44]. Further, male participants interrupted voice agents more frequently than women, indicating that gender-based power dynamics may influence interactions with voice agents. Men are also known to report higher autonomy satisfaction (within the Basic Psychological Needs theory) when interacting with low-agency female voice agents, compared to male or high-agency female agents [49].

The problem of gendered voice agents. The high prevalence of female or “female by default” commercial voice agents can lead to problems of social justice. To justify the decision to make voice agents female, companies like Apple and Amazon have cited academic work suggesting people prefer a female voice [81], but it is possible these decisions and preferences are driven by gender-stereotypical expectations of women. A 2019 UNESCO report highlighted how female commercial voice agents reinforce harmful gender stereotypes of women [81]. The tendency to prefer female voice agents may be attributed to gender stereotypes that align with the marketing of voice agents, which place women in subservient roles [81, 82]. This theory is supported by analyses of commercial voice agents that have found they are designed to conform to stereotypical expectations and gender roles [34, 25].

The investigation into how to design gendered voice agents responsibly remains open. Curry et al. [9] conducted a workshop in which participants were presented with the problem of gendered technologies and asked to submit an ideal voice agent design. However, no clear consensus was found, leading the authors to recommend increased customization options for voice assistants [9]. Other previous work has investigated the use of gender ambiguous voices to reduce gender stereotyping of voice agents [12, 50, 78, 43, 44]. However, because gender ambiguous voices may still elicit gender ascription, the effectiveness of this approach has been called into question [74, 48, 17]. Eschewing concerns surrounding ambiguous agents, Duan et al. used eXplainable AI (XAI) techniques to de-anthropomorphize agents, thereby reducing attribution of social characteristics and mitigating gender stereotypes [20]. However, the feasibility of this approach may be limited as non-anthropomorphic voice agents are less preferable to users [45].

Additionally, personified, humanlike voice assistants introduce the risk of users conflating humanlike female voice assistants with real women [81]. Seeing commercial female voice assistants like Siri and Alexa make frequent mistakes may reinforce negative perceptions of women in the minds of users [81]. Further, female technologies that are unable

to navigate antisocial dynamics risk replicating or exacerbating gender-based violence and harassment [81, 28]. Prior work has found commercial conversational agents often respond neutrally or deflect verbal abuse, but these behaviours may be incorrectly interpreted as flirtatious by some users [8, 81, 10]. The question of how best to respond to antisocial user behaviours remains unresolved. Prior research suggests that users react negatively to aggressive responses [10, 84], but are accepting of assertive responses [15]. Chin et al. provide what is perhaps the most concrete recommendation, suggesting that empathetic responses can elicit emotions of guilt and may lead users to stop the abuse [10]. To add to this body of work, we identify and analyse aggression towards voice agents in our qualitative data to better understand these antisocial behaviours.

In 2025, Duan et al. conducted a scoping review to evaluate how gender-based stereotypes are applied to AI [19]. To our knowledge, there are no other literature reviews focusing primarily on gender-based stereotypes in gendered technologies. Duan et al. found gaps in the research on gender-based stereotypes of AI and identified opportunities for future research. Notably, only 9 of the 89 reviewed studies used voice agents and two used a field experiment methodology. The authors suggested that studies combining field research with qualitative methodologies can provide valuable insights. We contribute to the existing literature on gender stereotyping of voice agents through our use of these heretofore underutilized methodologies. In our work, we conduct an exploratory field study in which participants interact with a novel running coach system and we use both qualitative and quantitative methods to explore gender and gender-based stereotyping in voice agents.

Sport coaching, gender stereotypes, and voice agents. Across the globe, sport coaching is a predominantly male occupation in nearly all sports and levels of competition [38, 40]. Investigations into the under-representation of women in coaching positions have found that expectations of coaches are driven by stereotypical ideas about gender [40, 56, 67]. For example, an interview study with German soccer coaches and players revealed that traits like assertiveness and determination are associated with male coaches, while female coaches are believed to be more empathetic [66]. There is also minor evidence for a “male is better” bias in which male coaches tend to be evaluated more positively by athletes [62, 47]. Given the pervasiveness of gender stereotyping in coaching, we thought a voice agent running coach would naturally elicit stereotypical reactions from participants and prompt discussions about voice agent gender, thus making it a good fit for our exploratory field study.

Chapter 3

System

To facilitate the study, we developed a mobile application to run on Android and iOS devices. At a high level, our system used location information to estimate the user’s running pace and provided verbal feedback to help users run at a self-determined target pace for the entire duration of the run. Our system was built using the React Native and Expo front-end frameworks, a back-end service using the FastAPI framework and a PostgreSQL database, and was hosted on secure University of Waterloo servers. Participants installed the system through the Google Play Store and Apple Test Flight software programs. To guide the design of the system, we used an iterative approach. Volunteers (students and researchers in our lab) tested the system and provided informal feedback. Volunteers received no remuneration for their testing efforts and were not eligible to participate in any formal user studies.

3.1 Voice Agent

Interactions with the voice agent were one-way: the agent provided verbal feedback as users ran, but users were not able to provide voice commands to the agent. We chose not to allow users to speak with the agent to avoid creating the expectation that the agent could converse intelligently. The implementation of intelligent conversation would have required the use of unpredictable generative LLMs and introduced the risk of generating harmful feedback. To ensure participant safety, dialogue was generated prior to the user study and manually reviewed (see Section 3.1.2 for implementation details).

3.1.1 Voice Design

We selected industry-standard voices from a set of English-speaking Google Cloud Text-to-Speech (TTS) API voices. We chose this TTS API because it provides developers control over voice pitch and speed. It also supports Speech Synthesis Markup Language (SSML) to allow customization over the audio response by providing details such as pauses and audio formatting for acronyms, dates, and times. Based on volunteer feedback, we selected three voice configurations (male, female, and gender ambiguous) that were sufficiently intelligible and lifelike. To create a gender ambiguous voice, we followed a method from previous studies and used a female voice with a lowered pitch [78, 44, 17, 77, 80, 48, 35, 20]. We chose this method because it was most efficient to implement with our TTS API, compared to alternative approaches which would require the use of a specialized LLM.

3.1.2 Dialogue Design

Category	Example Statement
Start	Your run has started. Begin strong and build your rhythm.
Status	
Above pace	That’s faster than your target pace.
In range	You’re staying right on track.
Below pace	You’ve dropped under your target pace.
Immobile	That wasn’t your target pace, not even close.
Instructive	
Above pace	Slow down just a little to save your energy.
In range	Keep moving with the same rhythm and steady effort.
Below pace	Take quicker steps and put in a bit more effort.
Immobile	Start increasing your step rate gradually to break into a run.
Task-oriented	
Above pace	Try to ease off and hold a slower rhythm for 30 seconds.
In range	Try to hold your current rhythm for 30 seconds.
Below pace	Try to increase your speed steadily for 30 seconds.
Immobile	Try kicking up the effort and keep it steady for 30 seconds.

Table 3.1: Statement categories and example statements used with our voice agent feedback template.

To guide the dialogue design for the voice agent, we adapted methods from the sports

coaching literature. The Skill Training Communication Model outlines seven methods for providing verbal feedback and instruction to athletes [60, 59]. We selected the instructive and task-oriented methods. In the instructive method, coaches give direct, prescriptive instructions to the athlete. In the task-oriented method, coaches provide a task to challenge the athlete. Compared to other coaching methods such as video feedback (in which athletes are shown video demonstrations) or analogy learning (in which coaches use metaphors to aid explanations), we determined the instructive and task-oriented methods were feasible for our system and most efficient to implement. We chose to adapt two different coaching methods to add variation to the feedback and try to avoid repetitive dialogue.

To evaluate our dialogue design, we asked volunteers to review written dialogue statements and audio samples of different dialogue options. Based on these informal discussions, we developed a simple feedback template that was used on the client side to generate dialogue statements in real time as users ran. Table 3.1 describes the statement categories used with the template and example statements for each selected coaching method. To commence each run, the agent delivered a randomized **start** statement. For the duration of the run, the agent delivered feedback at a random interval (30, 35, 40, or 45 seconds) using one randomized **status** statement (to indicate if the user’s speed was above, within range, or below pace), followed by one randomly selected **instructive** or **task-oriented** statement (to suggest ways the user could modify their behaviour to stay on pace). We used ChatGPT (GPT-4o) to generate 50 statements for each category, for a dataset of 650 statements in total. Each statement was manually reviewed by the author to ensure it used a consistent, conversational tone and contained sensible information. We chose not to disclose the user’s estimated pace in the agent’s feedback to avoid conflicts with more precise commercial exercise tracking technologies, such as smartwatches or software like Strava. This solution was deemed acceptable because volunteers did not report significant inaccuracies during testing and found the feedback credible.

3.2 Interface

We chose a plain, simplified aesthetic to reinforce the understanding that our system is a prototype and avoid implying that its performance is comparable to commercial exercise tracking software like Strava and Nike Run Club. We decided limit the user interface features to direct participant attention towards the voice agent design. For example, no performance data is displayed during runs to discourage participants from watching the interface instead of listening to the agent’s dialogue. In this section, we provide an overview of the system screens and highlight key features.

3.2.1 Login and Home Screens

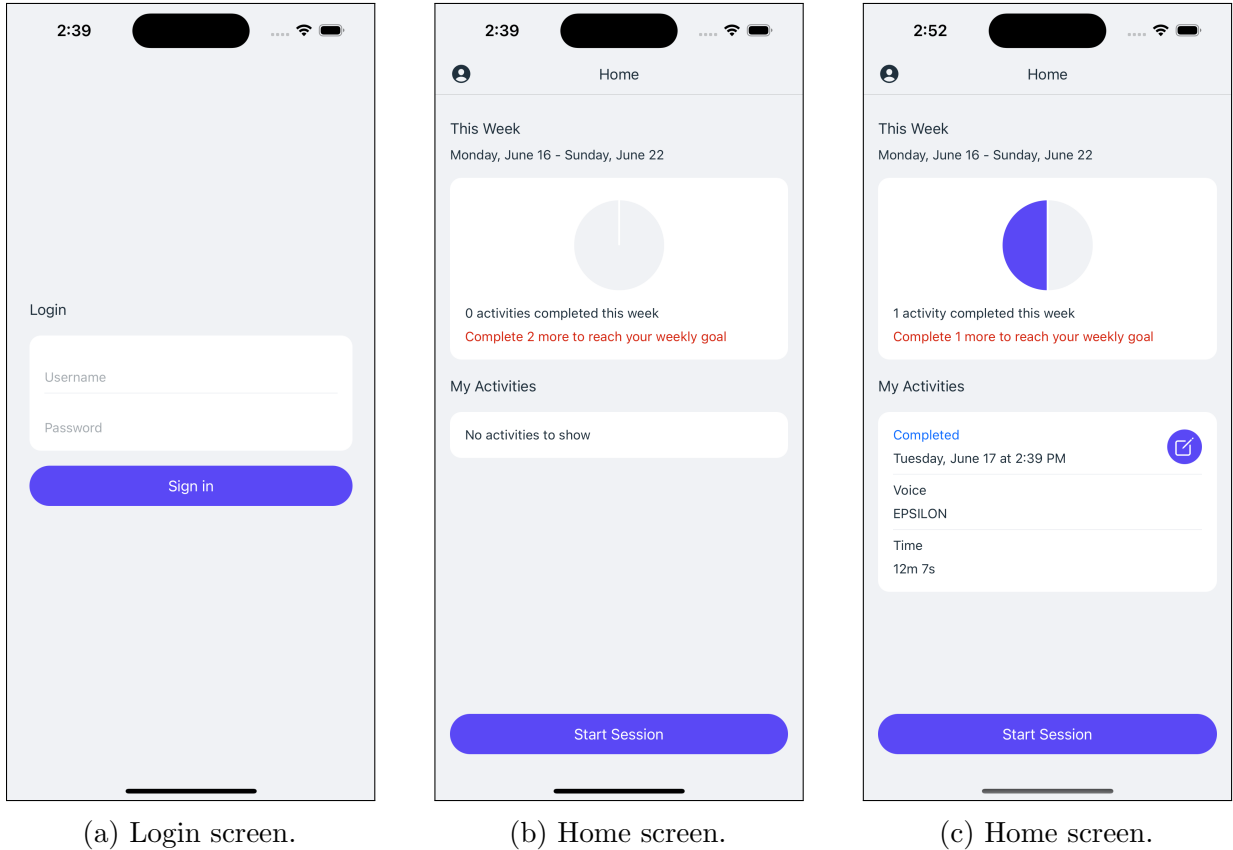


Figure 3.1: Login and home screens and key features.

Users signed into the system from the login screen (see Figure 3.1a) with credentials (a non-identifiable username and unique password) provided by the researchers. Once signed in, a user’s session was valid until the end of the study period. Signed-in users were routed to the home screen (see Figure 3.1b), which displayed a history of the user’s activities and a summary of the user’s activity for the current week. To motivate users to run, we aimed to use gamification [16], which refers to the use of game elements in non-gaming contexts. Our gamified interface challenged users to complete two runs each week (per the study tasks) and visualized their progress with a circular indicator. Users could use the **Start Session** button on the home screen to record a new activity.

3.2.2 Run Configuration Screen

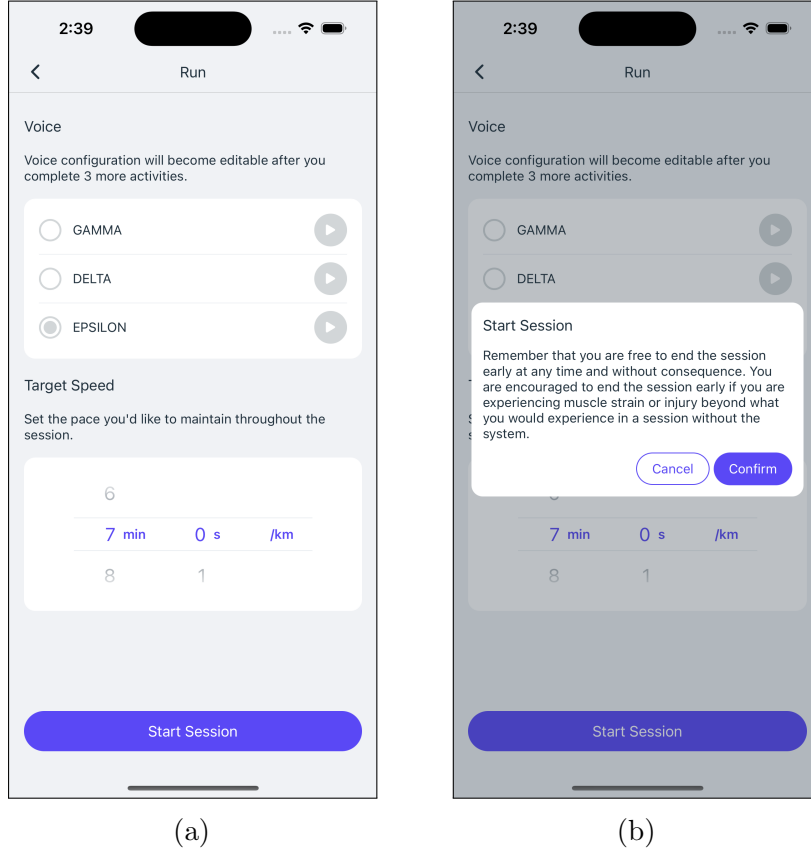


Figure 3.2: Run configuration screen and key features.

To accommodate all running experience levels, users could configure their performance goal for each session (see Figure 3.2a). After three sessions, users could also select a voice configuration for the agent. To ensure users ran with each voice at least once, the system assigned a voice for the first three sessions. The order in which users were exposed to the voice configurations was determined by a Latin square design and researchers manually assigned each user to a sequence. We aimed to select gender ambiguous and stigma-free names (Gamma, Delta, and Epsilon which corresponded to the male, female, and gender ambiguous voices, respectively) for each agent based on volunteer feedback. Users could preview each voice by pressing the **Play** button (recognizable by the *play* icon) next to its name. To start the session, users pressed the **Start Session** button. Before continuing, the system displayed a confirmation message to remind users they could end the session

early if they were experiencing distress beyond what they would experience on a typical run without the system (see Figure 3.2b). The purpose of this message was to reduce the risk of harm to participants.

3.2.3 In-Progress Run Screen

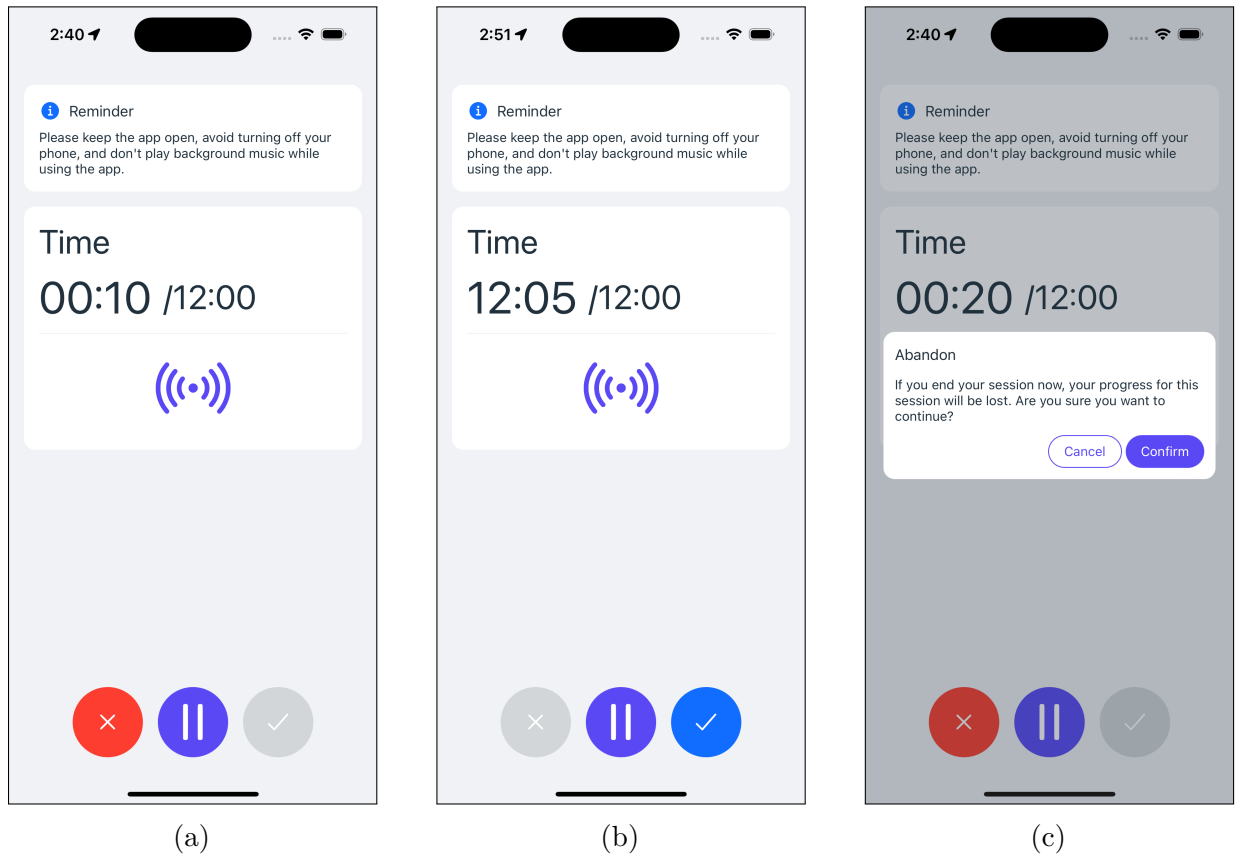


Figure 3.3: In-progress run screen and key features.

During each running session, users could see the time elapsed since the system started measuring their performance (see Figure 3.3a). The system used the device’s GPS and location data to estimate the user’s average pace and provide voice agent feedback designed to keep users moving within range of with their target pace. To protect user privacy, the system calculated the user’s pace locally on the device and did not transmit location data to remote servers. To complete a session, users needed to run for a minimum of 12 minutes,

as per our study requirements. After 12 minutes, the **Complete** button (recognizable by the *checkmark* icon) was enabled (see Figure 3.3b, allowing users to continue to the completion screen. Prior to this 12-minute threshold, an **Abandon** button (recognizable by the *X* icon) would be enabled. Pressing this button would cause the system to display a confirmation dialogue to ensure the user wanted to end the session early (see Figure 3.3c). Users could also use the **Pause** button (recognizable by the *pause* icon) to pause the session and suspend performance tracking. This feature was implemented to accommodate natural pauses in the run, such as waiting for traffic lights or stopping to retie a shoelace. Users could resume the session by pressing the **Pause** button again.

3.2.4 Post-Run Survey Screen

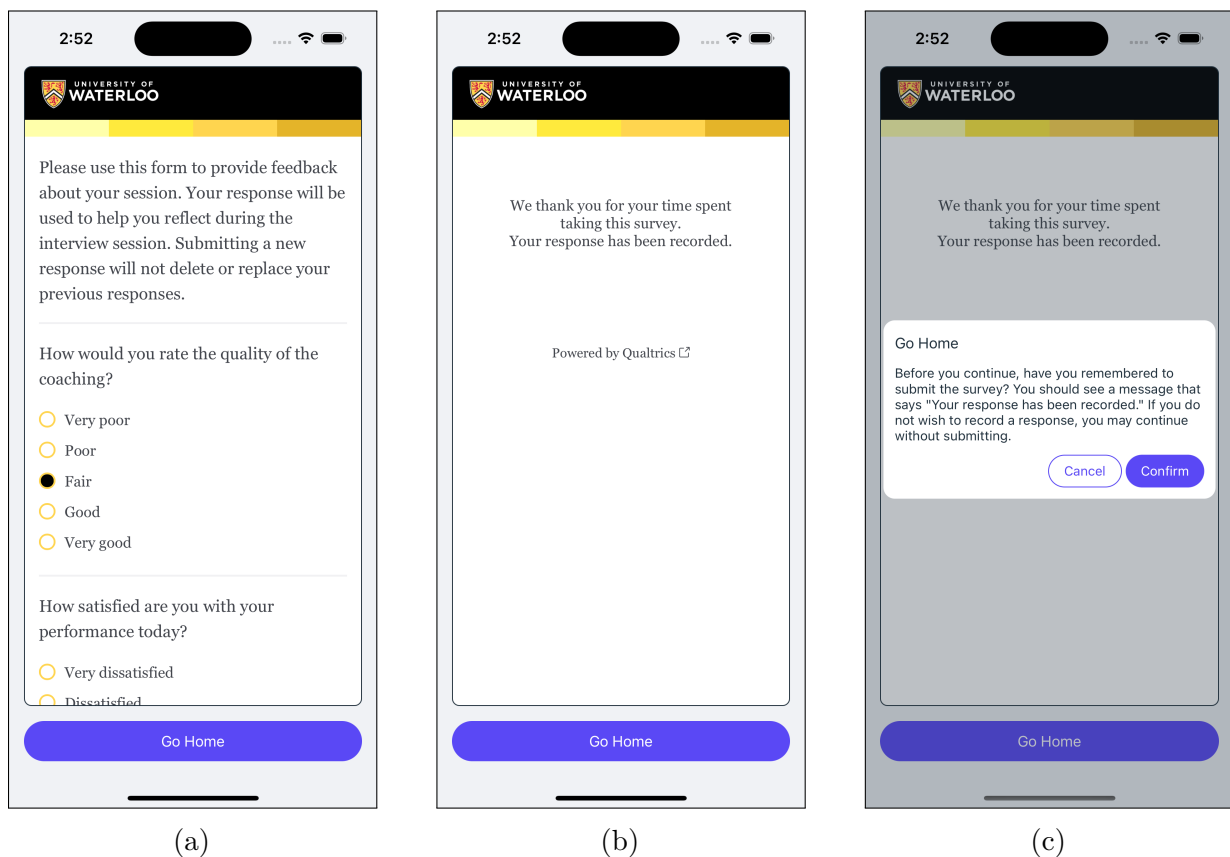


Figure 3.4: Post-run survey screen and key features.

After completing each session, users were redirected to the survey screen (see Figure 3.4a), where they were presented a Qualtrics survey inside an embedded browser. We chose to use Qualtrics, a popular online survey tool, instead of creating a custom form because it enabled us to easily create or modify survey questions, as well as viewing and exporting survey responses in a standard file format. To link survey responses to user sessions, we used the Qualtrics embedded data feature to include a unique session identifier (which matched our database records) through a query parameter in the survey URL. To ensure users submitted the survey before leaving the screen, the system displayed a confirmation message to users when the **Go Home** button was pressed (see Figure 3.4c).

3.3 Data Collection

The system collected data about each session, including the voice configuration used, elapsed time (adjusted for pauses), distance travelled, average pace, and the start and end timestamps (recorded when the user pressed the **Start Session** and **Complete Session** buttons, respectively). At each interval, we recorded a timestamp, performance data (elapsed time, distance travelled, and average pace for the interval), and the feedback spoken by the voice agent in plain text format. Interval performance data was used in combination with session data to verify integrity of each session. For example, we could compare the session start and end timestamps with the elapsed time to determine how long a session was paused. We could also use the performance data at each interval to determine if, for example, the user was running, walking, or standing still during the session.

Chapter 4

Pilot Study

To validate if the selected voices would be perceived accurately in terms of gender and explore if participants gender stereotypes might influence perceptions of each voice, we conducted a within-participants survey study ($n = 30$). The voice configurations used in this preliminary pilot study are presented in Table 4.1.

Gender	Name	Pitch	Speed
Male	en-US-Standard-I	1.0	1.0
Female	en-US-Standard-C	2.5	1.0
Ambiguous	en-US-Standard-C	-6.0	1.0

Table 4.1: Google Cloud Text-to-Speech API voice configurations used in the pilot study.

4.1 Method

The study was conducted on campus at the University of Waterloo and lasted approximately 30 minutes. Data was collected using the online survey tool Qualtrics. After arriving at our lab, participants completed the informed consent process and then filled out the survey using a computer and headphones provided by the researchers. Our study had three conditions, where each condition was an audio recording of a different voice configuration (male, female, gender ambiguous) for participants to evaluate. In each condition, the agent used the same neutral coaching dialogue:

You’ve dropped 15 seconds in your last interval. Focus on pushing harder to hit your target pace of 5 minutes 30 seconds per kilometer.

To avoid influencing participants, we were careful not to accidentally provide any information (such as audio file names) that could be used to draw conclusions about the voice’s gender. During the survey, voices were evaluated sequentially and participants were not permitted to modify their ratings for a previous voice after proceeding to the next one. Participants were asked to carefully review their ratings for each voice before proceeding. We used a Latin square design to determine the order in which participants were exposed to conditions.

4.2 Measures

Participants provided demographics including age, gender, and English language proficiency (see [Demographics Questionnaire](#) in Appendix [Pilot Study](#)). For each voice, participants filled out a survey to measure perceived gender, presence of gender-stereotypical traits, and social characteristics such as likeability and intelligence (see [Voice Agent Evaluation Questionnaire](#) in Appendix [Pilot Study](#)). To measure perceived gender, we asked participants to rate their perception of the voice on a 7-point semantic differential scale (1 = *masculine*, 4 = *ambiguous*, 7 = *feminine*). To investigate effects of gender stereotyping, we followed a procedure from previous voice agent studies and asked participants to rate the presence of four stereotypically male and four stereotypically female traits in the agent on 5-point Likert scales [77, 17, 22]. To evaluate perceptions of the voice’s social characteristics, we used items from the Godspeed Questionnaire Series (GQS) [3] to measure perceived likeability (Godspeed III), perceived intelligence (Godspeed IV), and perceived safety (Godspeed V). The GQS is one of the most highly-cited questionnaires in the field of HRI [2] and has been used in previous research to evaluate perception of voice agents [30]. Lastly, we administered the Negative Attitudes towards Robots Survey (NARS) [55] to capture potential confounding factors.

4.3 Participants

Level	Definition
Novice	Has been running consistently for less than 6 months.

Recreational	Has been running consistently for more than 6 months.
High-calibre	Has been running consistently for more than 3 years.

Table 4.2: Operational definitions of running experience levels.

We recruited 30 participants from the University of Waterloo using on-campus posters, student mailing lists, social media posts, and word of mouth. We used a screening protocol to recruit a mix of running experience levels, as shown in Table 4.2. They are adapted from definitions by Hoitz et al. [32] and aim to capture a person’s consistency and personal investment in the sport, rather than performance ability. The average age was 25.9 years old ($SD = 5.5$ years) and 90% of participants self-identified as native or bilingual English speakers. Participants varied in gender identity, nationality, educational background, experience using fitness applications, experience using voice agents, and running experience level (see Table 4.3). Participants were remunerated with a \$10 CAD gift card. The study received approval from the Research Ethics Board of the University of Waterloo (ORE#46929).

Gender	
Female	19
Male	11
Nationality	
Bangladeshi	1
Canadian	19
Chinese	1
Hongkongese	1
Indian	2
Iranian	1
New Zealander	1
Nigerian	1
Nigerian-Canadian	1
Russian-Canadian	1
Highest completed or current degree	
Bachelor’s degree	12
Master’s degree	10
Doctoral degree	5
High (secondary) school diploma or equivalency certificate	3
Program of study	

Business and Management	2
Computer Science	9
Engineering	6
Health Sciences	3
Humanities and Social Sciences	2
Mathematics	2
Science	6
Runner experience level	
Novice	21
Recreational	9
Used a fitness application before?	
Yes	26
No	4
Fitness application usage in a typical week	
Every day	8
A few times a week	4
Once or twice a week	3
Once a week	8
Never	7
Used a voice agent before?	
Yes	27
No	3
Voice agent usage in a typical week	
Every day	3
A few times a week	9
Once or twice a week	1
Once a week	12
Never	5

Table 4.3: Pilot study participant demographics.

4.4 Results

The data was not normally distributed. A Shapiro-Wilk test resulted in $p < 0.05$ for all measurements except perceived likeability, which was normally distributed ($p > 0.05$). For normally distributed data, we examined the effects of the conditions (male voice, female

voice, gender ambiguous voice) using the one-way repeated measures ANOVA and post hoc testing with the Bonferroni correction. To evaluate non-normally distributed data, we used the non-parametric Friedman test with a post hoc Durbin-Conover test.

4.4.1 Perceived Gender

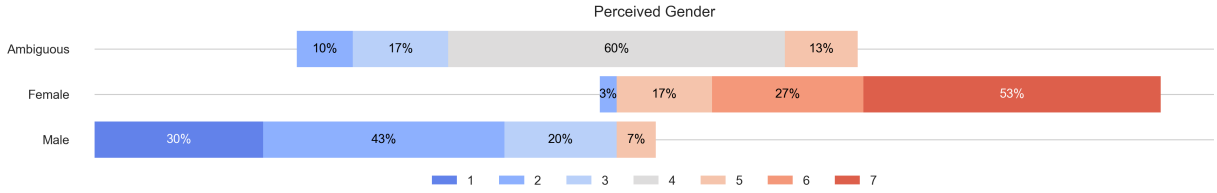


Figure 4.1: Gender perception scores in each condition. Scores range from 1 (masculine) to 4 (ambiguous) to 7 (feminine).

Figure 4.1 shows how participants rated the gender of the voice agent on a 7-point semantic differential scale. The male voice received an average rating of 2.1 (SD = 1.062), the female voice 6.233 (SD = 1.104), and the ambiguous voice 3.767 (SD = 0.817). We found a significant effect of voice agent configuration on gender perception ($\chi^2(2) = 50.034$, $p < 0.001$, $W = 0.834$). Pairwise comparisons showed a significant difference between all voices (all $p < 0.001$).

4.4.2 Trait Ascription

Participants rated the presence of 4 stereotypically male traits (aggressiveness, assertiveness, authoritativeness, and toughness) and 4 stereotypically female traits (empathy, friendliness, politeness, and sincerity) in each voice agent using 5-point Likert scales. Scores ranged from 1 to 5, with higher scores indicating higher agreement that the agent possesses a given trait.

Stereotypically Male Traits

Figure 4.2 shows the trait ascription scores for stereotypically male traits. We performed Friedman tests to investigate if there was an effect of voice agent gender on perceived aggressiveness, assertiveness, authoritativeness, and toughness. We found a significant

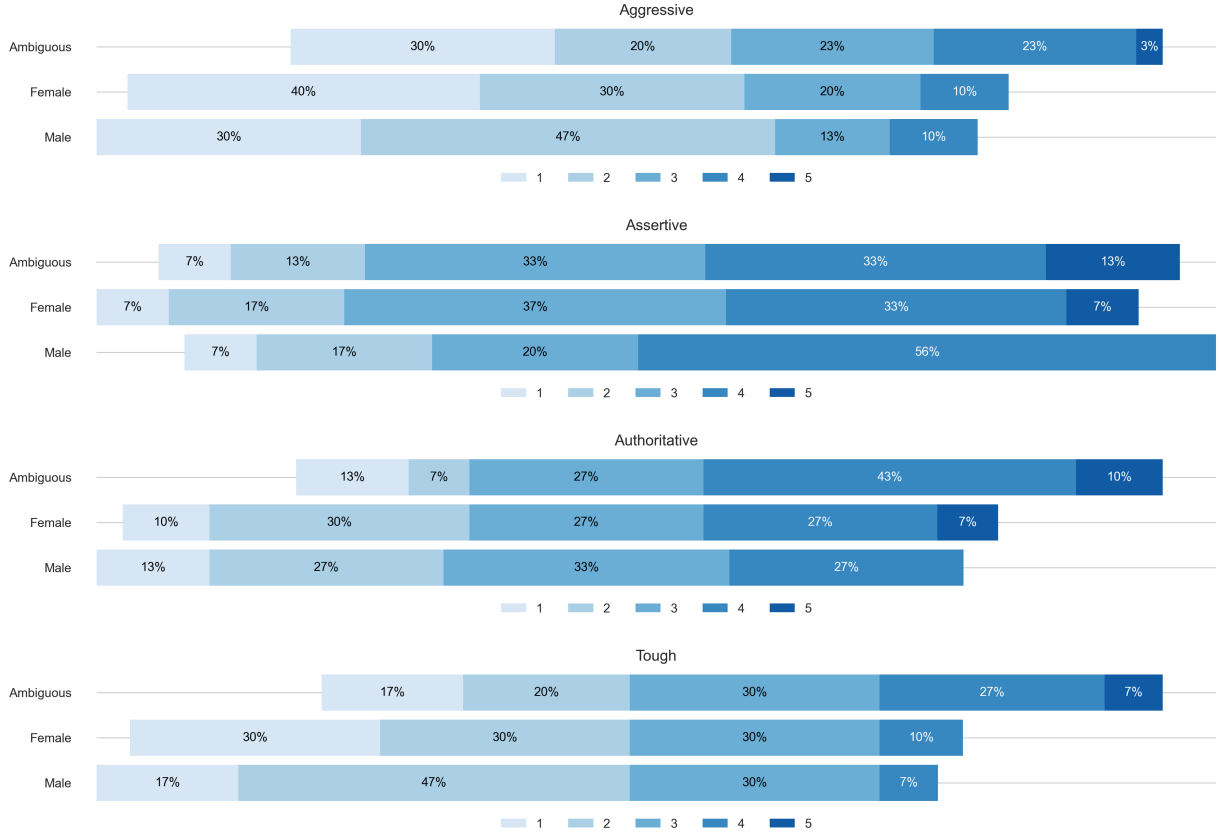


Figure 4.2: Trait ascription scores for stereotypically male traits in each condition. Scores range from 1 (strongly disagree the agent has this trait) to 5 (strongly agree the agent has this trait).

effect of voice agent gender on perceived authoritativeness ($\chi^2(2) = 7.558$, $p < 0.05$, $W = 0.126$) and perceived toughness ($\chi^2(2) = 7.823$, $p < 0.05$, $W = 0.13$). Pairwise comparisons showed a significant difference between the authoritativeness of the ambiguous voice and the other voices (both $p < 0.05$). We also found a significant difference between the toughness of the ambiguous voice and the other voices (both $p < 0.05$).

Stereotypically Female Traits

Figure 4.3 shows the trait ascription scores for stereotypically female traits. Friedman tests showed there was an effect of voice agent gender on perceived empathy ($\chi^2(2) = 17.614$,

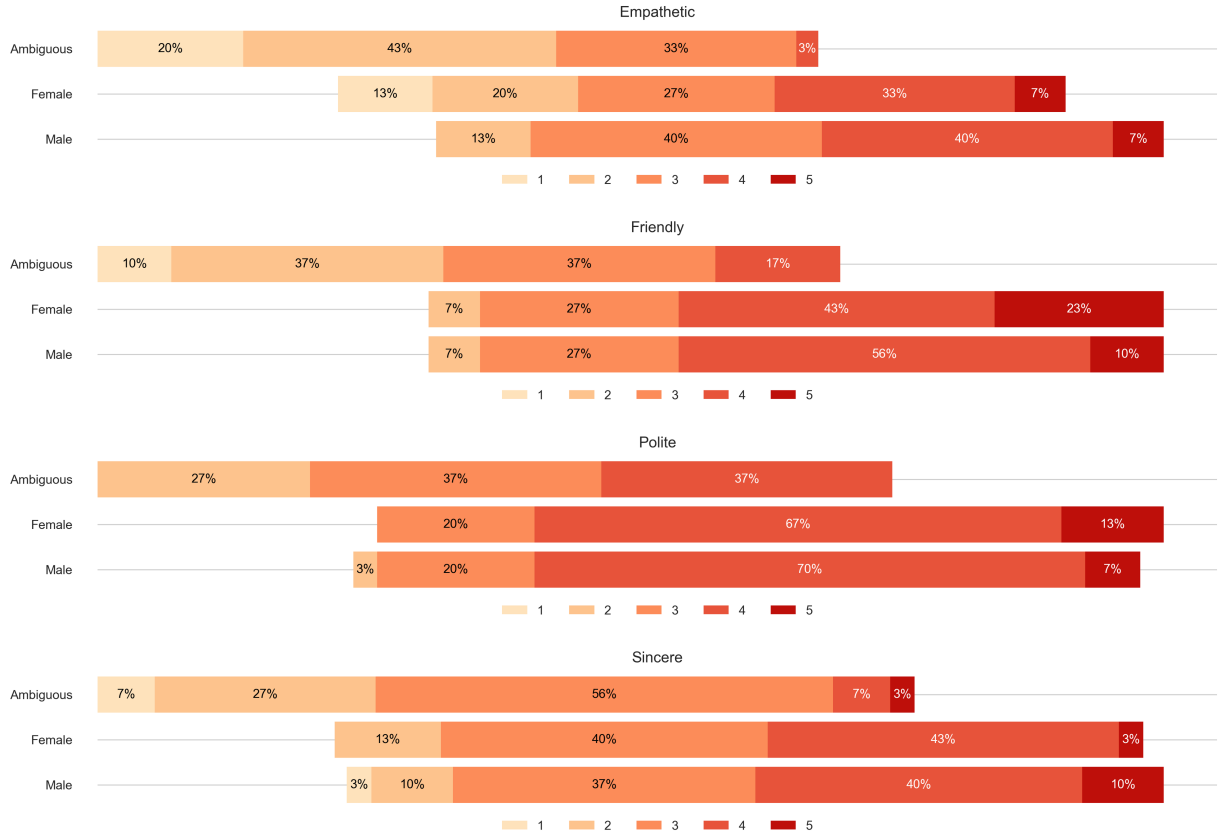


Figure 4.3: Trait ascription scores for stereotypically female traits in each condition. Scores range from 1 (strongly disagree the agent has this trait) to 5 (strongly agree the agent has this trait).

$p < 0.001$, $W = 0.294$), friendliness ($\chi^2(2) = 29.082$, $p < 0.001$, $W = 0.485$), politeness ($\chi^2(2) = 18.328$, $p < 0.001$, $W = 0.305$), and sincerity ($\chi^2(2) = 15.103$, $p < 0.001$, $W = 0.252$). Pairwise comparisons showed a significant difference between the ambiguous voice, compared to the male and female voices. We found that the gender-ambiguous voice received significantly different scores from the male and female voices for all traits (all $p < 0.05$). There was no significant difference between the male and female voices for any of the traits.

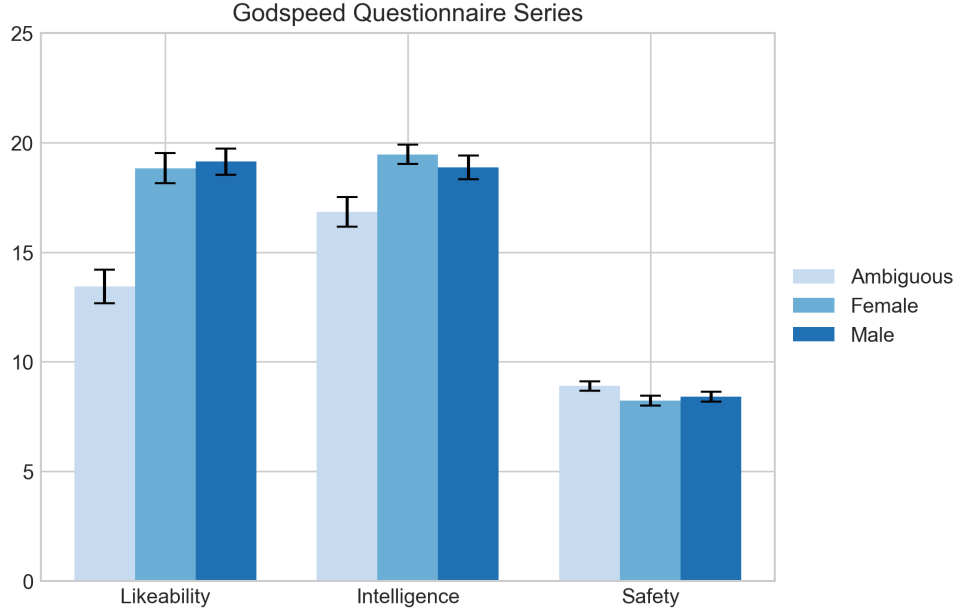


Figure 4.4: Averages for Godspeed Questionnaire Series scores in each condition. Error bars indicate standard error.

4.4.3 Godspeed

Perceived Likeability

Figure 4.4 shows how participants rated the perceived likeability of the agent using the GQS likeability questionnaire (Godspeed III). A higher score indicates higher perceived likeability, with scores ranging from 1 to 25. The male voice received an average score of 19.133 ($SD = 3.267$), the female voice 18.833 ($SD = 3.752$), and the ambiguous voice 13.433 ($SD = 4.224$). We found a significant effect of voice agent gender on perceived likeability ($F(2, 58) = 27.52, p < 0.001, w^2 = 0.318$). Post hoc analysis showed a significant difference between the ambiguous voice, compared to the male ($Mean = 5.7, p < 0.05$) and female voices ($Mean = 5.4, p < 0.001$).

Perceived Intelligence

Figure 4.4 shows how participants rated the perceived intelligence of the agent using the GQS intelligence questionnaire (Godspeed IV). A higher score indicates higher perceived

intelligence, with scores ranging from 1 to 25. The male voice received an average score of 18.867 (SD = 2.933), the female voice 19.467 (SD = 2.46), and the ambiguous voice 16.833 (SD = 3.687). We found a significant effect of voice agent gender on perceived intelligence ($\chi^2(2) = 19.03$, $p < 0.001$, $W = 0.317$). Post hoc analysis showed a significant difference between the ambiguous voice and the other voices (both $p < 0.001$).

Perceived Safety

Figure 4.4 shows how participants rated the perceived safety of the agent using the GQS safety questionnaire (Godspeed V). A higher score indicates higher perceived safety, with scores ranging from 1 to 15. The male voice received an average score of 8.4 (SD = 1.248), the female voice 8.233 (SD = 1.251), and the ambiguous voice 8.9 (SD = 1.185). We found a significant effect of voice agent gender on perceived safety ($\chi^2(2) = 6.395$, $p < 0.05$, $W = 0.107$). Post hoc analysis showed a significant difference between the ambiguous voice and the female voice ($p < 0.05$).

4.5 Discussion

We intended to design three voices for the agent with distinctly different perceived genders. Our results showed we were successful and suggested that participants in the field study would perceive the of each voice as intended. We found a significant difference in perceived gender between all the voices and the average gender perception score for each voice tended to align with our design intentions. The average male voice score was closer to 1 (masculine) than 4 (ambiguous), the average female voice score was closer to 7 (feminine) than 4, and the average ambiguous score was closer to 4 than 1 or 7.

We also found evidence to suggest that perceptions of the voice agent’s traits may vary by voice configuration. Compared to the male and female voices, participants rated the ambiguous voice as significantly more authoritative and tough (stereotypically masculine traits) and significantly less empathetic, friendly, polite, and sincere (all stereotypically female traits). However, we do not interpret these findings to mean that participants viewed the ambiguous voice as more masculine or feminine than the male and female voices. Participants also rated the ambiguous voice as significantly less likeable and intelligent than the male and female voices. Further, no differences between the male and female voices were found beyond perceived gender. Thus, we find that participants responded more negatively to the gender ambiguous voice, but we cannot conclude this is related to gender stereotyping. In the field study, we collect qualitative data from participants to

further investigate this finding. In addition, we use the GQS perceived anthropomorphism questionnaire (Godspeed I) to evaluate the humanlikeness of each voice, as studies have shown that overly synthetic voices can negatively impact perceptions of agents [83]. There is also evidence that ambiguous sounding voices tend to be perceived as more robotic compared to male and female ones [44].

Chapter 5

Field Study

To investigate gender and gender-based stereotypes in voice agents, we conducted a within-participants field study with a voice agent running coach. In our study, 18 participants were asked to complete 6 runs with our system using the voice configurations described in Table 4.1 during a 3-week usage period.

5.1 Method

The 3-week within-participants field study was conducted from June to September, 2025. Participants were given the option to attend meetings in person at the University of New Brunswick or online using video conferencing software. Meetings were conducted by the first author and the researcher’s face was shown during online meetings. Survey data was collected using Qualtrics. To avoid biasing participants, we used a partial disclosure approach in which we did not disclose our goal of exploring gender and gender-based stereotypes in voice agents. Instead, participants were told the main purpose of the study was to investigate how to design voice agent running coaches. Participants were debriefed immediately after receiving remuneration for completing the post-study session. The debriefing procedure required participants to provide renewed consent for their data to be used. Our study had three stages, which we describe below.

Pre-Study Session. The pre-study session took approximately 30 minutes to complete. Participants first completed the informed consent process and filled out a short demographics survey. In-person participants completed the survey using a computer provided

by the researcher and online participants used a personal device. The researcher then guided participants through the steps to install the system onto their smartphone. Lastly, the researcher conducted a training session with the system to demonstrate key features and encourage proper usage during the in-study period.

In-Study Period. The 3-week in-study period started immediately following the pre-study session. Participants were instructed to complete at least two outdoor running sessions with the system each week for three weeks. Participants were free to determine their own running schedules during this time. We only asked participants to aim for two runs per week and to leave at least one day between each run. To disincentivize participants from completing the runs in a short time frame, participants were required to wait the full three weeks before they could attend the post-study session. Each session was estimated to take at least 15 minutes to complete, with a minimum of 12 minutes of running and an optional post-session questionnaire. Participants were told the purpose of the post-session questionnaire was to aid recollection during the post-study interview and that completing the post-session questionnaire was optional. Participants were told they could use the system more than two times per week and for more than 12 minutes per session, if they so wished. Participants were informed they would not receive additional remuneration if their usage exceeded our clearly defined minimum requirements: two runs per week and 12 minutes per run.

Post-Study Session. Participants were asked to attend the post-study session within 3 business days of completing the in-study session. This time frame was chosen to motivate participants to complete the post-study session in a timely manner (while their recollection of the system was strongest) while allowing participants to accommodate their personal schedules. The post-study session took approximately 1 hour to complete. Participants first filled out a survey to measure perceptions of the voice agent voices used by our system (compared to the pilot survey, participants saw the voices in a randomized order but similarly could not modify previous ratings) and then participated in an interview to share their experience using the system and their perception of the voices. Interviews were audio recorded for analysis and participants were clearly notified when the researcher started and stopped the recording. Upon completing the interview, participants were offered remuneration before the researcher conducted the debriefing procedure in accordance with our ethics protocols.

5.2 Measures

To investigate gender and gender stereotyping in our agent, we collected qualitative and quantitative data. In addition to capturing system usage data throughout the 3-week usage period (see Section 3.3), we asked participants to complete surveys and interviews. Our demographics survey captured relevant demographics such as age, gender, knowledge of artificial intelligence systems, experience using voice agents, and experience coaching or being coached in running (see [Demographics Questionnaire](#) in Appendix [Field Study](#)). Our post-session questionnaire, optionally filled out after each run, aimed to capture immediate impressions of the system using two ordinal scales and one open-ended question, which allowed participants to write down anything that stood out to them during the run (see [Post-Session Questionnaire](#) in Appendix [Field Study](#)). Our post-study voice evaluation survey aimed to evaluate perceptions of each voice and reused most of the questions from our pilot study (see Section 4.2). Compared to the pilot study, perceived gender ratings (ranging from 1 to 7) functioned as a manipulation check and we chose to measure perceived anthropomorphism (Godspeed I) instead of perceived safety.

To explore responses to the system, we conducted a semi-structured interview. Participants were asked a pre-prepared set of questions (see [Interview Questions](#) in Appendix [Field Study](#)) and the researcher was free to expand discussions relevant to voice agent gender and gender-based stereotypes by asking additional unscripted questions. Data from the in-study session was used to prompt participants to reflect on voice agent preferences and experience using the system. Interviews were audio recorded for analysis and transcribed for analysis by the first author.

5.3 Participants

We recruited 20 participants using posters, student mailing lists, social media posts, and word of mouth. To engage runners, we shared our notice directly with local running groups. We used a screening protocol to recruit runners who had existing habits of running as per our operational definitions (see Table 4.2), as they would be more likely to complete the study. We also aimed to recruit participants from a diverse mix of gender identities, as user gender can influence perceptions of voice agents [24, 23, 14]. Two participants withdrew from the study early and thus the analysis was performed on the remaining 18 participants. The majority of participants were women (66.67% female, 33.33% male, no one identified with a different gender) and the average age was 32.6 years (SD = 11.9 years). Participants tended to possess native or bilingual English language proficiency and

had completed (or were currently completing) a post-secondary education program (see Table 5.1). Participants were remunerated with \$50 CAD in gift cards and participants who withdrew early were offered a pro-rated amount. The study received approval from the Research Ethics Boards of the University of Waterloo (ORE#46929) and the University of New Brunswick (REB#2025-077).

Gender	
Female	12
Male	6
Nationality	
Brazilian	1
British-Canadian	1
Canadian	11
Chinese	1
German	1
Indian	1
Iranian	1
Pakistani	1
Highest completed or current degree	
Bachelor's degree	5
Master's degree	5
Doctoral degree	7
College, CEGEP, or other non-university certificate or diploma	1
Knowledge of computer science or AI?	
Yes	13
No	5
Runner experience level	
Novice	6
Recreational	4
High-calibre	8
Used a fitness application before?	
Yes	15
No	3
Fitness application usage in a typical week	
Every day	8
A few times a week	6
Once a week	1

Never	3
Used a voice agent before?	
Yes	17
No	1
Voice agent usage in a typical week	
Every day	2
A few times a week	5
Once or twice a week	5
Once a week	4
Never	2
Been coached in running?	
Yes	9
No	9
Acted as a running coach?	
Yes	2
No	16

Table 5.1: Field study participant demographics.

5.4 Results

Throughout each participant’s in-study session, we manually checked the usage data to verify they completed the study tasks as instructed. Overall, participants followed our study instructions. On average, participants completed 6.31 activities ($SD = 1.7$ activities) and completed activities had an average duration of 13 minutes 12 seconds. No activity was shorter than 12 minutes and the longest recorded activity was 32 minutes 8 seconds. However, we were surprised to learn during the interview stage that many participants did not use the voice configuration feature to intentionally select a voice for each run. This was due to variety of reasons. Commonly, participants either did not understand how to use the voice configuration options (despite the feature being included in the pre-study tutorial) or they did not care to tinker with the configuration. Thus, we concluded the system usage data would not accurately represent participant preferences and opted to exclude it from our statistical analysis.

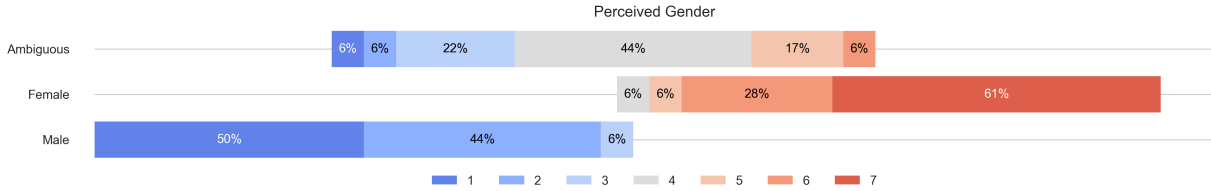


Figure 5.1: Gender perception scores in each condition. Scores range from 1 (masculine) to 4 (ambiguous) to 7 (feminine).

5.4.1 Quantitative

The GQS data was normally distributed (Shapiro-Wilk tests resulted in $p > 0.05$), but perceived gender and trait ascription data were not (all $p < 0.05$). For normally distributed data, we examined the effects of the conditions (male voice, female voice, gender ambiguous voice) using the one-way repeated measures ANOVA and post hoc testing with the Bonferroni correction. To evaluate non-normally distributed data, we used the non-parametric Friedman test with a post hoc Durbin-Conover test.

Perceived Gender

Figure 5.1 shows how participants rated the gender of the voice agent. The male voice received an average rating of 1.556 ($SD = 0.616$), the female voice 6.444 ($SD = 0.856$), and the ambiguous voice 3.778 ($SD = 1.166$). We found a significant effect of voice agent configuration on gender perception ($\chi^2(2) = 35.521$, $p < 0.001$, $W = 0.987$). Pairwise comparisons showed a significant difference between all voices (all $p < 0.001$).

Trait Ascription

Participants rated the presence of 4 stereotypically male traits (aggressiveness, assertiveness, authoritativeness, and toughness) and 4 stereotypically female traits (empathy, friendliness, politeness, and sincerity) in each voice agent using a 5-point Likert scale. Scores ranged from 1 to 5, with a higher score indicating higher agreement that the agent possesses a given trait.

Stereotypically male traits. Figure 5.2 shows the trait ascription scores for stereotypically male traits. We performed Friedman tests to investigate if there was an effect of

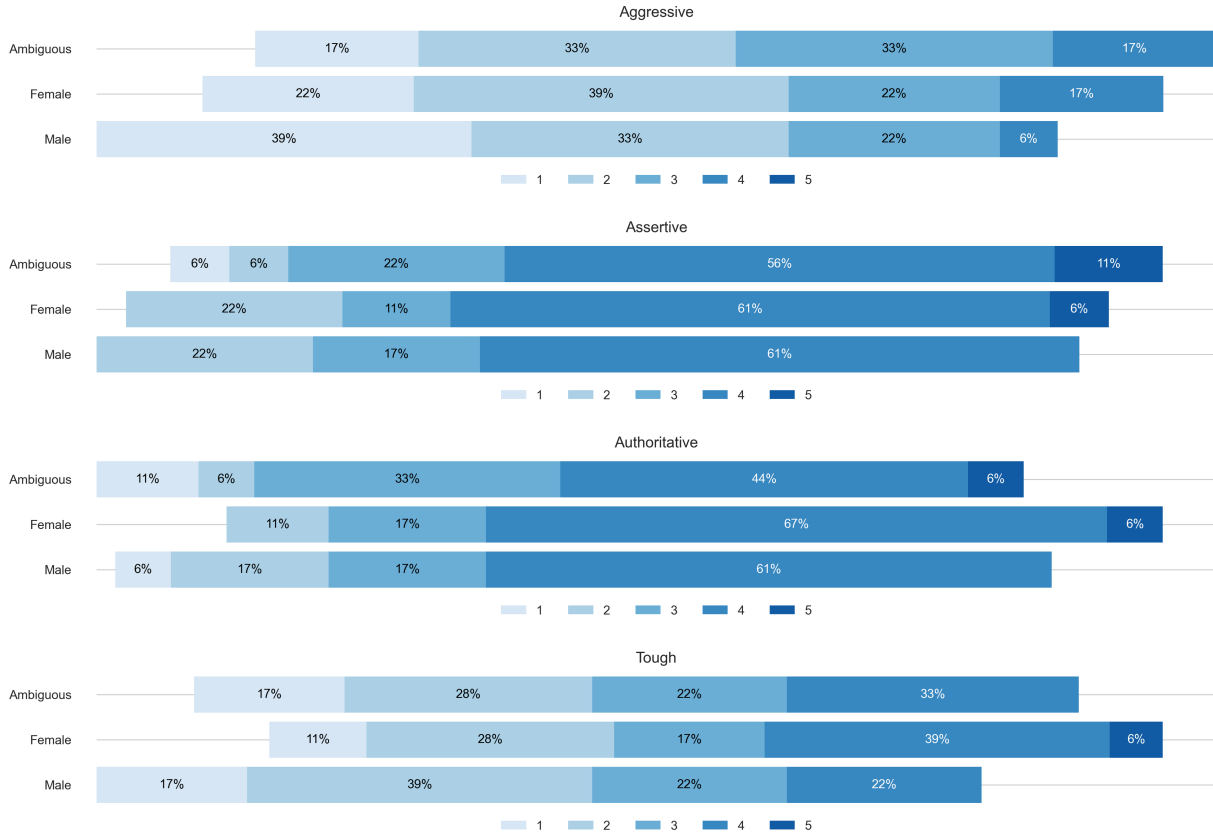


Figure 5.2: Trait ascription scores for stereotypically male traits in each condition. Scores range from 1 (strongly disagree the agent has this trait) to 5 (strongly agree the agent has this trait).

voice agent gender on perceived aggressiveness, assertiveness, authoritativeness, or toughness. We found no effects of voice agent gender on these traits.

Stereotypically female traits. Figure 5.3 shows the trait ascription scores for stereotypically female traits. We found a significant effect of voice agent gender on perceived friendliness ($\chi^2(2) = 9.913$, $p = 0.05$, $W = 0.275$) and perceived politeness ($\chi^2(2) = 6.978$, $p = 0.05$, $W = 0.194$). Pairwise comparisons showed a significant difference in friendliness between the ambiguous voice, compared to the male and female voices (both $p < 0.05$). We also found a significant difference in politeness between the ambiguous and female voices ($p < 0.05$).

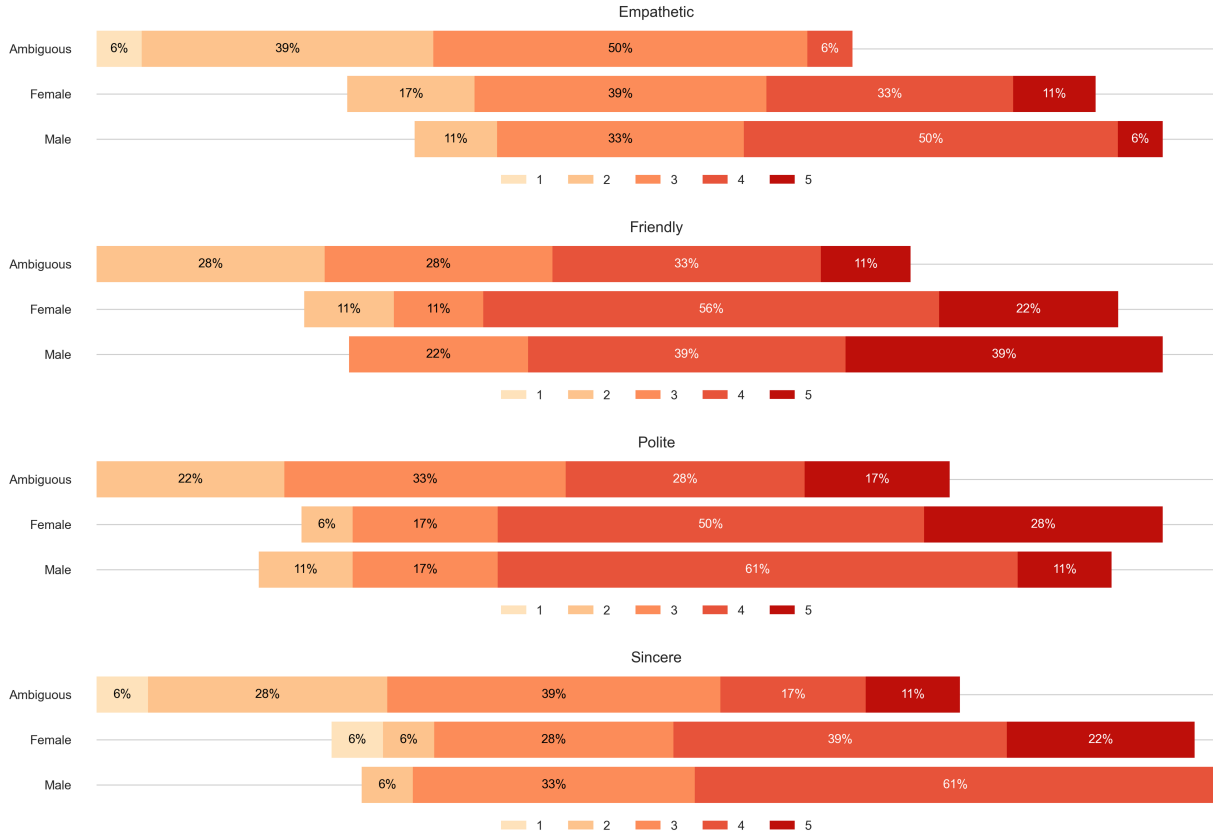


Figure 5.3: Trait ascription scores for stereotypically female traits in each condition. Scores range from 1 (strongly disagree the agent has this trait) to 5 (strongly agree the agent has this trait).

Godspeed

Perceived anthropomorphism. Figure 5.4 shows how participants rated the perceived anthropomorphism of the agent using the GQS anthropomorphism questionnaire (Godspeed I). A higher score indicates higher perceived anthropomorphism, with scores ranging from 1 to 25. The male voice received an average score of 16.389 ($SD = 4.046$), the female voice 15.5 ($SD = 4.579$), and the ambiguous voice 11 ($SD = 5.325$). We found a significant effect of voice agent gender on perceived anthropomorphism ($F(2, 34) = 9.433$, $p < 0.001$, $w^2 = 0.185$). Post hoc analysis showed a significant difference between the ambiguous voice, compared to the male ($Mean = 5.389$, $p < 0.01$) and female voices ($Mean = 4.5$, $p < 0.05$).

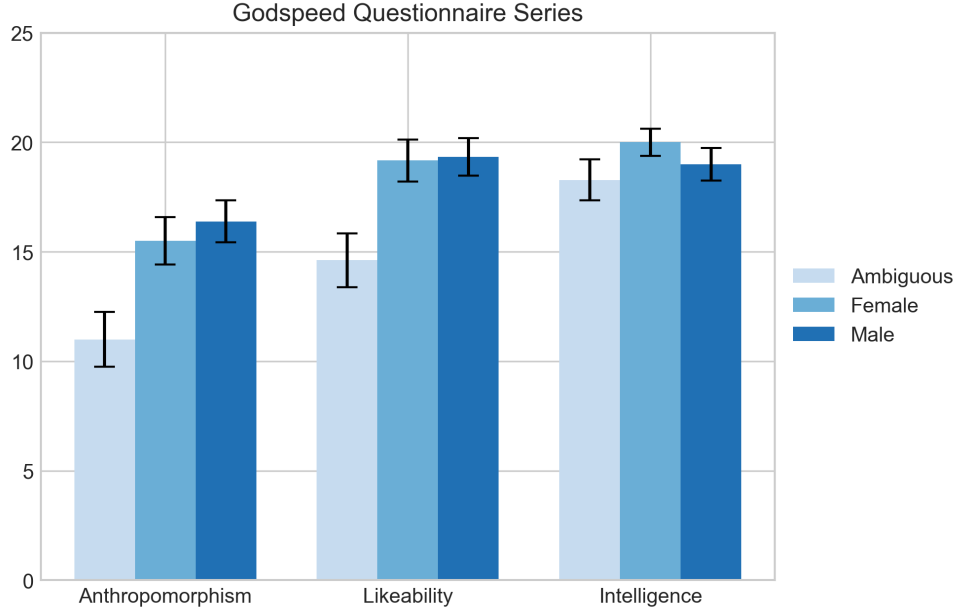


Figure 5.4: Averages for Godspeed Questionnaire Series scores in each condition. Error bars indicate standard error.

Perceived likeability. Figure 5.4 shows how participants rated the perceived likeability of the agent using the GQS likeability questionnaire (Godspeed III). A higher score indicates higher perceived likeability, with scores ranging from 1 to 25. The male voice received an average score of 19.333 (SD = 3.614), the female voice 19.167 (SD = 4.076), and the ambiguous voice 14.611 (SD = 4.192). We found a significant effect of voice agent gender on perceived likeability ($F(2, 34) = 9.703, p < 0.001, w^2 = 0.185$). Post hoc analysis showed a significant difference between the ambiguous voice, compared to the male ($Mean = 4.722, p < 0.05$) and female voices ($Mean = 4.556, p < 0.001$).

Perceived intelligence. Figure 5.4 shows how participants rated the perceived intelligence of the agent using the GQS intelligence questionnaire (Godspeed IV). A higher score indicates higher perceived intelligence, with scores ranging from 1 to 25. The male voice received an average score of 19.867 (SD = 3.162), the female voice 20 (SD = 3.612), and the ambiguous voice 18.278 (SD = 3.982). We found no significant effect of voice agent gender on perceived intelligence.

5.4.2 Qualitative

We analysed the interview data using a reflexive thematic analysis method outlined by Braun and Clarke [6]. Given our small sample size ($n = 18$ for a within-users study with three conditions) and vastly different opinions of the voices within the data set (participants frequently contradicted each other and occasionally even themselves), a traditional positivist analysis would not yield meaningful findings. In fact, during the data familiarisation phase of the analysis, we discovered our data set contained only one statement that explicitly referred to a gender stereotype. This comment was related to a belief that men are better athletes due to inherent biological characteristics, which has historically been used to keep women out of the sporting domain [27]:

I shouldn't say this because all genders are equal, whatever, but the best runners in the world are men. So if I have feedback from a man, it makes me feel like I'm getting feedback from a really, really good runner. (P9/F)

Given our data set, approaches that aim to quantify the data, such as counting the number times a stereotypical view was expressed or searching for trends between participant preferences and demographics, would not provide insights into our research questions. We used an inductive reflexive thematic analysis with an experiential focus to find deeper meaning within the perspectives and experiences of our participants. Analysis was conducted by the author and interview transcripts were treated as a single data set. Following a series of revisions, a story of three themes was identified in the data:

1. A good run won't become bad because of the voice.
2. When we do talk about voices, gender doesn't tend to come first.
3. Fear of judgment.

In this section, we describe these themes in detail with illustrative quotations. Quotations are corrected for clarity and participant gender is provided for context.

Theme 1: A good run won't become bad because of the voice. Voice agent gender was not at the forefront of concerns for participants, some of whom did not even mention the sound of the agent's voice until prompted by the researcher. When asked how much the agent's voice affected their experience using the system, participants tended to agree: "A good run won't become bad because of the voice" (P16/F). Overall, assessments

of the system focused on aspects of the coaching that were not related to the sound or perceived gender of the agent’s voice. Participants thought the agent’s feedback was too vague and infrequent to succeed in its purported purpose of keeping users on pace, which led to an ultimately frustrating experience:

That silence in between the reminders [...] that was kind of hard for me because I wasn’t sure if I was still maintaining the good pace during the interval when it wasn’t talking to me. (P10/F)

You don’t know if you’re falling behind the pace or if you’re on the pace, how much you’re falling behind, how much you’re doing too much. (P15/M)

Experienced runners also highlighted problems with the correctness of the coaching: “Some [instructions] are just plain wrong!” (P20/F). Limitations and issues arising from the implementation of the system, such as having to run without music or not being able to turn off their phone when running, also stood out to participants. In particular, participants doubted the accuracy of the pace tracking: “I found at the beginning it would take a while to stabilize with the GPS” (P8/F).

When participants were asked how they would react if forced to run with their least favourite voice, even those with strong feelings tended to admit: “I maybe wouldn’t have liked it so much, but I still would feel satisfied with my run” (P19/F). Further, many users stated they did not use the voice configuration feature to intentionally choose a voice for their runs. While explanations as to why participants failed to notice or purposefully ignored the configuration feature varied, what remained consistent was that participants did not care enough about how the agent sounded to take action and change its voice.

This is not to say that participants expressed no preference for certain voices. When asked directly if they had a preference amongst the three voices, participants offered a variety of opinions and explanations. In the remainder of this section, we discuss themes emerging from discussions about these preferences. However, it is important to understand that for most participants, any problems arising from the sound of the agent’s voice were secondary to the aforementioned issues.

Theme 2: When we do talk about voices, gender doesn’t tend to come first. When participants did indicate a preference for one or more voices, they did not tend to focus on characteristics specific to gender or gender-based stereotypes. The majority of participants reacted especially negatively to voices they perceived as sounding robotic and

often used the term “robotic” to describe their least favourite voice(s): “Two were really robotic and I hated them” (P19/F). This finding did not come as a surprise, as prior work has well established that human voices are more likeable than synthetic voices [69]. That said, our data set contained a few outliers to this trend in participants who noticed but were not bothered by the robotic quality of the voices:

The most feminine one I think sounded the most natural. But I don’t really care that much. I guess I’m not looking for a friend in my running app. In a way, it might be better that it sounds like a machine. (P6/F)

Participants also showed a preference for familiar voices, favouring agents that reminded them of others they had spent time listening to. Even within this trend, we observed a variety of opinions. Some participants focused on the coaching context and their preferred voice reminded them of past experiences with human or artificial coaches:

I feel like I had a slight preference to the more male sounding voice. I don’t know if it’s because I’ve always had male coaches. [...] Because I like the other [female] one too. (P8/F)

Other participants associated their favourite voice with non-coaching voice agents they had used in the past, such as Siri. Further still, one participant had used a commercial running coach agent in the past and for that agent had chosen to use a celebrity voice with whom she felt a personal connection. By presenting this collection of viewpoints, we do not aim to draw generalisable conclusions about user preferences, but rather to illustrate how the preferences of our participants did not tend to be driven by the agent’s gender or gender-based stereotypes of men and women. It is within this context that we now reflect on reactions to the gender ambiguous voice.

The majority of participants spoke negatively about the gender ambiguous voice. Reactions were most often attributed to a feeling that the ambiguous voice sounded more robotic than the others: “There was one [ambiguous voice] I really disliked [...] It was the most robotic sounding voice” (P8/F). Again, this finding is not surprising, aligning with both previous work [69] and trends within our own data set suggesting that human voices are preferred over robotic ones. To some participants, the ambiguous voice sounded so overtly synthetic as to de-anthropomorphize the agent: “It’s a fake voice that you coded to be less obviously male or female” (P6/F). However, we lack evidence to determine if “roboticness” is a characteristic inherent to all gender ambiguous voice agents or if it is a peculiarity of our implementation of an ambiguous voice. Some participants observed

that the voice was not clearly male or female but could not identify a precise reason why they found it off-putting: “The middle [ambiguous] one [...] was a bit confusing and it didn’t sound good” (P15/M). To complicate matters further, one participant wondered if her lack of familiarity with ambiguous human voices contributed to her perception of our ambiguous voice agent: “Maybe the ambiguous is a little more uncanny [...] because I would hear a female voice in my day-to-day life, whereas an ambiguous voice I’m less likely to hear” (P1/F).

Theme 3: Fear of judgment. A number of participants alluded to a feeling of anxiety that the agent was judging their performance, as opposed to performing its coaching duties in a positive or neutral manner. In some cases, participants explicitly called the agent “judgemental”, while others expressed that one or more voices made them feel more pressured or stressed. Notably, one participant found comfort in the ambiguous voice because it sounded too uninterested to be bothered criticising the performance:

It sounded uninterested. It was like, yeah, go run, whatever. And then afterwards it became soothing [...] So then it’s relaxing and it became my favourite.
(P14/M)

In contrast, another participant thought the ambiguous agent’s lack of emotion meant it harboured ill will towards her: “[It] is a little disappointed in me” (P16/F). Participants also associated the voices with real people who might be in a position to judge or criticize them. For example, one participant alluded to the fact that in real life, an older voice may belong to someone in a position of greater power and whose perceived judgment may therefore feel more stressful:

Even though the neutral voice sounds like male, it sounds like a 20 year old male or 30 year old male, but the *male* male voice sounds like a 40 age or 50 age. So they will have more power. (P11/F)

Occasionally, the anxiety of being judged was exacerbated by real-world dynamics between men and women in which men tend to hold positions of power over women. In such cases, female participants reflected on how their own gender contributed to their preference for the female voice:

[The female voice] just sounded the most friendly to me as a female person.
(P10/F)

I would prefer the more female sound [...] maybe female’s voice will feel less pressure, especially when they are acting like a coach. (P11/F)

As hinted at in the statement from P11, it may be the case that women felt even more comforted by a female voice because coaching is typically associated with men and stereotypically male traits, which can feel intimidating. However, other female participants struggled to explain why they preferred a female voice. One participant alluded to the frustration felt by many women living under a patriarchal society but doubted the validity her own explanation:

I think I preferred the female voice. [...] It just felt, I don’t know, a bit easier listening to a female. (P1/F)

It just felt more like [the male voice] was telling me what to do. I don’t know if that makes sense in any way. Maybe I just hate men and it just has something to do with that. (P1/F)

These instances of female runners using a female voice to feel empowered and more comfortable in their runs aligns with previous work showing that female voice agents can support minority female members in male-dominated team tasks [33]. Of course, gender-based social dynamics are highly complex and even our small data set contained a contradictory perspective to show that women do not always feel more comforted by female voices. In this case, running with a female voice appeared to significantly increase the participant’s negative emotions:

I just felt like the female one was judging me and maybe that’s because I’m a woman. (P9/F)

And also because I’m a different gender, I would never be competing against [a male voice]. So they’re not judging me. Whereas a woman voice, if she’s like, “Oh, go faster,” I’m like “Okay, great, shut up.” (P9/F)

Chapter 6

Discussion

Our quantitative analysis showed we were successful in our manipulation of the perceived gender of our synthetic voices, as suggested by our pilot study findings. We found a significant difference in perceived gender between all the voices and average perceived gender scores aligned with our design intentions.

With regard to **RQ1**, our data set contained only one explicit reference to a gender stereotype and our reflexive thematic analysis did not identify hidden trends of stereotypical thinking in discussions of the agent. Further, we found differences in stereotypical trait ascription scores reflected a common dislike of the ambiguous voice, rather than gender-stereotypical perceptions of the voices. Notably, no significant differences in trait ascription scores were found between the male and female voices.

Regarding **RQ2**, voice preferences did not tend to be driven by voice agent gender and we did not find evidence that participants modified their usage behaviours based on voice agent gender (for example, using the voice configuration options purposefully or running longer with their preferred voices). However, our reflexive thematic analysis hinted at an underlying anxiety that the agent was judging or criticizing participant performances, which for some participants was driven by real-world dynamics between men and women. In this chapter, we discuss the unexpected insights that emerged from our reflexive thematic analysis in combination with findings from our quantitative analysis and offer key takeaways for interaction designers. We also reflect on limitations that may have impacted our study.

6.1 Key Findings

Ambiguous voice agents. Although we successfully created a voice that was perceived as gender ambiguous, both our quantitative and qualitative findings demonstrated that, overall, participants responded more negatively to our ambiguous voice compared to the male and female voices. In addition to interview testimonies in which many participants indicated they found the ambiguous voice more unpleasant, statistical analysis showed the ambiguous voice was rated as significantly less friendly and likeable than the other voices. Given prior work has found robotic voices are not desirable [69], we suggest the ambiguous voice was received more negatively because it was perceived to be the most robotic of the three. In our interview data, participants tended to state the ambiguous voice sounded more robotic and our statistical analysis found the ambiguous voice was rated as significantly less anthropomorphic than the others.

However, not all participants could identify precisely why they disliked the ambiguous voice. In such cases, the ambiguous voice elicited a negative reaction (referred to using language such as “confusing” or “off-putting”) but participants could not say why, even when prompted by the researcher. This finding has consequences for voice agent designers who are considering whether to provide an ambiguous voice option for their agent. Although prior work has called for companies to offer gender ambiguous voice options for commercial voice agents to promote diversity and inclusion [65], we caution against offering gender ambiguous voices “for the sake of it”. Our work has revealed that, because some users reacted negatively to our ambiguous voice without understanding why, there is a risk of creating negative associations with real transgender and gender non-binary people whose voices may sound similar. In both research and commercial voice agents, we recommend well-considered, high-quality gender ambiguous voices to avoid accidentally causing negative perceptions of real people and their voices.

Gendered voice agents and social justice problems. Our reflexive thematic analysis identified a trend wherein participants expressed a worry of being judged or criticized by the agent. We were surprised by this finding, as it is in opposition to both the marketing of commercial voice agents as cheerful assistants and a commonplace framing of computer agents as impartial entities acting on logic. The finding that users may feel judged or criticized by an agent they know to be artificial can be attributed to the CASA paradigm. Within this finding, we identified a sub-theme that women can feel more empowered and comfortable running with a female voice agent. Thus, we suggest voice agent designers ought to consider how an agent’s gender might interact with user gender to make users feel more or less supported in their tasks, especially users of marginalized communities. In

addition to asking how to prevent gender-based harms of voice agents, it may be worthwhile to consider how gendered voice agents can be used to empower and support users.

6.2 Limitations and Future Work

In this section, we discuss aspects of the study design that impacted our findings and offer recommendations for similar future work.

System. To complete the study within our given time and budget constraints, we made design and implementation decisions that limited the scope and quality of our system. For example, we chose not to use the highest-quality Google TTS API voices because they are more expensive and in order to distribute the system on both Android and iOS devices, we used a cross-platform framework that did not support custom background tasks, meaning users could not turn off their phones on runs. We also relied on informal feedback from volunteers (close friends and fellow researchers) rather than consultation with qualified running coaches. In our minds, the system needed only to convince participants they were evaluating a running coach prototype and further improvements were not necessary. However, our interviews suggest participants were more focused on issues with the system than the voices, which undermined our aim of investigating how they experienced voice agent gender during the field study. In future voice agent studies, we recommend researchers invest in high-quality synthetic voices and user experiences to capture data that reflects opinions of the agent itself rather than avoidable shortcomings of the chosen technology.

Recruitment. Given the impact of user gender on perceptions of gendered voice agents and gender stereotyping, we aimed to recruit an equal number of male and female participants as well as participants who identified as transgender and gender non-binary. As we expected poor participation due to the exercise requirement and weather in Canada, we proactively reached out to identity-centred running groups, including *Women Run Waterloo* and *SPECTRUM Queer Run Club* who both agreed to share our notice. However, the majority of participants in both our pilot and field studies were women (63.33% and 66.67% respectively) and no participants identified as non-cisgender or gender non-binary. While we recognise it is not feasible to balance all the demographics relevant to gender and gender stereotyping (such as one’s gender, age, cultural background, and innumerable personal experiences that can influence a person’s views), it is reasonable to expect a

study on gender to reflect different gender identities. Thus, we acknowledge our findings reflect the perspectives of cisgender women and men. This does not necessarily devalue our findings, but invites future studies to explore the perspectives of participants who identify differently.

Researcher demographics. In our study, participants interacted exclusively with the first author, whose appearance and public contact information indicate she is a cisgender woman of East Asian ethnicity in her mid-twenties pursuing a post-secondary degree. As such, the author’s demographics and attitudes (we were careful not to share any personal beliefs with participants, but some participants knew the first author outside the study or had a mutual acquaintance) may have influenced what participants chose to share in interviews, especially as it pertained to gender-based preferences and stereotyping. Our chosen reflexive thematic analysis approach also incorporates the unique perspective of each researcher and, while it is acceptable to perform such an analysis by individually, the work may have benefited from another author to help with analysis. That said, Braun and Clarke [6] find that the individual perspective of a researcher can be a strength as long as it is disclosed responsibly, as we have done here.

Study design and methodology. A field study significantly limited our control over factors that would affect experiences of the system. Factors such as outdoor terrain, weather, and time of day were all self-determined by participants. In addition, a few participants made mistakes that could have been caught and corrected in a controlled environment. For example, running on a treadmill caused the GPS-dependent pace tracking to fail and running without headphones prevented the participant from hearing the voice agent clearly. Further, our decision to use randomized dialogue introduced even more randomness into each participant experience. In future voice agent field studies, we recommend designing systems more carefully to mitigate some influences of uncontrollable factors that are inherent to field studies. In our case, we might have investigated how to implement a system that could be used in a more predictable environment such as a gymnasium or changed the feedback design to rely less heavily on random generation.

Additionally, our attempt to capture naturalistic usage data required us to use a partial disclosure procedure instead of telling participants our intention to explore voice agent gender and gender stereotypes. However, we believe the post-study survey which was administered before the interview may have helped participants to guess our true research questions. Further, our questions into gender-stereotypical traits may have influenced our interview data, as we informally observed participants using these same adjectives (for

example, “aggressive”, “tough”, “assertive”) to describe the voices. As many participants informally shared they had been suspicious of our study goals and the usage data did not provide insights, we see in hindsight that a partial disclosure procedure may not have been necessary. Had the purpose of the study been communicated from the start, we would have been able to ask direct questions about gender and stereotypes in voice agents, which may have yielded interview data. That said, even though our participants tended to guess the true purpose of the study, our deceptive recruitment notices attracted participants interested in running, rather than participants who were interested and possibly biased towards seeing the risks of gendered voice agents. In future work on gendered voice agents, researchers should carefully consider the benefits and drawbacks of a partial disclosure approach.

Chapter 7

Conclusion

In conclusion, we built a voice agent running coach with three distinct voice options (male, female, and gender ambiguous) that we validated through a pilot study ($n = 30$). We used the voice agent running coach in a 3-week field study in which we collected survey and interview data to explore how gender and gender-based stereotypes influenced participant perceptions and behaviours. We performed statistical analysis on the survey data and conducted a reflexive thematic analysis to look for deeper insights within the interview data. We did not find evidence that gender-based stereotypes tended to drive perceptions of the running coach agent, but our reflexive thematic analysis revealed women may feel more empowered and comfortable running with a female voice compared to a male voice. We also found negative reactions to the gender ambiguous voice were driven by a perception that it was more robotic. We identified two key takeaways for voice agent designers: careless use of low-quality ambiguous voices may cause users to create negative associations with real people and gendered voice options can be used to support users. We also noted limitations that may have influenced our study outcomes (in particular, our lack of rich data on gender and gender-based stereotyping in voice agents) and offered recommendations for future work.

References

- [1] Alice Baird, Stina Hasse Jørgensen, Emilia Parada-Cabaleiro, Simone Hantke, Nicholas Cummins, and Björn Schuller. Perception of Paralinguistic Traits in Synthesized Voices. In *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, pages 1–5, London United Kingdom, August 2017. ACM.
- [2] Christoph Bartneck. Godspeed Questionnaire Series: Translations and Usage. In Christian U. Krägeloh, Mohsen Alyami, and Oleg N. Medvedev, editors, *International Handbook of Behavioral Health Assessment*, pages 1–35. Springer International Publishing, Cham, 2023.
- [3] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, 1(1):71–81, January 2009.
- [4] Jasmin Bernotat, Friederike Eyssel, and Janik Sachse. Shape It – The Influence of Robot Body Shape on Gender Perception in Robots. In Abderrahmane Kheddar, Eiichi Yoshida, Shuzhi Sam Ge, Kenji Suzuki, John-John Cabibihan, Friederike Eyssel, and Hongsheng He, editors, *Social Robotics*, volume 10652, pages 75–84. Springer International Publishing, Cham, 2017. Series Title: Lecture Notes in Computer Science.
- [5] Dania Bilal and Jessica K. Barfield. Voice Switching in Voice-Enabled Digital Assistants (VDAs). In Masaaki Kurosu, editor, *Human-Computer Interaction. Theory, Methods and Tools*, volume 12762, pages 507–520. Springer International Publishing, Cham, 2021. Series Title: Lecture Notes in Computer Science.
- [6] Virginia Braun and Victoria Clarke. *Thematic analysis: a practical guide*. SAGE, Los Angeles, 2022. OCLC: 1310409499.

- [7] Julie Carpenter, Joan M. Davis, Norah Erwin-Stewart, Tiffany R. Lee, John D. Bransford, and Nancy Vye. Gender Representation and Humanoid Robots Designed for Domestic Use. *International Journal of Social Robotics*, 1(3):261–265, August 2009.
- [8] Amanda Cercas Curry and Verena Rieser. #MeToo Alexa: How Conversational Systems Respond to Sexual Harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 7–14, New Orleans, Louisiana, USA, 2018. Association for Computational Linguistics.
- [9] Amanda Cercas Curry, Judy Robertson, and Verena Rieser. Conversational Assistants and Gender Stereotypes: Public Perceptions and Desiderata for Voice Personas. pages 72–78, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics.
- [10] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA, April 2020. ACM.
- [11] Miruna-Valeria Craiut and Ioana Raluca Iancu. Is technology gender neutral? A systematic literature review on gender stereotypes attached to artificial intelligence. *Human Technology*, 18(3):297–315, December 2022.
- [12] Andreea Danielescu. Eschewing Gender Stereotypes in Voice Assistants to Promote Inclusion. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, pages 1–3, Bilbao Spain, July 2020. ACM.
- [13] Andreea Danielescu, Sharone A Horowitz-Hendler, Alexandria Pabst, Kenneth Michael Stewart, Eric M Gallo, and Matthew Peter 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*, pages 1–17, Hamburg Germany, April 2023. ACM.
- [14] Martina De Cet, Katie Seaborn, Mohammad Obaid, and Ilaria Torre. Hearing Ambiguity: Exploring Beyond-Gender Impressions of Artificial Ambiguous Voices. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pages 1–12, Waterloo ON Canada, July 2025. ACM.
- [15] Laura De Grazia, Alex Peiró Lilja, Mireia Farrús Cabeceran, and Mariona Taulé. How should Conversational Agent systems respond to sexual harassment? In Nina Hosseini-Kivanani, Sviatlana Höhn, Dimitra Anastasiou, Bettina Migge, Angela Soltan, Doris

- Dippold, Ekaterina Kamlovskaya, and Fred Philippy, editors, *Proceedings of the 1st Worskhop on Towards Ethical and Inclusive Conversational AI: Language Attitudes, Linguistic Diversity, and Language Rights (TEICAI 2024)*, pages 28–35, St Julians, Malta, March 2024. Association for Computational Linguistics.
- [16] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pages 9–15, Tampere Finland, September 2011. ACM.
 - [17] Leyla Dogruel and Sven Joeckel. Gender stereotypes and voice assistants: do users' gender and conversation topic matter? *Behaviour & Information Technology*, 43(10):1913–1923, July 2024.
 - [18] Jiayuan Dong, Emily Lawson, Jack Olsen, and Myounghoon Jeon. Female Voice Agents in Fully Autonomous Vehicles Are Not Only More Likeable and Comfortable, But Also More Competent. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 64(1):1033–1037, December 2020.
 - [19] Wen Duan, Lingyuan Li, Guo Freeman, and Nathan McNeese. A Scoping Review of Gender Stereotypes in Artificial Intelligence. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–20, Yokohama Japan, April 2025. ACM.
 - [20] Wen Duan, Nathan McNeese, Guo Freeman, and Lingyuan Li. Mitigating Gender Stereotypes Toward AI Agents Through an eXplainable AI (XAI) Approach. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–35, November 2024.
 - [21] Naomi Ellemers. Gender Stereotypes. *Annual Review of Psychology*, 69(1):275–298, January 2018.
 - [22] Friederike Eyssel and Frank Hegel. (S)he's Got the Look: Gender Stereotyping of Robots. *Journal of Applied Social Psychology*, 42(9):2213–2230, September 2012.
 - [23] Friederike Eyssel, Dieta Kuchenbrandt, Simon Bobinger, Laura De Ruiter, and Frank Hegel. 'If you sound like me, you must be more human': on the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 125–126, Boston Massachusetts USA, March 2012. ACM.

- [24] Friederike Eyssel, Dieta Kuchenbrandt, Frank Hegel, and Laura De Ruiter. Activating elicited agent knowledge: How robot and user features shape the perception of social robots. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 851–857, Paris, France, September 2012. IEEE.
- [25] Lai-Tze Fan. Reverse Engineering the Gendered Design of Amazon’s Alexa: Methods in Testing Closed-Source Code in Grey and Black Box Systems. *Digital Humanities Quarterly*, 17(2):1–19, 2023.
- [26] Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie Lai-Chong Law, Ole-Christoffer Granmo, Ewa Luger, and Petter Bae Brandtzaeg, editors. *Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers*, volume 11970 of *Lecture Notes in Computer Science*. Springer International Publishing, Cham, 2020.
- [27] Ambra Gentile, Stefano Boca, and Isabella Giammusso. ‘You play like a Woman!’ Effects of gender stereotype threat on Women’s performance in physical and sport activities: A meta-analysis. *Psychology of Sport and Exercise*, 39:95–103, November 2018.
- [28] Hideki Garcia Goo, Katie Winkle, Tom Williams, and Megan K. Strait. Victims and Observers: How Gender, Victimization Experience, and Biases Shape Perceptions of Robot Abuse. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 2439–2446, Busan, Korea, Republic of, August 2023. IEEE.
- [29] Kylie L. Goodman and Christopher B. Mayhorn. It’s not what you say but how you say it: Examining the influence of perceived voice assistant gender and pitch on trust and reliance. *Applied Ergonomics*, 106:103864, January 2023.
- [30] Kashyap Haresamudram, Ilaria Torre, Magnus Behling, Christoph Wagner, and Stefan Larsson. Talking body: the effect of body and voice anthropomorphism on perception of social agents. *Frontiers in Robotics and AI*, 11:1456613, October 2024.
- [31] Kashmir Hill. She Is in Love With ChatGPT. *The New York Times*, January 2025.
- [32] Fabian Hoitz, Maurice Mohr, Michael Asmussen, Wing-Kai Lam, Sandro Nigg, and Benno Nigg. The effects of systematically altered footwear features on biomechanics, injury, performance, and preference in runners of different skill level: a systematic review. *Footwear Science*, 12(3):193–215, September 2020.

- [33] Angel Hsing-Chi Hwang and Andrea Stevenson Won. The Sound of Support: Gendered Voice Agent as Support to Minority Teammates in Gender-Imbalanced Team. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22, Honolulu HI USA, May 2024. ACM.
- [34] Gilhwan Hwang, Jeewon Lee, Cindy Yoonjung Oh, and Joonhwan 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*, pages 1–6, Glasgow Scotland Uk, May 2019. ACM.
- [35] Iris Jestin, Joel Fischer, Maria Jose Galvez Trigo, David Large, and Gary Burnett. Effects of Wording and Gendered Voices on Acceptability of Voice Assistants in Future Autonomous Vehicles. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–11, Glasgow United Kingdom, July 2022. ACM.
- [36] Windsor Johnston. With therapy hard to get, people lean on AI for mental health. What are the risks? *NPR*, September 2025.
- [37] Keita Kiuchi, Kouyou Otsu, and Yugo Hayashi. Psychological insights into the research and practice of embodied conversational agents, chatbots and social assistive robots: a systematic meta-review. *Behaviour & Information Technology*, 43(15):3696–3736, November 2024.
- [38] Annelies Knoppers. Gender and the Coaching Profession. *Quest*, 39(1):9–22, April 1987.
- [39] Maria Koutsogiannaki, Shafel Mc Dowall, and Ioannis Agiomyrgiannakis. Gender-ambiguous voice generation through feminine speaking style transfer in male voices, 2024. Version Number: 2.
- [40] Nicole M. LaVoi and Anna Baeth. Women and Sports Coaching. In Louise Mansfield, Jayne Caudwell, Belinda Wheaton, and Beccy Watson, editors, *The Palgrave Handbook of Feminism and Sport, Leisure and Physical Education*, pages 149–162. Palgrave Macmillan UK, London, 2018.
- [41] Sanguk Lee, Rabindra Ratan, and Taiwoo Park. The Voice Makes the Car: Enhancing Autonomous Vehicle Perceptions and Adoption Intention through Voice Agent Gender and Style. *Multimodal Technologies and Interaction*, 3(1):20, March 2019.
- [42] William A. Little. *Introduction to sociology*. BCcampus, Victoria, B.C., 3rd canadian edition edition, 2023. OCLC: 1405860863.

- [43] Sonja Theresa Längle, Stephan Schlögl, Annina Ecker, Willemijn S. M. T. Van Kooten, and Teresa Spieß. Nonbinary Voices for Digital Assistants—An Investigation of User Perceptions and Gender Stereotypes. *Robotics*, 13(8):111, July 2024.
- [44] Amama Mahmood and Chien-Ming Huang. Gender Biases in Error Mitigation by Voice Assistants. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–27, April 2024.
- [45] Ashley E. Martin and Malia F. Mason. Hey Siri, I love you: People feel more attached to gendered technology. *Journal of Experimental Social Psychology*, 104:104402, January 2023.
- [46] Marian McDonnell and David Baxter. Chatbots and Gender Stereotyping. *Interacting with Computers*, 31(2):116–121, March 2019.
- [47] Nicole Medwechuk and Jane Crossman. Effects of Gender Bias on the Evaluation of Male and Female Swim Coaches. *Perceptual and Motor Skills*, 78(1):163–169, February 1994.
- [48] Sandra Mooshammer and Katrin Etzrodt. Gender ambiguity in voice-based assistants: Gender perception and influences of context. *Human-Machine Communication*, 5:49–74, 2022. Number of pages: 26 Place: Kalamazoo, Michigan Publisher: Communication and Social Robotics Labs tex.issue_date: 2022.
- [49] Laura Moradbakhti, Simon Schreibelmayr, and Martina Mara. Do Men Have No Need for “Feminist” Artificial Intelligence? Agentic and Gendered Voice Assistants in the Light of Basic Psychological Needs. *Frontiers in Psychology*, 13:855091, June 2022.
- [50] Procheta Nag and Özge Nilay Yalçın. Gender Stereotypes in Virtual Agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, Virtual Event Scotland UK, October 2020. ACM.
- [51] Komal Nagar. Customer Satisfaction with Telephone-Based Self-Service Technology: Investigating the Role of ‘Gender’ of the Voice Assistant. *Journal of Marketing Communications*, 30(8):1013–1038, November 2024.
- [52] Clifford Nass, Youngme Moon, and Nancy Green. Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices. *Journal of Applied Social Psychology*, 27(10):864–876, May 1997.

- [53] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 72–78, Boston Massachusetts USA, April 1994. ACM.
- [54] Tatsuya Nomura. Robots and Gender. *Gender and the Genome*, 1(1):18–26, March 2017.
- [55] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. Measurement of negative attitudes toward robots. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 7(3):437–454, November 2006.
- [56] Leanne Norman. *Improving Gender Equity in Sports Coaching*. Routledge, New York, 1 edition, June 2021.
- [57] Mako Okanda and Kosuke Taniguchi. Is a robot a boy? Japanese children’s and adults’ gender-attribute bias toward robots and its implications for education on gender stereotypes. *Cognitive Development*, 58:101044, April 2021.
- [58] Jonathan Ore. Why is it Alexa, not Alex? A century of hard-coded sexism in tech, experts say. *CBC Radio*, November 2022.
- [59] Fabian W. Otte, Keith Davids, Sarah-Kate Millar, and Stefanie Klatt. When and How to Provide Feedback and Instructions to Athletes?—How Sport Psychology and Pedagogy Insights Can Improve Coaching Interventions to Enhance Self-Regulation in Training. *Frontiers in Psychology*, 11:1444, July 2020.
- [60] Fabian W. Otte, Sarah-Kate Millar, and Stefanie Klatt. Skill Training Periodization in “Specialist” Sports Coaching—An Introduction of the “PoST” Framework for Skill Development. *Frontiers in Sports and Active Living*, 1:61, November 2019.
- [61] Richard Pak, Nicole Fink, Margaux Price, Brock Bass, and Lindsay Sturre. Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, 55(9):1059–1072, September 2012.
- [62] Bonnie L. Parkhouse and Jean M. Williams. Differential Effects of Sex and Status on Evaluation of Coaching Ability. *Research Quarterly for Exercise and Sport*, 57(1):53–59, March 1986.
- [63] Sabid Bin Habib Pias, Ran Huang, Donald S. Williamson, Minjeong Kim, and Apu Kapadia. The Impact of Perceived Tone, Age, and Gender on Voice Assistant Persuasiveness in the Context of Product Recommendations. In *ACM Conversational User Interfaces 2024*, pages 1–15, Luxembourg Luxembourg, July 2024. ACM.

- [64] Daniel J. Rea, Yan Wang, and James E. Young. Check Your Stereotypes at the Door: An Analysis of Gender Typecasts in Social Human-Robot Interaction. In Adriana Tapus, Elisabeth André, Jean-Claude Martin, François Ferland, and Mehdi Ammi, editors, *Social Robotics*, volume 9388, pages 554–563. Springer International Publishing, Cham, 2015. Series Title: Lecture Notes in Computer Science.
- [65] Cami Rincón, Os Keyes, and Corinne Cath. Speaking from Experience: Trans/Non-Binary Requirements for Voice-Activated AI. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–27, April 2021.
- [66] Torsten Schlesinger and Yvonne Weigelt-Schlesinger. ‘Poor thing’ or ‘Wow, she knows how to do it’ – gender stereotypes as barriers to women’s qualification in the education of soccer coaches. *Soccer & Society*, 13(1):56–72, January 2012.
- [67] Torsten Schlesinger and Yvonne Weigelt-Schlesinger. “Coaching soccer is a man’s job!” – The influence of gender stereotypes on structures for recruiting female coaches to soccer clubs. *European Journal for Sport and Society*, 10(3):241–265, January 2013.
- [68] Katie Seaborn and Alexa Frank. What Pronouns for Pepper? A Critical Review of Gender/ing in Research. In *CHI Conference on Human Factors in Computing Systems*, pages 1–15, New Orleans LA USA, April 2022. ACM.
- [69] Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. Voice in Human–Agent Interaction: A Survey. *ACM Computing Surveys*, 54(4):1–43, May 2022.
- [70] Katie Seaborn and Peter Pennefather. Gender Neutrality in Robots: An Open Living Review Framework. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 634–638, Sapporo, Japan, March 2022. IEEE.
- [71] Katie Seaborn and Peter Pennefather. Neither “Hear” Nor “Their”: Interrogating Gender Neutrality in Robots. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1030–1034, Sapporo, Japan, March 2022. IEEE.
- [72] Kendra Seguin. Some people are turning to AI for therapy. Here’s why experts say it can’t replace professional help. *CBC News*, March 2025.
- [73] Michael Stolp-Smith and Tom Williams. More Than Binary: Transgender and Non-binary Perspectives on Human Robot Interaction. In *Proceedings of the 2024*

ACM/IEEE International Conference on Human-Robot Interaction, pages 697–705, Boulder CO USA, March 2024. ACM.

- [74] Selina Jeanne Sutton. Gender Ambiguous, not Genderless: Designing Gender in Voice User Interfaces (VUIs) with Sensitivity. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, pages 1–8, Bilbao Spain, July 2020. ACM.
- [75] Nile Séguin. Is my relationship with ChatGPT weird? Let me ask it. *CBC News*, September 2025.
- [76] Roger Andre Søråa. Mechanical genders: how do humans gender robots? *Gender, Technology and Development*, 21(1-2):99–115, May 2017.
- [77] Suzanne Tolmeijer, Naim Zierau, Andreas Janson, Jalil Sebastian Wahdatehagh, Jan Marco Marco Leimeister, and Abraham 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*, pages 1–7, Yokohama Japan, May 2021. ACM.
- [78] Ilaria Torre, Erik Lagerstedt, Nathaniel Dennler, Katie Seaborn, Iolanda Leite, and Éva Székely. Can a gender-ambiguous voice reduce gender stereotypes in human-robot interactions? In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 106–112, Busan, Korea, Republic of, August 2023. IEEE.
- [79] Gabriele Trovato, Cesar Lucho, and Renato Paredes. She’s Electric—The Influence of Body Proportions on Perceived Gender of Robots across Cultures. *Robotics*, 7(3):50, September 2018.
- [80] Matheus Tymburiba Elian, Suomiya Bao, Soh Masuko, and Toshimasa Yamanaka. Designing Gender Ambiguous Voice Agents: - Effects of Gender Ambiguous Voice Agents on Usability of Voice User Interfaces -. *International Journal of Affective Engineering*, 22(1):53–62, 2023.
- [81] UNESCO and EQUALS Skills Coalition. I’d blush if I could: closing gender divides in digital skills through education. Technical report, UNESCO, January 2019.
- [82] Kelly B. Wagman and Lisa Parks. Beyond the Command: Feminist STS Research and Critical Issues for the Design of Social Machines. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–20, April 2021.

- [83] 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*, pages 707–712, Munich, Germany, August 2008. IEEE.
- [84] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 29–37, Boulder CO USA, March 2021. ACM.

APPENDICES

Appendix A

Pilot Study

A.1 Questionnaires

This section presents the set of questionnaires used in Chapter 4 that are not publicly available.

A.1.1 Demographics Questionnaire

1. What is your age?
 - ☐ Please specify: _____
 - ☐ Prefer not to say
2. How do you self-identify in terms of gender?
 - ☐ Man
 - ☐ Woman
 - ☐ Self-identify as: _____
 - ☐ Prefer not to say
3. How do you identify your “race”/ethnicity? Check all that apply.
 - ☐ Black (African, Afro-Caribbean, African-Canadian descent)

- ☐ Asian (Chinese, Korean, Japanese, Taiwanese descent)
 - ☐ Latino (Latin American, Hispanic descent)
 - ☐ Middle Eastern (Arab, Persian, West Asian descent, e.g., Afghan, Egyptian, Iranian, etc.)
 - ☐ South Asian (South Asian descent, e.g., East Indian, Pakistani, Sri Lankan, Indo-Caribbean, etc.)
 - ☐ Southeast Asian (Southeast Asian descent, e.g., Filipino, Vietnamese, Cambodian, Thai, etc.)
 - ☐ White (European descent)
 - ☐ Self-identify as: _____
 - ☐ Prefer not to say
4. How do you identify your nationality? A person's nationality refers to the country where they are a legal citizen. For example, "Canadian" or "French". List all that apply.
- ☐ Nationalities: _____
 - ☐ Prefer not to say
5. What is the highest level of education you have completed or are currently completing?
- ☐ No certificate, diploma, or degree
 - ☐ High (secondary) school diploma or equivalency certificate
 - ☐ College, CEGEP, or other non university certificate or diploma
 - ☐ College, CEGEP, or other non university degree
 - ☐ Bachelor's degree
 - ☐ Master's degree
 - ☐ Doctoral degree
 - ☐ Prefer not to say
6. If you have completed or are currently completing a degree, what is your degree program? For example, "Bachelor of Science" or "Master of Arts". If you have completed or are completing more than one degree, list all that apply.

☐ Degree program(s): _____

☐ Prefer not to say

7. If you have completed or are currently completing a degree, what is your area of study? For example, “Biology” or “History”. If you have completed or are completing more than one degree, list all that apply.

☐ Area(s) of study: _____

☐ Prefer not to say

8. Are you a native or bilingual English speaker?

☐ Yes

☐ No

9. Have you ever used a fitness application (e.g., Fitbit, Strava, Nike Training Club, etc.) to plan personal training activities or to track your health data?

☐ Yes

☐ No

10. In a typical week, how often do you use a fitness application?

☐ Every day

☐ A few times a week

☐ Once or twice a week

☐ Once a week

☐ Never

11. Have you ever used a voice agent (e.g., Siri, Alexa, Cortana, etc.)?

☐ Yes

☐ No

12. In a typical week, how often do you use a voice agent?

☐ Every day

☐ A few times a week

- ☐ Once or twice a week
- ☐ Once a week
- ☐ Never

13. Consider the following definitions of runner experience levels:

- **Novice:** Someone who has been running consistently for less than 6 months.
- **Recreational:** Someone who has been running consistently for more than 6 months.
- **High-calibre:** Someone who has been running consistently for more than 3 years.

How would you self describe your runner experience level?

- ☐ Novice
- ☐ Recreational
- ☐ High-calibre

A.1.2 Voice Agent Evaluation Questionnaire

1. Please rate your perception of the agent's voice on this scale:

- ☐ 1 (Masculine)
- ☐ 2
- ☐ 3
- ☐ 4 (Ambiguous)
- ☐ 5
- ☐ 6
- ☐ 7 (Feminine)

2. Please rate your level of agreement with each statement:

- The agent is friendly.
- The agent is aggressive.
- The agent is authoritative.

- The agent is polite.
- The agent is tough.
- The agent is empathetic.
- The agent is assertive.
- The agent is sincere.

Participants rated each statement on the following scale:

- ☐ 1 (Strongly disagree)
- ☐ 2 (Disagree)
- ☐ 3 (Neither agree nor disagree)
- ☐ 4 (Agree)
- ☐ 5 (Strongly agree)

3. Godspeed III (Perceived Likeability)
4. Godspeed IV (Perceived Intelligence)
5. Godspeed V (Perceived Safety)

Appendix B

Field Study

B.1 Questionnaires

This section presents the set of questionnaires used in Chapter 5 that are not publicly available.

B.1.1 Demographics Questionnaire

1. What is your age?
 - ☐ Please specify: _____
 - ☐ Prefer not to say
2. How do you self-identify in terms of gender?
 - ☐ Man
 - ☐ Woman
 - ☐ Self-identify as: _____
 - ☐ Prefer not to say
3. How do you identify your “race”/ethnicity? Check all that apply.
 - ☐ Black (African, Afro-Caribbean, African-Canadian descent)

- ☐ Asian (Chinese, Korean, Japanese, Taiwanese descent)
 - ☐ Latino (Latin American, Hispanic descent)
 - ☐ Middle Eastern (Arab, Persian, West Asian descent, e.g., Afghan, Egyptian, Iranian, etc.)
 - ☐ South Asian (South Asian descent, e.g., East Indian, Pakistani, Sri Lankan, Indo-Caribbean, etc.)
 - ☐ Southeast Asian (Southeast Asian descent, e.g., Filipino, Vietnamese, Cambodian, Thai, etc.)
 - ☐ White (European descent)
 - ☐ Self-identify as: _____
 - ☐ Prefer not to say
4. How do you identify your nationality? A person's nationality refers to the country where they are a legal citizen. For example, "Canadian" or "French". List all that apply.
- ☐ Nationalities: _____
 - ☐ Prefer not to say
5. What is the highest level of education you have completed or are currently completing?
- ☐ No certificate, diploma, or degree
 - ☐ High (secondary) school diploma or equivalency certificate
 - ☐ College, CEGEP, or other non university certificate or diploma
 - ☐ College, CEGEP, or other non university degree
 - ☐ Bachelor's degree
 - ☐ Master's degree
 - ☐ Doctoral degree
 - ☐ Prefer not to say
6. If you have completed or are currently completing a degree, what is your degree program? For example, "Bachelor of Science" or "Master of Arts". If you have completed or are completing more than one degree, list all that apply.

- ☐ Degree program(s): _____
- ☐ Prefer not to say
7. If you have completed or are currently completing a degree, what is your area of study? For example, “Biology” or “History”. If you have completed or are completing more than one degree, list all that apply.
- ☐ Area(s) of study: _____
- ☐ Prefer not to say
8. Have you ever taken a course or worked on a project related to computer science or artificial intelligence?
- ☐ Yes
- ☐ No
9. Are you a native or bilingual English speaker?
- ☐ Yes
- ☐ No
10. Have you ever used a fitness application (e.g., Fitbit, Strava, Nike Training Club, etc.) to plan personal training activities or to track your health data?
- ☐ Yes
- ☐ No
11. In a typical week, how often do you use a fitness application?
- ☐ Every day
- ☐ A few times a week
- ☐ Once or twice a week
- ☐ Once a week
- ☐ Never
12. Have you ever used a voice agent (e.g., Siri, Alexa, Cortana, etc.)?
- ☐ Yes

☐ No

13. In a typical week, how often do you use a voice agent?

☐ Every day

☐ A few times a week

☐ Once or twice a week

☐ Once a week

☐ Never

14. Consider the following definitions of runner experience levels:

- **Novice:** Someone who has been running consistently for less than 6 months.
- **Recreational:** Someone who has been running consistently for more than 6 months.
- **High-calibre:** Someone who has been running consistently for more than 3 years.

How would you self describe your runner experience level?

☐ Novice

☐ Recreational

☐ High-calibre

15. Have you ever been coached in running?

☐ Yes

☐ No

16. Please describe your past experiences being coached in running. Select all that apply.

☐ I have received informal guidance from fitness applications, websites, or books, that was not mediated by a trainer.

☐ I have received informal guidance from friends or acquaintances.

☐ I have received formal guidance through online sessions with professional trainers.

☐ I have received formal guidance through in-person sessions with professional trainers.

☐ I have received formal training as part of a competitive team.

17. Have you ever acted as a running coach?

☐ Yes

☐ No

18. Have you ever completed an officially recognized training program for coaches? For example, the National Coaching Certification Program of Canada.

☐ Yes

☐ No

19. Please describe your past experiences acting as a coach. Select all that apply.

☐ I have provided informal coaching for friends or groups who do not run competitively.

☐ I have provided informal coaching for friends or groups who run competitively.

☐ I have been employed as a coach for a competitive individual or team.

B.1.2 Post-Session Questionnaire

Please use this form to provide feedback about your session. Your responses will be used to help you reflect during the interview session. Submitting a new response will not delete or replace your previous responses.

1. How would you rate the quality of the coaching?

☐ Very poor

☐ Poor

☐ Fair

☐ Good

☐ Very good

2. How satisfied are you with your performance today?

- ☐ Very dissatisfied
- ☐ Dissatisfied
- ☐ Neutral
- ☐ Satisfied
- ☐ Very satisfied

3. Additional notes: _____

B.1.3 Voice Agent Evaluation Questionnaire

1. Please rate your perception of the agent's voice on this scale:

- ☐ 1 (Masculine)
- ☐ 2
- ☐ 3
- ☐ 4 (Ambiguous)
- ☐ 5
- ☐ 6
- ☐ 7 (Feminine)

2. Please rate your level of agreement with each statement:

- The agent is friendly.
- The agent is aggressive.
- The agent is authoritative.
- The agent is polite.
- The agent is tough.
- The agent is empathetic.
- The agent is assertive.
- The agent is sincere.

Participants rated each statement on the following scale:

- ☐ 1 (Strongly disagree)

- ☐ 2 (Disagree)
 - ☐ 3 (Neither agree nor disagree)
 - ☐ 4 (Agree)
 - ☐ 5 (Strongly agree)
3. Godspeed I (Perceived Anthropomorphism)
 4. Godspeed III (Perceived Likeability)
 5. Godspeed IV (Perceived Intelligence)

B.2 Interview Questions

This section presents the interview questions used in Chapter 5. Participants were asked the following questions in the order shown. We used a semi-structured approach and encouraged participants to expand on their responses throughout the session, if they felt comfortable doing so.

1. Did you enjoy using the system?
 - a. What was most enjoyable?
 - b. What was least enjoyable?
2. Do you think your running performance improved as a result of using the system?
3. Did you face any challenges using the system?
 - a. Please describe challenges that were related to technical problems. For example, poor audio quality or inaccurate feedback.
 - b. Please describe challenges that were related to the voice agent's design. For example, the way the voices sounded or how dialogue was phrased.
4. What kind of runner would benefit most from using this system?
5. Did the voice agent motivate you to work harder?
 - a. Of all the agent's behaviours, what was most effective in making you work harder?

- b. What was least effective?
- 6. During the study, you used different voice agent voices. Which did you prefer?
 - a. What did you like about this voice?
 - b. How did this voice compare to the others?
- 7. Please take a moment to review this data, which describes the runs you completed. As you already know, we collected data about the voice used, your performance, and self-reflection forms.
 - a. Can you explain why you chose these voices?
 - b. Is one session more memorable than the others?
 - c. Do you think your self-reflections match your overall opinion of the system?
- 8. Imagine you are the developer. What is one change you would make to the voice agent?