2022

A sociophonetic analysis of female-sounding virtual assistants

Alyssa Allen

A Sociophonetic Analysis of Female-Sounding Virtual Assistants

by

Alyssa Allen

Thesis

Submitted to the College of Arts and Sciences

Department of English Language & Literature

Eastern Michigan University

in partial fulfillment of the requirements

for the degree of

MASTER OF ARTS

in

English Linguistics

Thesis Committee:

Eric Acton, PhD

T. Daniel Seely, PhD

May 24, 2022

Ypsilanti, Michigan

Abstract

As conversational machines (e.g., Apple's Siri and Amazon's Alexa) are increasingly anthropomorphized by humans and viewed as active interlocutors, it raises questions about the social information indexed by machine voices. This thesis provides a preliminary exploration of the relationship between human sociophonetics, social expectations, and conversational machine voices. An in-depth literature review (a) explores human relationships with and expectations for real and movie robots, (b) discusses the rise of conversational machines, (c) assesses the history of how female human voices have been perceived, and (d) details social-indexical properties associated with F0, vowel formants (F1 and F2), -ING pronunciation, and /s/ center of gravity in human speech. With background context in place, Siri and Alexa's voices were recorded reciting various sentences and passages and analyzed for each of the aforementioned vocal features. Results suggest that sociolinguistic data from studies on human voices could inform hypotheses about how users might characterize conversational machine voices and encourage further consideration of how human and machine sociophonetics might influence each other.

Keywords: sociolinguistics, sociophonetics, HMC, conversational machines, language and gender

Table of Contents

Abstract ii

List of Tables v

List of Figures vi

Chapter 1: Introduction 1

1.1 Introduction 1

1.2 Relevant Framework 2

1.3 Purpose and Research Questions 4

1.4 Remaining Chapters 4

1.5 Definition of Terminology 5

Chapter 2: Literature Review 6

2.1 Movie Robots and Real Robots 6

2.2 The Rise of Conversational Agents 13

2.3 Social Expectations Related to Female Voices 22

2.4 Relevant Sociophonetic Features 32

Chapter 3: An Analysis of Siri and Alexa's Vocal Traits 45

3.1 Overview of Stimuli 45

3.2 General Methodology 46

3.3 Study 1: F0 47

3.4 Study 2: Formants (F1 and F2) 58

3.5 Study 3: /s/ 63

3.6 Study 4: -ING 66

3.7 Summary 70

Chapter 4: General Discussion 71

Chapter 5: Summary, Considerations, and Looking Ahead 78

5.1 Summary 78

5.2 Considerations 79

5.3 Looking Ahead 79

References 82

Appendices 98

Appendix A: Stimuli Sentences for Chapter 3 99

Appendix B: Rainbow Passage 100

Appendix C: The Boy Who Cried Wolf 101

Appendix D: Stimuli for Study 1.2 102

List of Tables

Table Page

1 Siri: Difference Between F0 of First and Final VI Token (Hz) 49

2 Alexa: Difference Between F0 of First and Final VI Token (Hz) 50

3 Mixed-Effects Model: Average F0 of Utterances 54

4 Differences in Average F1 values per VI between Siri and Alexa 59

5 Differences in Average F2 values per VI between Siri and Alexa 60

6 Center of Gravity (Hz) for Word-Initial /s/ in "The Rainbow Passage" for Siri and Alexa ... 65

7 [iŋ] Versus [in] in Read -ING Final Words for Siri and Alexa 68

List of Figures

Figure Page

1 Comparison of F0 Values per VI Category and Overall F0 Average for Each Virtual Assistant 51

2 Comparison of Overall F0 Values per Stimulus Sentence for Each Virtual Assistant 52

3 Comparison of Results from Study 1.1's Average F0 Values per Sentence and Study 1.2's Average F0 Values per Prompt-Type for Each Virtual Assistant (Siri and Alexa) 53

4 F1-F2 Vowel Space Mapping of Siri and Alexa 61

Chapter 1: Introduction

Chapter 1 aims to provide context necessary for the work. Section 1.1 provides an overview of female-sounding virtual assistants (such as Apple's Siri and Amazon's Alexa) in our society and why a sociophonetic analysis of such voices would be a valuable focus of study. Section 1.2 then provides details on the relevant framework needed in order to make sense of the arguments laid out in this work. This chapter also presents primary research questions, an overview of content, and key definitions.

1.1 Introduction

The popularity and accessibility of virtual assistants such as Apple's Siri (referred to as Siri) and Amazon's Alexa (referred to as Alexa) has contributed to a growing interest in human-machine communication as an area of study, particularly as it pertains to voice-user-interface (VUI). The field of human-machine interactions has, to this point, largely focused on how humans align their speech styles when speaking with a machine-driven conversational agent (Zellou et al., 2021; Bell, 2003; Branigan et al., 2011; Cohn & Zellou, 2019). Insights from previous research provide evidence that humans are beginning to perceive virtual assistants as active participants in the conversation, as interlocutors (Zellou et al., 2021). This concept is discussed further in section 1.2.

If humans perceive conversational machines as active interlocutors, the disembodied voices of virtual assistants should also be studied in order to fully understand human-machine interactions. For example, the perceived gender of the machine voice may impact the way a human user converses with the machine. One could also imagine that a virtual assistant's dialect may have an impact on the human user's perception of the assistant. Nearly all popular virtual

assistant voices are programmed to have a default female-sounding voice, which likely on its own affects human users' perceptions of the role of females and assistants.

As a clarifying point when discussing "female voices," a female voice does not necessarily indicate the speaker was assigned female at birth (Zimman, 2017). It is especially important to note as machine voices are discussed that the devices do not have a biological or anatomical sex. The devices were never born. Therefore, gender perception and societal expectations become a key consideration of machine voice research as human listeners assign gender and other social or personality traits to the devices (e.g., polite, kind, wise, funny, non-threatening). This work focuses on how socio-indexical properties associated with human phonetic qualities relate to voice-driven conversational machines. The phonetic qualities of virtual assistants offer a starting point in a much larger conversation about the potential impact machine voices can have on user perceptions of a conversation, the machine, and social roles more generally.

Gaining a more nuanced linguistic understanding of a machine-driven voice's vocal qualities can provide insights into humans' approach to conversations with machines. From a technology perspective, this research can also be used by programmers and developers to create more inclusive and dynamic technology.

1.2 Relevant Framework

This thesis relies on the computers are social actors (CASA) framework. A high-level approach to this framework, as laid out by Nass et al. (1994), aims to show that a machine exhibiting "a limited set of characteristics associated with humans provides sufficient cues to encourage users to exhibit behaviors and make attributions toward computers that are nonsensical when applied to computers but appropriate when directed at other humans" (p. 72).

Treating computers as active interlocutors can elicit social behaviors, such as forms of politeness or gender stereotype perpetuation, from humans. These social behaviors seem to occur even when the human consciously knows that the machine has no sense of "self," gender, or other inherently human motivations (Nass et al. 1994). Subfields of CASA include human-machine communication (HMC), human-computer interaction (HCI), human-robot interaction (HRI), human-agent interaction (HAI), and media effects (Gambino et al., 2020).

Following CASA, further research has been conducted to better understand how humans apply gender stereotypes to machines that are encoded with minimal gender cues. Nass et al. (1997) conducted an experiment where human subjects were assisted by computers on four distinct tasks: a preparation session, a tutoring session, a testing session, and an evaluation session. During the tutoring, testing, and evaluation sessions, subjects were randomly assigned a male-or-female-sounding machine. In the tutoring session, subjects were given facts by the computer on the topics "computers" and "love and relationships." The testing computer then administered a test on the material. The evaluation computer then reviewed how well subjects did on the test and how well the computers did at guiding the subjects. Nass et al. (1997) found that humans applied the following gender-based stereotypes to disembodied machine voices, at minimum (discussed in more depth in section 2.3): "evaluation from males is more valid than evaluation from females," "dominance is more desirable in men than women," and (for male-voiced evaluators) "women know more about subjects that are typically regarded as 'feminine'" (p. 874). Additional research supports the finding that humans apply gender stereotypes to male and female voices (Mullennnix et al., 2003).

CASA provides a basis for this work to apply sociophonetic trends found in human language to disembodied machine voices. For example, if humans apply gender-based

stereotypes to a voice-based conversational agent, the phonetic qualities of that machine's voice should be perceived in ways that relate to findings from phonetic studies conducted on human voices. Exploring the phonetic qualities of popular virtual assistants through the lens of sociophonetics can lead to a more nuanced understanding of human-machine interactions in regards to both the human and the machine interlocutors.

1.3 Purpose and Research Questions

The purpose of this thesis is to explore the relationship between human sociophonetics, social expectations, and conversational machine voices. In order to gain a more nuanced understanding of the socio-indexical characteristics indexed by disembodied virtual assistant voices, this work will also include a phonetic analysis of particular vocal qualities (F0, F1, F2 and /s/ realization, and -ING realization) of Siri and Alexa. Questions of particular interest include the following: What socio-indexical characteristics historically associated with human female voices are applicable to Siri and Alexa's voices? How do phonetic properties of female-sounding virtual assistant voices contribute to creating a particular kind of persona for the devices?

1.4 Remaining Chapters

The remainder of this thesis is organized as follows. Chapter 2 presents a review of previous research on how movie robots shape human beliefs about real robots, the evolution of machine-driven conversational agents, perceptions of “the female voice” throughout history, and social characteristics indexed by certain phonetic features. Chapter 3 details a series of experiments showing how Siri and Alexa produce the aforementioned phonetic features, providing insights into which social characteristics may be indexed by the machine voices.

Subsections of Chapter 3 will focus on F0, F1 and F2, center of gravity of /s/, and -ING pronunciation. Chapter 4 consists of a general discussion, and Chapter 5 provides readers with a summary, considerations, and suggestions for future research.

1.5 Definition of Terminology

Within this thesis, the terms below are defined as follows:

- *Virtual assistant*: any machine-powered voice-user interface taking on the role of an assistant that does not have any physical representation or manifestation attached to it. Virtual assistants are completely operated by machine-learning capabilities and have the goal of providing humans with information or to perform simple tasks digitally.
- *Conversational agent*: any machine-learning device capable of conversing with a human in a written or spoken capacity. By this definition, virtual assistants are a type of conversational agent.
- *Machine bias*: information the machine program receives during initial programming or via user input that results in the virtual assistant speaking in a manner that perpetuates a potentially harmful human stereotype. Biases can also manifest in the form of increased misunderstandings of one dialect versus another.
- *Movie robot*: "an artificial entity that can sense and act as a result of a (real-world or fictional) technology and [has] played a role in at least one movie [or television show]" (Saffari et al., 2021, p. 3).

Chapter 2: Literature Review

Chapter 2 provides an in-depth literature review of material relevant to this work's main research questions and purpose, as identified in section 1.3. Section 2.1 focuses on the extent to which movie robots have influenced human perception of how a robot should behave in reality, including robot personas, female movie robots and sexism in the real world, and the cyclical relationship between creating movie robots and inventing real robots. Section 2.2 then discusses the rise of conversational agents (defined in section 1.5) as it pertains to historical context on automated speech recognition technology, human tendency to anthropomorphize machines, and potential issues with machine learning bias. With background context in place, section 2.3 assesses social expectations for female voices, including a historical understanding of the female voice as "noise" and social discourse expectations for females. Lastly, section 2.4 discusses specific phonetic features examined in this work. Phonetic features represented in section 2.4 will be the primary focus of experiments presented in Chapter 3.

2.1 Movie Robots and Real Robots

The development of virtual assistants like Siri and Alexa catalyzed widespread public adoption and regular use of VUI robot technology. As discussed in this chapter, it is plausible that movie robots depicted in science fiction film and television have influenced human expectations for and perceptions of real-world robot behavior. In some cases, human perceptions of robot capabilities may go beyond current actual robot functionality.

2.1.1 Relationship Between Humans and Movie Robots

2.1.1.1 Fictional Robot Inferiority. Movie robots are not limited by real-world technological capabilities. One result of the limitless potential for robot power is that some humans fear a robot takeover. That fear, in turn, powers many storylines involving robots taking

over human civilization (Bartneck, 2013). As a means to off-set these concerns, a set of suggested rules was developed by science fiction author Issac Asimov in 1991 to define boundaries for fictional human-robot relations—formally called Asminov's laws (Bartneck, 2017). There has been some debate on how to interpret and follow the guidelines, but the rules set a foundation for the way many robots were depicted in science fiction films, television, books, etc (Salge & Polani, 2017). Asimov's laws state that a robot (a) "may not injure a human being or, through inaction, allow a human being to come to harm," (b) "must obey orders given it by human beings except where such orders would conflict with the First Law," and (c) "must protect its own existence as long as such protection does not conflict with the first or second law" (Anderson, 2003, p. 477-478). These rules served as a foundation for science fiction creators to develop storylines that either reinforced these rules or created tension by breaking the rules. Asimov's laws also position robots as inferior or in a servant-like position to humans. People have therefore grown accustomed to this dynamic via science fiction robots and have come to expect a functioning robot to act accordingly, in real life and in movies.

Psychologically, movie robots often either want to be human or are unaware that they are not human (Bartneck, 2013). This dynamic furthers the framing of robots as inferior to humans regardless of their potential to be all powerful. Movie robots may also be portrayed as having difficulty comprehending emotions or common sense but also as highly skilled in computational problem solving (Bartneck, 2013). A lack of emotional intelligence may be one quality that begins to shape human expectations and perceptions of robot behavior. For example, one could imagine a task-driven scenario in the real world as one that a humanoid robot might be able to do better than a human, but it becomes more unsettling to imagine a humanoid robot becoming a therapist.

2.1.1.2 Movie Robot Personas. Movie robots may be characterized by viewers based on features of their physical manifestation. For example, judgments made regarding the social character of the movie robot (e.g., good or evil, masculine or feminine, sweet or sassy, subservient or dominant) may occur based on if it has a humanoid or mechanical body. Viewers make these same assumptions about human characters based on physical appearance in films and television.

In the same way, voice has an impact on the persona of movie robots. The voice of *The Jetsons* robot maid, Rosey, has a mechanical element to her voice (e.g., minimal pitch fluctuation, short sentence structures; Barbera & Hanna, 1962). Compare that with a "fembot" from the film *Austin Powers: International Man of Mystery*, whose voice sounds more human. Both have distinct and highly contrastive physical manifestations: Rosey as a matronly (even called homely in the show), mechanical, cartoon machine versus Austin Powers' fembots as hyper-sexualized, seductive, blonde females—played by human female actors. Fembot voices are breathy and lower in pitch, playing into a seductive female stereotype (Roach, 1997). Each voice adds to a target stereotype and furthers the characterizations of the movie robot. Judgments on non-human characters such as robots are possible because of real-world stereotypes that exist within a society. Stereotypes can be defined as "a gestalt view of individual perception that emphasizes the notion that certain traits, characteristics, or prototypes are more central and important in organizing our perceptions of other people than other traits" (Tay et al., 2014, p. 76).

Robots' voices can also be used to further a plotline. Movie robots may adopt an increasingly low pitch to signify that the machine is corrupt or has evil intentions (Humphry & Chesher, 2021). This vocal shift happens with VIKI (a female-sounding machine) in iRobot

(Proyas, 2004). Throughout the film, VIKI is a system monitoring an army of robot servants. When it becomes clear that VIKI has ambitions to reprogram the robot servants to kill all humans, her voice distorts, becoming lower in pitch and less human-sounding (Brown, 2019). Shifting the voice from natural to mechanical potentially leads human viewers to distinctly perceive VIKI as non-human. This vocal shift also happens with HAL 9000 from Stanley Kubrick's *2001: A Space Odyssey* where HAL speaks with an increasingly slower cadence throughout the film and eventually distorts its voice to be even slower while lowering its pitch (Kubrick, 1970).

As it relates to the focus of this paper, virtual assistants such as Siri and Alexa have pitch values aligned with averages among human females (Allen, 2022). This design choice to make Siri and Alexa sound more human and less similar to distorted voices of evil movie robots may be motivated by a desire to make the virtual assistants sound less threatening. Interestingly, before Siri was introduced to the public, the development team brainstormed calling Siri "Hal" and using the tagline "HAL's back --- but this time he's good" (Sterling, 2020).

2.1.2 Gender Expectations on Female Movie Robots and Real World Implications

The following section describes how female-presenting movie robots relate to three female-focused stereotypes that exist in the United States: (a) sexualized females cause trouble, (b) women are objects, and (c) women should be maternal. Female-presenting movie robots are of particular interest considering this thesis' later discussion of Siri and Alexa.

One of the first female movie robots was Maria in the 1927 German expressionist silent film *Metropolis*. In the film, Maria was created by a male who wanted to resurrect a deceased loved one (Brown, 2019). While initially desired, the robot Maria became dangerous and caused chaos among workers (Lang, 1927). Since this film, sexualized female robots have typically been

characterized as dangerous in some capacity. As noted above, the fembots in the film *Austin Powers: International Man of Mystery* are presented as hyper-sexual, feminine, seductive females; and their ultimate goal is to kill Austin Powers (Roach, 1997). This highlights a common US societal stereotype that sexualized women should not be trusted—as seen in a femme fatale trope.

Female movie robots are also commonly seen as objects or inferior to men in movies and television. In the HBOMax show *WestWorld*, female robots run the brothel in a town where humans can do whatever they please to the human-like robot women without repercussions. The show has influenced the real-world sex robot industry, where men expect a similar level of abusive fantasy without repercussion (Brown, 2019). In the movie *Her,* the female-voiced robot Samantha is treated as an object. Her role is to be emotionally supportive and even emotionally intimate with the human user, but the robot is literally carried around in a "a pocket-sized device resembling a vintage cigarette case or compact mirror" (Kidd, 2021, p. 61). So even though Samantha is expected to be an emotionally invested companion, she is still reduced to an object when not needed. This dynamic positions female companions as objects. Cases of a female-voiced emotional companion in movies being treated as inferior or non-human may be seen as giving humans in the real-world permission to treat sex workers and other women who are emotional companions the same way (Brown, 2019). Even though the female movie robots are non-human, the social dynamic presented via the characters roles influences the way actual human females (particularly sex workers) are treated (Brown, 2019).

A third stereotype attached to female movie robots is that the robot should act as a maternal figure. In Disney's movie *Smart House,* the female-voiced disembodied machine PAT assumes the position of mother in the family living in the house. In PAT's effort to protect her

family, she eventually becomes an overprotective anti-hero that locks the family inside to keep them away from harm (Burton, 1999). When the shift happens, PAT's voice distorts and becomes more machine-like, indicating that she can no longer be trusted. PAT's actions mirror VIKI's narrative as mentioned in 2.1.1.2. These examples underpin the impact of voice on human perception of a machine's intentions and character. A female-sounding movie robot can be accepted as a maternal figure, until the voice distorts. At that point, even if the intentions are maternal in nature, the robot is no longer trust-worthy.

As explored above, there are many types of female personas. When discussing female-sounding machine voices, it is helpful to consider the personality traits indexed by the machine voice in order to understand what the human perception of that device may be.

2.1.3 Connecting Fantasy and Reality

2.1.3.1 Human Perceptions of Robots. In the real world, female-sounding robots are perceived as more trustworthy than male-sounding robots in domesticated roles. Humphry and Chesher (2021) discuss Siri and Alexa as sounding like middle-class women in their assessment of virtual assistants in the home but do not establish this premise linguistically. Later in this thesis, there will be a phonetic analysis which demonstrates that Siri and Alexa's voices have vocal qualities consistent with being perceived as women and potentially as middle-class. Humphry and Chesher (2021) argue that using voices for virtual assistants that embody the persona of a domestic middle-class female results in devices that families are comfortable having in their homes daily without fear of surveillance. Virtual assistant voices were designed to create trust and be transparent in intention (Phan, 2017). Understanding how voice impacts user perceptions of trustworthiness, potential threat, and transparency can lead to design choices that may help combat human fear of a robot takeover (a fear potentially fueled by the limitlessness of

movie robots as discussed in 2.1.1) or issues with privacy or security (Liang & Lee, 2017). It's important to note that current socially intelligent robots are clearly distinguishable from humans in terms of capability and competence (Breazeal, 2003).

Building on human personality and interactional expectations of robots based on movie robots, humans also commonly anthropomorphize inanimate objects and apply social models to them (discussed further in section 2.2), particularly when those inanimate devices can converse (Breazeal 2003). For example, gender stereotypes have been found to be applied to occupational robots. One study showed humans prefer a male-voiced robot for a security-related job, while a female-voiced robot was deemed more suited for a caretaker position (Tay et al., 2014). Additionally, the study showed evidence for people rating their interactions with a robot more favorably if the robot's voice and gender matched occupational expectations. Social expectations and stereotypes for humans are seemingly transferred onto robots, even in cases where the robot's programming and capabilities are the same regardless of voice. Studies have shown humans make judgments about the "humanness" of a robot based on gender stereotypes as well (Borau et al., 2021; Sørra, 2017; Bernotat et al., 2021; Belanche et al., 2021).

2.1.3.2 Technology Development. With human perceptions of robots in mind, it is also valuable to consider that fictional robots do not adhere to real world technological limitations. Science fiction writers can create characters and robots based on real or fictional technology. Machine learning engineers can pull from those creations as inspiration and develop technology to make it possible in reality. In turn, science fiction writers can find inspiration in new technology to create new fictional capabilities (Saffari et al., 2021). A cycle is thus created connecting fictional robot creation and real-world robotic engineering.

As discussed previously, a robot's voice helps create the persona of the movie robots. As voice-driven virtual assistants became more ubiquitous, voice-user-interface developers have also recognized that the machine voice impacts user perception of the device (Humphry & Chesher, 2021). For example, the general target perception of virtual assistants should be a female helper who is "helpful without being obsequious, warm without being sexual, intelligent without being arrogant" (Humphry & Chesher, 2021, p. 1980). The "speak when spoken to" nature of virtual assistants also aligns with perceptions of domestication and non-threatening behavior (Humphry & Chesher, 2021). The turn-taking restrictions on when a virtual assistant speaks also aligns with the laws of robotics described in section 2.1.1, positioning robots are servants to humans. As discussed by Humphry and Chesher (2021) studies have shown humans apply social gender stereotypes to robots, influencing the decision to make many virtual assistant voices female-sounding (Bergen, 2016; Strengers & Nicholls, 2017; Phan, 2019; Woods, 2018; West et al., 2019). Virtual assistants have been positioned as a gendered helper. A female-sounding voice leaves users with an impression that the robot matches the assistant-occupation and that the device will be harmless (Humphry & Chesher, 2021). These personality traits should be kept in mind for section 2.4 and Chapter 3.

2.2 The Rise of Conversational Agents

Building on section 2.1's examination of the connection between movie robots and real-world robots, this section discusses the origins of speech recognition technology and conversational agents (as defined in section 1.5), previous research on human-machine relationships, and potential machine bias issues. Each of these factors influences how humans have grown to understand the role of virtual assistants in their lives, whether consciously or otherwise. Establishing background research in the aforementioned areas allows for a more

nuanced and contextualized evaluation of socio-indexical properties conveyed by virtual assistant voices, as explored in Chapter 3.

2.2.1 The Journey Towards Conversational Agents

2.2.1.1 A Brief History of Speech Recognition Technology. Before more complex natural-language processing technology or conversational agents were created, automated speech recognition technology (ASR) was developed. ASR can be defined as a computer transcribing spoken language into readable text in real or near-real time (Topaz et al., 2018). The ASR system is able to transcribe spoken language due to the signals given by soundwaves (Jurafsky & Martin, 2020). For example, "The first machine that recognized speech was a toy from the 1920s. 'Radio Rex' … was a celluloid dog that moved (by means of a spring) when the spring was released by 500 Hz acoustic energy" (Jurafsky & Martin, 2020). In other words, audio input is provided to the ASR system, the system then interprets the soundwaves to trigger a response. In the case of Radio Rex, the response was non-linguistic. Text to Speech (TTS) technology does the opposite of ASR. It uses a text-based input and maps those written symbols to soundwaves—enabling machines to speak (Jurafsky & Martin, 2020). Both types of technology are required for a virtual assistant (as defined in 1.5) to operate. For the purposes of this paper, ASR and TTS technology is discussed as back-end advancements for conversational technologies. It is not focused on conversational agents attached to robotic or humanoid forms. This section is meant to provide a brief history and is by no means exhaustive.

With ASR and TTS capabilities, human-computer conversations (to varying degrees) became possible. Command and control is one type of interaction model in which the user provides the system with a command or asks a question within the limits of pre-defined phrases known by the machine (Paek & Chickering, 2007). Command and control became more widely

used with the advent of voice dialing on mobile devices and is potentially the simplest form of human-machine communication. Another form of ASR was dictation systems. In the 1960s, IBM launched IBM Shoebox—a digital ASR tool—that recognized 16 words and digits (Mutchler, 2018). Then, in the early 1990s, dictation systems became more advanced. By 1993, dictation systems could understand up to 40,000 words but could only be used in restricted environments, such as a quiet office with headphones on the user (Rudnicky et al., 1994). Dragon Dictate, the first ASR product for consumers was also launched in the 1990s by a company called Dragon and cost $6,000 (Mutchler, 2018). Dragon Dictate illustrates how costly and inaccessible this technology was to the public still.

Interactive voice response capabilities (IVR) soon followed dictation systems. An IVR is "an automated telephone system that combines pre-recorded messages or text-to-speech technology with a dual-tone multi-frequency (DTMF) interface to engage callers, allowing them to provide and access information without a live agent" (IBM Cloud Education 2021, What is interactive voice response?, para. 1). Moviefone is an example of a successful IVR in which users could call the system, provide their zip code, and hear a list of available movies and showtimes nearby (IBM Cloud Education 2021). Virtual assistants came later in the evolution of ASR and TTS technology, with Apple introducing Siri in 2011, Google with Google Now in 2012, Microsoft with Cortana in 2013, and Amazon with Alexa in 2014. Of note, all of the aforementioned virtual assistants were released with female-sounding voices as the default setting. From 2015 onward there has been a boom in voice-driven ASR products made available for consumers at a relatively affordable price (Mutchler, 2018).

2.2.1.2 The Rise of Voice-Driven Virtual Assistants. It was predicted that by 2021, 123 million people would regularly use voice-driven virtual assistants (Choi & Drumwright, 2021,

citing Petrock, 2021). It was also predicted that Amazon would continue to dominate the market share on smart speakers through 2021 with about 70% of smart speakers owned in the US being the Amazon Echo device (Perez, 2020). In 2019, 147 million smart speakers were sold globally and 157 million smart speakers existed in US homes (Sterling, 2020). Smart speaker adoption is relevant in the case of Amazon Echo since each device is equipped with Alexa.

As of 2020, there were over half a billion Siri-capable devices in existence and about half a billion Google Assistant users (Perez, 2020). This widespread adoption and growth demonstrates how accessible and cost-effective these ASR and TTS devices are compared to earlier iterations discussed in section 2.2.1.2. Mobile devices made it more convenient for people to use speech as the input modality with conversational agents because users could more easily speak to the device while being able to perform other tasks at the same time. Mobile devices also allowed ASR technology to be scaled to smaller and cheaper forms (Paek & Chickering, 2007). Without widespread adoption, it is possible human relationships with conversational agents would not be as prevalent today as section 2.2.2 discusses.

2.2.2 Human Relationships and Perceptions of Conversational Agents

Following an overview of the types of ASR technology humans have been exposed to over the years and an understanding of how prevalent voice-driven virtual assistants are in US households, human relationships with and perceptions of these devices can be contextualized and further explored.

2.2.2.1 Anthropomorphism and CASA. As discussed in section 1.2, the CASA paradigm posits that humans exhibit behaviors towards computers that would be typically found in human-to-human conversation. This framework suggests that humans are capable of viewing computers as active interlocutors (Nass et al., 1994). Research built on the CASA paradigm

usually also makes reference to a human's ability to anthropomorphize inanimate objects. Studying anthropomorphism in cases of human-computer communication can help uncover the extent to which humans view machines as active interlocutors. Previous research has shown that people use similar politeness cues with computers as they would with humans and can view computers as cooperative teammates and confidants (Choi & Drumwright, 2021; Nass et al., 1996; Reeves & Nass, 1996; Kleinberg, 2018).

Virtual assistants are often placed within the user's home (as mentioned in 2.2.1.2) and are a part of users' daily lives. Within the first year of Alexa being on the market, half a million users told "her" they loved her and another half a million users proposed marriage to Alexa (Schweitzer et al., 2019, citing Risley, 2015; Murdoch, 2016). Matrimony and love are not concepts often associated with computers, yet users are interacting with Alexa in a way that mirrors a typical human-human interaction. That said, developers of Alexa did make design choices (such as naming the device with a gendered name and personality) which played a role (on top of its language capabilities) in the likelihood that anthropomorphism would occur (Schweitzer et al., 2019). As anthropomorphism is considered in this work, it should be noted that Siri and Alexa's voices initially came from human voice actors who were White adult females (Henry, 2022; Vincent, 2021). Potential user awareness of voice origin may further lead to perceptions of the device voices being human-like. More in-depth research would be needed to fully understand how much users associate Siri and Alexa's voices with the women whose voices were recorded for the devices.

In a report by Druga et al. (2017), findings show that children made assumptions about a conversational agent's intelligence based on personal conversations with the machines. For example, if discussing a topic the child knows a lot about and the machine did not know all of

the answers to the child's questions, the child ranked the machine as less intelligent. Younger children even used he/she pronouns when discussing their impression of the machines, indicating they at least partially comprehended the human-machine conversation similarly to a human-human conversation.

In a study done by Schweitzer et al. (2019), participants describe their relationship with their virtual assistant (e.g., Siri). The responses fall into three categories: (a) device is servant, (b) device is partner, and (c) device is master. Participants in the first category view the agent as a "nice, friendly, helpful, reliable person with a ready-to-please character, who acts professionally, as well as somewhat subserviently, and remotely" (p. 703). Members of this group acknowledge that the agent only spoke when spoken to by the user first. Interestingly enough, feedback from this first group also includes feedback such as

> I think this woman is stupid and naive. She simply gave me too many stupid and senseless replies for me to think of her as an intelligent creature. Some of the interactions were wrong, and, for whatever reason, Siri simply can't process certain tasks. Perhaps the technology isn't sufficiently developed as yet (Schweitzer et al., 2019, p. 704).

Responses such as this indicate anthropomorphism has occurred. Siri does not have an anatomical sex or assigned gender, but this participant connects the device with being a woman. The participant also seems to acknowledge the device is not human yet refers to its intellect as if it was human. Participants in Group 2 describe Siri as "an attractive and likeable character" (Schweitzer et al., 2019). Members of Group 2 also seem to project human qualities and features onto the disembodied voices. Participants in Group 3 view themselves as beholden to the

goodwill of the device. These participants are more likely to believe the device would take on a life of its own and have a sense of distrust with the machine (Schweitzer et al., 2019).

In a content analysis study on human perception of Alexa (Purington, 2017), 587 product reviews on Amazon's website were analyzed for degree of personification, degree of sociability, integration, technical qualities and issues, and household characteristics. Users who implemented the female pronoun “her” when discussing Alexa in the review were more likely to describe Alexa with a higher level of sociability. Users who implemented “Amazon Echo” instead of “Alexa” and object pronouns typically described their interactions with the assistant as less sociable. Users who had children or other people in the home were more likely to anthropomorphize Alexa (Purington, 2017). These trends point to a relationship being built with the machine, particularly when there are younger users in the household.

Other research on the anthropomorphism of conversational machines has uncovered that humans are able to view virtual assistants as friends, conversational partners, or even family members (Choi & Drumwright, 2021, citing Purington et al., 2017; Rhee & Choi, 2020; Wang et al., 2020; Zhao & Rau, 2020). Humans have the ability to anthropomorphize virtual assistants and the CASA paradigm opens the door to delve into the relationships humans can build with machines. If humans view virtual assistants as active interlocutors, then socio-indexical characteristics conveyed by a machine's voice provide a new perspective on the interaction.

2.2.2.2 Linguistic Alignment to Conversational Machines. Many studies have focused on human-machine conversations that demonstrate how humans align their speech with the device (Zellou et al., 2021; Bell, 2003; Branigan et al., 2011; Cohn & Zellou, 2019). Building on the anthropomorphism and CASA paradigm discussed previously, speech alignment is evidence that human users see these conversational agents as active interlocutors.

User vocal alignment with the agent in areas such as speech patterns, tone, rate, amplitude or inflection have been found, consistent with the communication accommodation theory (CAT), "which proposes that speakers use linguistic alignment to emphasize or minimize social differences between themselves and their interlocutors" (Zellou et al., 2021, p. 2). Zellou et al. (2021) also found that female human speakers are more likely to align their speech to the conversational agent (e.g., Siri or Alexa) than a male human speaker. Children have also been found to quickly adapt their speech patterns if a conversational agent does not understand them (Druga et al., 2017). For example, a child might phrase their question differently or try an alternate word if there is a comprehension issue on the side of the machine.

Evidence for vocal alignment reinforces the relevance of the CASA paradigm and the value of examining and understanding the machine voice from a sociophonetic perspective.

2.2.3 Machine Learning (NLP)(ASR) Bias Issues

Having discussed machine-human communication at a high level and how the technology has evolved over time, the following section aims to understand the sources of machine learning biases (as defined in section 1.5) and identify examples of machine biases. Given this chapter's discussion on ASR technology and the anthropomorphized relationships humans have with conversational agents, devices that are programmed with or learn biases (related to gender, race, sexuality, etc.) have the potential to perpetuate stereotypes to the users (e.g., a female-sounding assistant could theoretically reinforce a stereotype that assistants should be female).

2.2.3.1 Training issues. A 2019 analysis of the technology companies that develop virtual assistants show that only between 10% and 15% of researchers on development teams were women (West et al., 2019). The field of machine learning and artificial intelligence, overall, is male-dominated. This gender gap means the people programming and training virtual

assistants are mostly male, leading to a lack of diverse perspectives and the likely creation of bias-prone technology. Especially considering that popular virtual assistants are female-sounding, the lack of diversity can lead to oversight of how to program responses for taboo topics or perpetuation of harmful stereotypes for women. A cycle is then created where there are fewer women in STEM working on these technologies, and more biased technologies are created, but then more biased technology exists, leading to less women in STEM (Wang, 2020).

Early versions of Siri and Alexa were not exceptions. The virtual assistants would employ flirtatious responses or ones that perpetuated the stereotype of a flirty, sexual, and submissive female assistant. Siri, for example, would only refuse a sexual request if the user outwardly asked Siri to engage in sexual intercourse. Her response: "You have the wrong type of assistant" (Fessler, 2017). This response is non-confrontational and alludes to there being a female assistant that should willingly accept the sexual request. Virtual assistants have the potential to create stereotype normalization by not directly renouncing sexist requests (Wang, 2020).

Following an assessment of studies done on NLP bias, Blodgett et al. (2020) propose steps to move forward and create more equitable training circumstances. The first encourages researchers to understand relevant literature in order to examine the relationship between language and social hierarchies. This approach places emphasis on understanding how language is used to create labels and shape how people see different groups in society. The second encourages researchers to be more explicit about the type of bias being discussed, how it is harmful, and what groups are being impacted. Without clarity, the issues cannot be properly addressed or resolved. Finally, the third encourages researchers to examine language in use. This recommendation focuses on the value and impact of speech communities, how language can shift

meaning depending on the audience, and understanding how the language could be potentially harmful to users.

In summary, section 2.2 has provided insights into the origins of ASR technology, the anthropomorphism of machine interlocutors, and machine bias issues that may reinforce potentially harmful stereotypes. Understanding the rise of conversational machines, and therefore virtual assistants, provides important context for examining socio-indexical qualities conveyed by virtual assistant voices.

2.3 Social Expectations Related to Female Voices

Following section 2.2's discussion on anthropomorphism and the acknowledgement that virtual assistants have disembodied voices, this chapter aims to focus on social judgments or expectations of female voices. Analysis and discussion provided in this chapter acknowledges there is no one version of a female voice. That said, there are assumptions attached to various linguistic features and femininity. Female voices are of particular interest since Siri and Alexa have female-sounding default voices, which will be relevant for Chapter 3's analysis and discussion.

2.3.1 Female Voice as Noise

2.3.1.1 Historical Representations of Female Speech. The presence of female voices throughout history has been treated as "marked" or less desirable to a man's speech. The sounds of women in classical literature were described as primal and uncontrolled outbursts (Carson, 1995). Even as early as 6th Century BCE, laws were made in Sicily to regulate female noise during funeral lament, including restrictions on “location, time, duration, personnel, choreography, musical content, and vocal content of the female's funeral lament” (Carson, 1995,

p. 127). Societal perception deemed female noise barbaric in nature and something that could spark disorder more broadly.

Literature in the Middle Ages commonly juxtaposed a woman's speech with a man's sense of literacy (Neufeld, 2021). By creating this dynamic, a female speaking in a conversation indicated the female lacked proper social etiquette. The mere presence of a female voice in social environments was a negative contribution. In modern literature and film, the juxtaposition can be seen in "The Philosopher and the Shrew" archetype where an unruly wife is paired with an educated man. For example, the plot of the classic film *My Fair Lady* centers around Eliza Dolittle being coached by Henry Higgins on how to speak (Neufeld, 2021). Henry Higgins claims that by correcting Eliza's speech, people will consider her to be a proper lady of society (Cukor, 1964). There are men in the film who speak with Eliza's dialect, but Eliza was the main character. This was perhaps because the plot of an educated man teaching an unruly female fit societal expectations that women needed to be tamed vocally.

Speaking women have also long been characterized as gossips, with society characterizing collective female speech as associated with "sensuality, irrationality, and rebelliousness," as seen in 15th century antifeminist satirical literature into the early modern period (Neufeld, 2021, p. 35). Enforcing the idea of collective public female speech as meaningless creates a depiction of female voices as white noise and the thoughts and opinions of women are taken as uneducated chatter.

Even in terms of early 19th century education about voice, the female voice was reduced to being described by speech pathologists as "a reproduction of the male an octave higher." (Hoegaerts, 2020, p. 447). Beliefs such as this position the power or significance of a female voice as solely existing in comparison to a male voice. In fact, more precise studies into the

differences between female and male voices existed minimally. Male voices were analyzed in depth (e.g., voice change during puberty) and female voices were reduced to being discussed as not doing what the male voice did (Hoegaerts, 2020). Again, this dynamic reinforces perceptions of female voices as being inferior or less than male voices.

During the 1830s, women began to advocate for a place in the male-dominated field of U.S. politics. As more women became vocal during the suffrage movement, men discussed the female voice in terms of tonality (Levander, 1998). In fact, "the preeminent linguist Otto Jesperson summarized and gave 'scientific' credence to the long--held view that 'woman's language' was incapable of accommodating factual information and thus was essentially tonal…" (Levander, 1998, p. 15). Because emphasis was taken away from *what* women said and focused on *how* women said their words, a shift occurred. Women were expected to sound elegant and be "sweet little beings" who sound like a nightingale (Levander, 1998, p. 16). According to Levander (1998), women who stepped into the political arena were characterized by men as deteriorating their vocal tonality. One prominent male figure stated, "American women are physically attractive, when they open their mouths, they produce a mean, thin, nasal, rasping tone, by which you are at once disenchanted" (Levander, 1998, p. 17). Instead of earlier associations of women having shrill voices, women's voices were expected to be pleasing to hear, but not contain any actual information. Framing women's voices as only suitable for private life reinforced the role of women in society as submissive and domestic, while men were able to continue in the political arena.

During this same period, ensuring a child's vocal health became a maternal responsibility: "Sweet motherly voices were not only seen as important early guides toward vocal health, but also as a civilizing influence, pointing children toward moral behavior and good taste"

(Hoegaerts, 2020, p. 448). Beliefs toward the female voice seem to be shifting during this point, where some responsibility and validity is being transferred to women and caretakers. While more emphasis is being placed on the nurturing impact of a female voice, women are being framed as subservient caretakers rather than powerful scholars and educators.

Female voices continued to be viewed as inferior in public arenas with the advent of recording devices, radios, telegrams, etcetera in the 20th century. As Amy Lawrence (1991) explores in her book "Echo and Narcissus: Women's Voices in Classical Hollywood Cinema," many assumptions that women's voices could not be recorded hinged on the fact that the voice-based technology was created by men, using male voices as the prototypical voice. A general distaste for female voices when expressing any intellectual content also persisted in relation to the radio or other recordings: "One radio executive proclaimed 'it is my opinion that women depend upon everything else but the voice for their appeal. Their voices are flat or shrill, and they are usually pitched far too high to be modulated correctly'" (Lawrence, 1991, p. 31).

Female recorded speech in the 20th century was seen as a challenge to male dominance and overall societal decorum. Lawrence (1991) conducted a review of films in which she found most female characters had to overcome obstacles in order to be heard. The character would face repressive challenges if she refused to be silenced. Often, female voices in 20th century films were silenced, dubbed, or otherwise manipulated (Hankins, 1992). Specifically in Hitchcock films such as, *The Birds* and *Psycho*, the female voice is made to be submissive or inferior to the male voice and perspective (Hankins, 1992).

In the 21st century, female voices continue to be seen as superfluous noise that is only valued in certain social scenarios. For instance, there is a common stereotype that women speak

more than men, but this is far from generally true. These assumptions about female speech demonstrate that historical ideologies around when and how a woman can speak have persisted.

When considering this historical background on perceptions of female voices, it is interesting that popular virtual assistants all used female-sounding voices as their initial default setting (e.g., Siri, Alexa, Google, Cortana). Virtual assistants are meant to share information and facts, yet historically speaking, women's voices have been associated with feelings and emotions instead of facts. A potential reason for the choice to give virtual assistants female-sounding voices is based in section 2.1's discussion of how female-sounding robots are perceived in movies and film, and how those expectations or perceptions relate to real robots. Female-sounding robots have the ability to be seen as maternal, trustworthy, and non-threatening (as mentioned in 2.1). These qualities may outweigh stereotypical associations with female voices as superfluous noise in public spheres. Restriction of what environments women are believed to be more apt to thrive in creates perceptions that females are more likely to be warm, friendly, and capable of empathizing with emotions (Borau et al., 2021). Additionally, Borau et al. (2021) points out that because women are stereotypically more closely associated with emotional behavior and intuitions than men, female voices may help in developers' goals to make virtual assistants sound more human. From this perspective, the historical restriction of where and how a woman should speak has shaped perception of female voices in a way that could create a more human-like interaction between users and virtual assistants, especially considering these devices are typically placed in personal locations such as a user's home—a domestic domain. Additional motivations behind why female voices may have been used for virtual assistants are presented in 2.3.1.2.

Chapter 3 will discuss some of the relevant social characteristics in more depth when analyzing specific linguistic variables.

2.3.1.2 Female Voices in Technology. Building on the historical evolution of societal perceptions of female human voices, female-sounding machine voices can be more carefully understood and analyzed. As mentioned throughout this work, conversational agents take up space within human society and humans have commonly anthropomorphized such devices. Therefore, perceptions and judgments made about human female voices likely apply to female-sounding conversational agents.

As technology has advanced, research into synthesized voice development has progressed. Early research focused specifically on male synthesized voices simply due to technology limitations—machines that generated synthetic speech were limited in F0 production (Karlsson, 1991). Vocal qualities such as breathiness and formant space also inhibited research into synthesized female speech, since the machine language technology could not accurately replicate the sounds (Karlsson, 1991). Because of technological limitations, synthesized male speech became more advanced than synthesized female speech and set a foundation for computer voices to be male-sounding. Even as conversations around creating human-sounding machines progressed, female voices seemed increasingly difficult to replicate due to limited F0 synthesis capabilities, lack of previous research, and difficulty separating F1 values from vocal tract measurements (Karlsson, 1991). It is interesting then that female-sounding computer voices are dominant as the original default settings for popular virtual assistants. One would expect the opposite to hold true if machine voice generation only relied on available data and ease of replication.

The study done by Nass et al. 1997 (discussed in 1.2) further demonstrates how gender perceptions and stereotypes are applied to female-sounding computer voices. For instance, participants viewed evaluations given by male-sounding machine voices to be more valid than evaluations given by female-sounding machine voices. Results also show that participants expected the female-sounding machine voices to know more about stereotypically feminine subject matter (Nass et al., 1997).

In a study by Mullenix et al. (2003), female-voiced computers were rated as more truthful but significantly less powerful than male-voiced computers. When polling only male listeners, female synthesized speech was rated as more truthful, knowledgeable, and involved than male synthesized speech. The opposite results proved true for female listeners, with male synthesized speech being rated as more truthful, knowledgeable, involved, and accurate than female synthesized speech. Both male and female listeners ranked female synthesized speech as less powerful than male synthesized speech.

In previous marketing research, passiveness, vulnerability, and submission were all qualities associated with female voices (synthesized or human) while dominance, self-assurance, and independence have all been associated with male voices (synthesized or human; Chebat et al., 2007). A study done by Kwon 2010, found that Korean human female voices were ranked as significantly more intelligible than male voices. The authors attributed this to gender differences in "fundamental frequency, fundamental frequency range, formant frequency, formant range, vowel working space area and vowel dispersion" that influence speech intelligibility (Kwon, 2010, pp. 74-75).

These findings suggest that a human female voice may yield ratings that categorize the voice as less intimidating, more truthful (especially when the listener is male), and more

intelligible for listeners than a human male voice. This result is even more likely if the conversational subject matter is stereotypically feminine, as discussed in 2.3.1.1. Given the discussion in section 2.1, a machine voice that ranks as less intimidating may help assuage human fear of robot superiority and serve as a credible servant for humans. Generally, female voices today seem to still carry perceptions of inferiority to male voices but are perceived as reliable when filling a servant or domestic position.

This work largely discusses virtual assistants as female-voiced machines, but it should be noted that there has been diversification of voice options offered by companies for assistants like Siri. With Siri's most recent update, there is no default voice set for users. Instead, users select from five voice options: two female-sounding, two male-sounding, and one non-binary-sounding voice (Porter, 2022). That said, it is still interesting that all major popular virtual assistants debuted with female-sounding voices and seemed to be built with common personality traits in mind (as discussed in 2.1.3.2).

2.3.2 Social Discourse Expectations for Females

2.3.2.1 Women's Language. With an understanding of perceptions of female voices taken into consideration and an awareness of female-sounding virtual assistants, social discourse expectations for females should also be assessed. Lakoff's description of "Women's Language" has become a foundation for work on gendered speech. According to Lakoff (1973), women's language includes stylistic choices such as using diverse lexical choices and descriptive words (e.g., "mauve" instead of "brown"), using weaker expletives (e.g., "oh dear" instead of "oh shit"), and employing more adjectives that reflect adoration (e.g., "adorable," "charming," "sweet"). Lakoff (1973) also claimed that women's language includes increased use of hedging, euphemisms, and tag questions. Women who follow this way of speaking may be deemed as

proper women by society, though this may contribute to stereotypes about how women should act in society. Stereotypes in this case can be described as theories "about how members of another group look, think, and act, and how and why these attributes are linked together" (Strand, 1999, p. 94). Of note, Lakoff's observations were based on intuitions, unstructured observations, and were only of White middle-class women (Crawford, 2003). Women who are not members of this demographic group may be expected to speak differently, and there is of course variation in speech style among White middle-class women

Subsequent work has further interrogated and challenged Lakoff's claims such as work on tag questions by McMillan et al. (1977), Lapadat and Seesahai (1977), Holmes (1982), Cameron et al. (1989) and gossip research by Cameron (1997) (Eckert and McConnell-Ginet 2013). Even so, Lakoff's work remains foundational and still serves as a reference point for much research on language and gender. Themes linking women's speech to lack of power and concepts like politeness continue to be prominent in research on language and gender to today.

2.3.2.2 Factors influencing speech expectations. As stated in the previous section, Lakoff's work focused primarily on the speech of White middle-class women. Contemporary social stereotypes in the U.S. often dictate that a woman's speech should be gentle, trivial, and polite (Gal, 1989). The same stereotypes dictate that a man's speech should be aggressive, forceful, and blunt (Gal, 1989). Stereotypical expectations have the power to then reinforce the way people actually speak. In turn, the stereotype is reinforced. There is a performative element to "speaking like a lady." Lakoff (1973) acknowledges social reinforcement as she discusses the way younger girls are socialized into understanding what constitutes acceptable female speech. For example, if a girl uses crass language that is expected from a boy, there would likely be a negative social result such as being scolded. The girl then learns how to speak in a way that is

socially acceptable in order to avoid future negative social consequences. In a similar way, men who use speech that is associated with women's language may be viewed as more feminine or posh based on the scenario (Lakoff, 1973). This social perception is due to the male violating social expectations or stereotypes on how a man should speak.

Beyond stereotypical reinforcement, there are communities that use gendered speech as a reflection of labor divisions or roles within that group. For example, people in the Southern Highlands of Papua New Guinea set expectations that women are meant to lead public ceremonies, engage in ceremonial songs, and be at the forefront of public expressions of grief (e.g., weeping) in cases of significant loss (Gal, 1989). These duties require the women to be incredibly vocal in the public domain. The American expectation for females to be silent or submissive in public settings (as discussed in 2.3.1) therefore contradicts expectations in this Papua New Guinea community. In any community, properly following social expectations around speech could positively impact a woman's social status.

2.3.2.3 Gender Performance. Because gendered speech is influenced by social expectations, it is also possible to use language as a performative tool to enact a target gendered persona. In the 1990s, 900-number fantasy hotlines allowed callers to dial-in and intimately speak with a woman. Callers were mostly male. Kira Hall (1995) found that “their training manuals for the job tell [the individuals working the hotlines] to create stereotypical characters such as bimbo, nymphomaniac, mistress, slave, lesbian, and virgin.” They are also instructed to be “bubbly, sexy, interesting, and interested” (Hall, 1995, pp. 190–191). Women working at these fantasy hotlines might have naturally spoken in a wide range of styles or dialects. For the job though, language was manipulated in order to provide customers with a certain kind of

experience that matched or potentially surpassed social expectations for how a sexualized female should sound in this scenario.

As mentioned earlier, dominance is a trait stereotypically expected from male speech. A woman embodying dominance in their speech in a public domain would be violating a long-standing social expectation that only men's words have meaning in such domains (Neufeld, 2021). A woman may speak with dominance intentionally then as a means to assert power in a situation where she felt otherwise powerless. For example, as discussed in 2.3.1.1, women in the 1830s were attempting to enter the male-dominated political sphere and needed to adopt speech expected from a man in order to combat stereotypes that women were meant to sound pleasing to the ear but say nothing of substance.

In summary, section 2.3 discussed how female voices were perceived throughout history and explored current social expectations for females during discourse. These dynamics are crucial to consider as this work continues to discuss conversational agent voices, especially given that virtual assistants like Siri and Alexa are conceptually understood by users as female-sounding, following the CASA paradigm.

2.4 Relevant Sociophonetic Features

Based on section 2.3's discussion on female voices and social expectations related to how women speak, it follows that a speaker’s voice can play a major role in assumptions and judgments made about that speaker. While social expectations may influence a speaker's stylistic choices, the act requires a hearer to react (Levon, 2014). The listener is therefore crucial in a voice having social meaning. Phonetic qualities of a voice will impact the assumptions made by the listener. It should also be noted that character traits such as polite, kind, nurturing, etc. may typically be tied to stereotypical perceptions of female voices, but that does not mean all female

voices index these qualities. Perceived social meanings of a particular acoustic feature can shift over time or based on larger social shifts in how a community is perceived (Villarreal, 2018).

Section 2.4 presents an overview of previous research focused on phonetic features that align with various socio-indexical properties and have been linked to gender. For the purposes of this work, the focus will be on F0, vowel formants (F1 and F2), /s/ center of gravity, and -ING pronunciation. It should be acknowledged that the aforementioned phonetic features described in this section do not create an exhaustive list of contributing features. Other vocal features that may impact perception of a voice's gender and other social qualities include phonation, breathiness, creakiness, intensity, pitch range, and pitch contours, among others (Smorenburg & Chen, 2020).

It should also be noted in relation to conversational agents that "machine learning algorithms utilizing known aspects of voice that are associated with gender perception have also been developed. These algorithms classify speech signals into groups according to gender" (Leung et al., 2021, p. 2601). By gaining a more acute understanding of the socio-indexical characteristics tied to phonetic features, it becomes clearer how programmers could design or manipulate virtual assistant voices that index various social and personality traits.

2.4.1 F0

A phonetic feature often associated with conversations about female voices and femininity is vocal pitch, or F0. In studies of American English, men typically have an average F0 from 100-120 Hz whereas women have an average F0 of 200-220 Hz (Zimman, 2017). In a study of native English speakers from England, Cartei and Reby (2013) found that listeners consistently rated adult voices with lower F0 values as belonging to more masculine individuals.

When the F0 values were artificially raised, listeners identified the voices as belonging to less masculine individuals.

As discussed in Allen (2022), average F0 is also used in work related to perceived speaker attractiveness. A 2011 study by Borkowska and Pawlowski found that Polish-speaking participants associated attractive female-sounding voices with F0 values ranging from 220-262 Hz, with F0 values below 220 Hz being perceived as least attractive. Voices within an F0 range of 220 Hz to 262 Hz were typically perceived as more feminine, youthful, and flirtatious (Re et al., 2012). In Scottish English, Feinberg et al. (2008) found that male listeners perceived female voices that had their F0 values synthetically raised as more attractive than the non-raised forms. The impact of raising F0 values on listener perception was the smallest for the highest range tested (raising a voice from 241 to 261 Hz). This attenuation suggests there may be a point at which raising the F0 value does not yield increased perceived attractiveness (Feinberg et al., 2008). These studies indicate that for female voices higher F0 values were generally viewed as more attractive than lower F0 values.

As discussed by Skuk and Schweinberger (2014), a 2010 study done by Honorof and Whalen had male and female North American English speakers deliberately produce /a/ within an overlapping F0 range, making some female speakers produce the phoneme at lower F0 measurements than the male speakers and vice versa. Overall, participants were able to accurately identify the gender of the speaker at about 72% of the time. Confidence ratings in participant selection were lowest for stimuli with low-pitch females or high-pitch males. Listeners make social judgments based on vocal qualities and may base those judgments on pre-existing expectations for how a gender *should* sound (Levon, 2007). There has been relatively little research done on how sexuality correlates with F0 values among women, but Moonwomon-

Baird (1997) presented anecdotal evidence which suggested heterosexual women have a less restricted F0 range than lesbian or bisexual women. Crain (2019) analyzed a small sample (21 participants) of queer women voices. Data from Crain suggested that the queer women in the study generally had F0 values below 220 Hz.

Beyond identifying speaker gender or sexuality per se, F0 is associated with many other social characteristics. Male speakers with lower pitch values have been associated with increased physical dominance: taller, heavier, and stronger (Kempe, 2013). Higher F0 variation has been tied to emotions such as happiness, anger, and fear. Lower F0 variation has been tied to passive emotions like sadness (Breitenstein et al., 2001). F0 values also lead to perceptions of size, with higher F0 values being associated with smaller sizes and lower F0 values being associated with larger sizes (Eitan et al., 2014). F0 can also indicate age of the speaker, with children typically having F0 values above 200 Hz (Barkana & Zhou, 2015). F0 ranges for vowels can also create perceptions beyond human characteristics. As D'Onofrio and Eckert (2021) discuss, F0 can be assessed from a synesthetic perspective to address domains such as light, shape, weight, and affect.

In summary, F0 is commonly discussed in conversations about speaker gender, but F0 values also index social characteristics such as size, emotional status, age, attractiveness, and dominance. Beyond human personalities, F0 values can also be used to contrast qualities like heaviness and lightness, roundness and sharpness, and positivity and negativity.

2.4.2 Formants (F1 and F2)

Vowel formant frequencies also contribute to listener perceptions of voice characteristics, including speaker gender. According to Smorenburg and Chen (2020), combining the first three formants of a voice or audio sample can lead to human listeners identifying speaker gender with

92% accuracy. Each formant alone has lower accuracy rates, but still provides valuable socio-indexical information.

One view on why formant frequencies provide insights on speaker gender is due to anatomical differences in the shapes and sizes of male and female vocal tracts (Hillenbrand et al., 1995). A study by Cartei and Ruby in 2013 counters Hillenbrand et al.'s claim by analyzing the speech of prepubescent children. The children spoke with similar F0 values, but boys typically had lower F1-F3 than females, signaling a narrower formant space overall. Since these children were prepubescent, the findings indicate that these differences may be due to behavioral changes rather than solely reliant on sexual dimorphism.

Regardless of cause, females typically have higher vowel formant frequency values than males (Gelfer & Bennett, 2013). That said, a study done by Munson et al. (2006) of English speakers from University of Minnesota and from the St. Paul and Minneapolis area connected non-high front vowel production with higher formant frequencies for gay males than for heterosexual males (Queen, 2007). It should also be stated that there is a difference between listeners' ability to identify biological sex category based on voice and ranking the voice on a masculinity-femininity scale. Leung et al. (2021) analyzed 417 audio recordings of read speech from Australian English speakers. Leung et al. (2021) tested for effects of change in F0, F1, F2, and F3. Participants (seventeen students who spoke Australian English) were asked to first identify speaker gender and then rate each recording on a sliding masculinity-feminity scale, where a *very feminine* ranking was to the far left of the scale and a *very masculine* ranking was to the far right of the scale. Participant ability to first identify speaker gender heavily impacted perceptions of vocal masculinity or femininity. For example, if the participant identified the voice as belonging to a female speaker, the voice would be ranked as at least somewhat

feminine. Beyond a potential priming effect, as formants increased, the likelihood of a listener perceiving a speaker as more masculine decreased (Leung et al., 2021).

Within male speech, lower, narrower formant spaces are typically associated with male voices and index qualities like increased masculinity and threat potential (Kempe et al., 2013, citing Bruckert et al., 2006; Evans et al., 2006; Puts et al., 2012; Wolff & Puts, 2010). Pierrehumbert et al. (2004) also found that Chicago-area lesbian and bisexual women produced variants of /u/ and /ɔ/ vowels that were more backed compared to how heterosexual women produced the vowels. A study by Munson et al. (2006) found that the differences were more subtle when studying speakers from University of Minnesota and the St. Paul/Minneapolis metropolitan area, where lesbian and bisexual women did have backer vowels or lower F2 values than heterosexual women, but the only monophthong vowel where the difference was significant was /ɛ/.

Concerning child-directed speech, American English-speaking mothers widen their vowel space when speaking with children who are in the early language acquisition stage. Widening the vowel space makes it easier for children to distinguish vowel sounds, leading wider vowel spaces to be associated with higher degrees of articulateness (Kempe et al., 2013, citing Kuhl et al., 1997; Liu et al., 2003).

Formant frequency variation can also index conversational intention. A study done by Eckert (2009) focused on a preadolescent girl's speech. The girl first spoke about a topic she liked and then initiated a conversation about boys being jerks. LOT and PRICE vowels were moved significantly higher and backer when talking negatively. Similarly, a study by Geenberg (2014) asked adult speakers of English to talk to a plush toy pig—first to praise the pig as cute and then to comfort the toy for having a hurt knee. Most participants had a higher F2 for vowels

in the first task than the second, indicating that higher formant frequencies may be connected to playful praise and adoration while lowering of formant frequencies may indicate a calmer affect.

Formant frequencies and speaker vowel space also impact perceived persona of the speakers. For example, speakers in California have been found to now use a backed pre-oral TRAP vowel (more broadly speaking, front-lax vowels in California English are backing and lowering) which, in combination with other vocal features such as speech rate and creaky voice, leads to perceptions of the "valley girl" persona (D'Onofrio, 2015, citing Kennedy & Grama, 2012). More broadly speaking, the California vowel shift (CVS) consists of a LOT/THOUGHT merger, higher F2 values for high and mid back vowels like GOOSE, and lowering of KIT. There also appears to be a nasal split for TRAP in which the vowel is raised before a nasal and backs elsewhere (Villarreal, 2018). Previous research suggests that the CVS is associated with California(ness) at least for some speakers, as documented in Villareal's (2018) perception study and D'Onofrio and Pratt's (2017) analysis of parodied Californian speech in Saturday Night Live's *The Californians*. Villarreal (2018) also notes that the CVS has been linked with qualities like carefreeness, fun, Whiteness, femininity, and privilege (Villarreal, 2018, citing Podesva, 2011; Fought, 1999; Eckert, 2008). Fought (1999) also found that GOOSE fronting in California English was associated with lower class speakers and more conservative pronunciations of the GOOSE vowel was associated with middle-class. Results are all dependent on the social context in which the speaker is using the CVS and what preconceived notions may already be held by listeners. According to Villarreal (2018), an average GOOSE vowel for White or Latina California-English-speaking females falls into an F2 range of 1,640-2,219 Hz and TRAP vowel for those speakers had an F2 range of 1,626–1,895 Hz. A TRAP F2 value of 1,895 Hz cannot

definitely be labeled as *extremely backed* though because backness is relative to listener perception compared to their own speech (Villarreal, 2018).

In summary, formant frequencies index social qualities such as masculinity or femininity, sexuality, articulateness, and can lead to persona judgments as seen with the California vowel shift. Because formant frequency values cannot fully be attributed to vocal tract size, formant frequency variation can signal social performance from a speaker, indicating links to certain traits, affects, or social groupings.

2.4.3 /s/

Previous research has also demonstrated how the phonetic realization of /s/ is connected with perceptions of gender. While earlier discussion of /s/ production attributed differences to anatomical variations (women have smaller vocal tracts), the physiological difference is actually minimal (Strand, 1999; Heffernan, 2004). Strand (1999) also found that prepubescent children were already aligning their fricative production with adult methods (females were fronting /s/ and males were more alveolar) even though there was even less of an anatomical reason at this young age for there to be a significant and consistent distinction. Additionally, /s/ production is heavily influenced by tongue placement, which occurs in the oral cavity and may lead to the vocal tract having a minimal impact on how the sound was produced by a speaker (Zimman, 2017). Differences are also exaggerated by speakers manipulating the roundness of the lips (D'Onofrio & Eckert, 2021). The remainder of this discussion will focus on /s/ frequencies, but other aspects have been examined; see for example Crist (1997) on duration of /s/ clusters.

Research on /s/ production for American-English speakers found a peak frequency between 6,500 and 8,100 Hz for women and between 4,000 and 7,000 Hz for men (Zimman, 2017, citing Flipsen et al., 1999; see also Schwartz, 1968). Production of /s/ is in turn linked to a femininity and masculinity scale, with lower peak frequencies (backed /s/) identified as more

masculine sounding and higher peak frequencies (fronted /s/) identified as more feminine sounding (Bekker & Levon, 2017).

Zimman (2017) found that transmasculine people who were receiving testosterone therapy saw small but significant shifts in /s/ center of gravity over time. Because some participants had an increase in /s/ center of gravity, some participants had a decrease, and some saw no change; the variability in center of gravity is likely due to intentional articulatory shifts instead of anatomical changes to participants. One participant began producing more fronted /s/ phonemes over time, which is typically aligned with more feminine attributes. Zimman suggests that the participants felt they could begin introducing more feminine characteristics into how they expressed themselves, once they were more broadly perceived as men.

Campbell-Kibler (2011) used sample recordings from four male speakers (two from North Carolina and two from California) in order to test responses to three features ((-ING), pitch, /s/ and /z/). Campbell-Kibler found the male voice with fronted /s/ or /z/ tokens were viewed as less masculine-sounding and were more likely than voice samples with mid or backed /s/ or /z/ to be classified by participants as "gay-sounding." Results also found that there was a strong correlation between fronted /s/ and decreased competency. The correlation weakens for certain character types at the extreme ends of this scale (the highly educated posh male vs the incompetent hyper-masculine male). Judgments such as this are examples of how social stereotypes influence listener perceptions. Particular linguistic features may indicate gayness, but these are contingent on social cues and stereotypes that the listener perceives. For the purposes of this work, it is also important to note that the correlation described above is for male voices.

Podesva and Van Hofwegen (2016) analyzed the speech of speakers in Redding, California. /s/ COG for straight versus gay women was tested. Findings showed that speakers

identifying as lesbians had significantly lower COG values than speakers identifying as straight women in this community. Lesbian speakers had a COG range of 4,955-6,120 Hz. Compare this to straight rural women who had an average COG range of 5,863-7,743 Hz and straight town women at 6,120-6,930 Hz. Limited research has been done in other communities comparing /s/ COG values based on speaker sexuality, but Podesva and Van Hofwegen's results offer a baseline comparison for the purposes of this work.

Podesva and Van Hofwegen (2016) also compared how speaker's being from rural versus town environments impacted /s/ COG. Results found that females consistently had higher COG values than males and younger speakers consistently had higher COG values than older speakers. The age variable had a more significant impact on speakers from rural communities than on speakers from town. Younger speakers in rural communities were more likely to adopt a fronter /s/ that was similar to /s/ production in town. Younger rural females had an average COG of 6,930 Hz versus older rural females at 5,544 Hz. Younger town female speakers had an average COG of 6,555 Hz versus older town female speakers at 6,120 Hz. Of note, these results are of a specific community in California but provide preliminary evidence that /s/ COG can be impacted by if the speaker lives in a rural or more urban environment.

Individual social expectations of communities should also be considered when discussing /s/ production. Results may vary depending on the community (e.g., rural vs. urban), age, or other socio-indexical qualities of a speaker's voice. Furthermore, /s/ can index class status in Arabic (Haeri, 1996) or even threat potential in Southeast England (Holmes-Elliott & Levon, 2017). A study done by Bekker (2007) found that a fronted /s/ was "more characteristic of wealthy northern suburbs of Johannesburg than elsewhere, with the northern suburbs being clearly associated in the local South African English consciousness with wealth and prestige"

(Bekker & Levon, 2017, p. 1110, citing Bekker, 2007). Stuart-Smith et al.'s (2003) analysis of Glaswegian speakers' production of /s/ also found correlations between /s/ quality and sex, age, and class. For instance, results showed that working-class women had a similar minimum /s/ frequency to male speakers (approx. 2,000-2,500 Hz), but middle-class women had higher minimum /s/ frequency (approx. 3,500-4,000 Hz).

In summary, /s/ production in English is tied to gender, age, location, and sexuality, with fronted /s/ typically being associated with femininity, straightness, youth, and a more urban setting. /s/ can also index other social information such as social class or threat potential.

2.4.4 -ING

Pronunciation of word-final -ING is another phonetic feature that provides listeners with socio-indexical information. While the variable does not directly index masculinity or femininity, -ING is used in conjunction with other linguistic factors to create various speaking styles (Gratton, 2016). Variants with an alveolar nasal will be referred to as *-in* and those with a velar nasal will be referred to as *-ing.*

In terms of -ING's connection to gender performance, Gratton (2016) conducted a study of two non-binary trans individuals in different social scenarios to test how -ING pronunciation fluctuated in relation to gender and setting. Participants were educated, White native speakers of English from middle-class families. Flynn was assigned female at birth and Casey was assigned male at birth. Gratton found that usage of -ING variants in queer-friendly environments were similar between the two speakers. In nonqueer public environments though, Flynn was more likely to use *-in* and Casey was more likely to use *-ing*. This pattern suggests that Flynn and Casey were more conscious of their speech in nonqueer public spaces and wanted to distance their speech from their respective assigned-at-birth genders.

According to Tagliamonte (2004), English-speaking females are more likely than male counterparts to use *-ing* and view it as prestigious (Tagliamonte, 2004, citing Fischer, 1958; Labov, 1966/1982; Trudgill, 1972). To further explore socio-indexical properties conveyed by -ING pronunciations, Tagliamonte analyzed 70 speaker samples of the York English corpus, where every speaker was natively from York and was a speaker of British English. Tagliamonte found that blue collar workers and students preferred *-in* usage while white collar workers disfavored the variant (Tagliamonte, 2004).

Labov et al. (2011) found that in addition to social class, -ING pronunciation can also be tied to conversational formality, or how much attention is being paid to speech. Results showed that in casual scenarios, participants in lower social classes were more likely to use *-in* while participants in higher classes were more likely to use *-ing*. Lower class speakers demonstrated about an 80% *-in* rate in casual speech, which was more than working class (40-50%), lower-middle class (30-40%), and upper-middle class speakers (0-20%) in the same conversational context. The same trend occurs when participants were asked to read a passage; however, the difference in *-in* production is minimized between classes: lower class speakers demonstrated slightly above 20% *-in* rate in casual non-read speech, which was more than working class (10-15%), lower-middle class (about 0%), and upper-middle class speakers (about 0%). In casual speech, the difference in *-in* usage between lower class and upper-middle-class speakers was about 60% (Labov et al., 2011). The difference in *-in* usage for read speech between lower class and upper-middle-class speakers was about 20%, though this number may be skewed by a floor effect (Labov et al., 2011). More generally, -ING as a sociolinguistic variable is more commonly associated with informality and lower social classes when pronounced as *-in*.

As Campbell-Kibler (2008) discusses, interpretation of -ING (and other phonetic features) is largely impacted by the listener's preconceived notions about dialectal variation. For example, in her study, participants were asked to make judgments about segments of speech from eight speakers who were individually recorded during informal hour-long interviews. Recordings were then altered using an extended matched guise technique in which the -ING variable was altered for the two variants. Participants judging the speaker "Elizabeth" disagreed on if the *-ing* usage created a compassionate or condescending effect. For example, participants who were from the south were more familiar with *-in* and therefore reacted less negatively to its usage than participants from California who were more familiar with hearing *-ing*.

In summary, *-ing* has a range of socio-indexical associations, including associations with female speakers, higher classes, and articulate speech in casual speaking contexts. Change in usage of the -ING variable can also indicate a speaker attempting to align speech patterns with a particular gender. That said, judgments made about speakers using *-ing* or *-in* can be highly variable depending on listener perceptions of speaker dialect or their own listener dialect.

Section 2.4 has discussed social qualities associated with F0, formants F1/F2, /s/, and -ING. Information in section 2.4 will be used throughout Chapter 3 as part of better understanding the social qualities that relate to and may be indexed by the voices of Siri and Alexa.

Chapter 3: An Analysis of Siri and Alexa's Vocal Traits

Building on literature presented in Chapter 2, Chapter 3 reports on Siri and Alexa's F0, formants F1 and F2, /s/ center of gravity, and -ING production in order to analyze socio-indexical qualities that may be conveyed to the human user through voice. The virtual assistants chosen for this study were Apple's Siri (iOS 14, American, Voice 4) and Amazon's Alexa (default setting). Voice 4 was chosen as Siri's voice for this study because it is most similar to Siri's original default voice.

Studies 1 and 2 focused on F0 and formants F1 and F2. Vowels included were /i/, /ɪ/, /e/, /ɛ/, /æ/, /u/, /ʊ/, /o/, /ɔ/, /ɑ/. Study 3 focused on Siri and Alexa's center of gravity for /s/. Study 4 focused on analyzing -ING production for each virtual assistant. Specific methodologies and results will be discussed in their respective sections throughout the chapter.

3.1 Overview of Stimuli

This section provides an overview of all stimuli used in Chapter 3 experiments, with the exception of stimuli used in 1.2 which will be described within that section. Vowels of interest mentioned above were analyzed using 10 distinct stimulus sentences (see Appendix A), each consisting of three to four words that include a specific vowel of interest (referred to in this paper as VI). For example, if the VI is /i/, the sample sentence for /i/ includes three to four words that include /i/. The environments for each VI were restricted to cases where the VI is followed and preceded by an obstruent or no sound in the carrier word (word initial or word final). Function words were not included in the analysis. Methodology for recording F0 and formants F1 and F2 are defined in sections 3.3 and 3.4.

Each virtual assistant was also prompted to recite the "Rainbow Passage" (see Appendix B), which was used, for instance, by Zimman (2017) in his study on gender, F0, and /s/. There

were 16 tokens of word-initial /s/. Each token was measured for center of gravity (COG). Methodology for recording COG is defined in section 3.5.

Siri and Alexa's recordings of "Rainbow Passage" were also analyzed for pronunciation of word-final -ING tokens. Siri and Alexa were then prompted to recite “The Boy Who Cried Wolf” (see Appendix C) to increase the number of word-final -ING tokens produced. There were six tokens of word-final -ING in each passage, resulting in 12 total word-final -ING tokens for Study 4. Methodology for classifying -ING pronunciation is defined in section 3.6.

3.2 General Methodology

The methodologies included in this section pertain to all studies detailed in Chapter 3.

3.2.1 Recording Equipment

The Zoom H2N and the Audio-Technica Pro 70 lavalier microphone were used to record each utterance from Siri and Alexa, all of which were emitted from the built-in speaker of an iPhone 11 (version iOS 14.8.1). Utterances were recorded with XY stereo and low-cut turned on. Individual .wav files were then uploaded into PRAAT for analysis. During each recording, the lavalier microphone was placed closely to the iPhone external speaker to ensure accurate audio was captured.

3.2.2 Virtual Assistant Versions

Siri and Alexa read the 10 sentences, the “Rainbow Passage,” and “The Boy Who Cried Wolf” as described in the stimuli section above using text-to-talk software on an iPhone 11 version iOS 14.8.1. The following two sections detail how each virtual assistant was prompted to speak.

3.2.3 Prompting Siri

Each sentence was typed into a Google Doc. The "Speak Screen" capability in iPhone's accessibility setting for spoken content was turned on. Each document consisted of one sample sentence or passage. Each document was opened on the iPhone via the Google Docs app. Then, the "Hey Siri, speak screen" command was given for each document. Siri read the text and the response was recorded using the lavalier microphone and Zoom H2N microphone as outlined in section 3.2.1.

3.2.4 Prompting Alexa

Alexa was used on the iPhone via the Amazon Alexa app version 2.2.454039. Within the app, each sample sentence or passage was written into a different note. The command "Alexa, read note" was given for each note. Alexa then responded with "I found one note. First note…" followed by the sample sentence text. Given a character limit in Amazon Alexa notes, the "Rainbow Passage" was written over the course of four notes and "The Boy Who Cried Wolf" was written over the course of three notes. Each response from Alexa was recorded using the lavalier microphone and Zoom H2N microphone as outlined in section 3.2.1.

3.3 Study 1: F0

F0 is often tied to conversations around speaker gender, emotion, size, femininity, attractiveness, and other social characteristics (as discussed in section 2.4.1). This study aims to understand how F0 values of Siri and Alexa relate to sociophonetic findings from previous linguistic research, as discussed in 2.4.1. Study 1 focused on the average F0 for each virtual assistant per sentence and per vowel of interest. Specific methodologies are outlined below.

3.3.1 Methodology

Following general methodology described in 3.2, Siri and Alexa were each prompted to recite 10 stimuli sentences (see Appendix A). Once each recitation was recorded, the .wav files were transferred from the recording device memory card onto a Praat-compatible computer. All 20 recordings were then opened in Praat and individually cut to ensure minimal recording existed before and after the virtual assistant's recitation. Average F0 for each sentence was then calculated using Praat. F0 values of each vowel of interest (VI) as outlined in Chapter 3's introduction were also taken for each isolated vowel. Findings from this portion of Study 1 will be referred to as Study 1.1 results.

Additionally, results from this study were compared with data collected in Allen 2022, which was an analysis of Siri and Alexa F0 values as related to vocal attractiveness and F0 variation across prompt type. In Allen 2022, Siri and Alexa were given five prompts from each of three different prompt types— emotional, factual, opinion-based— to test if type of prompt impacted utterance F0. Emotion-based prompts such as “I'm lonely” required Siri and Alexa to address human emotion, fact-based prompts such as “Who is the current president?” required informational responses, and opinion-based prompts such as “What should I have for dinner?” required the virtual assistant to offer an opinion (Allen, 2022). The goal of this research was to identify whether and how virtual assistants shift speech (particularly F0) based on conversational intent.

In order to create a comparable data set, the 15 prompts asked of Siri and Alexa in Allen 2022 (see Appendix D) were replicated using the methodologies outlined in section 3.2. Findings from the re-recorded answers will be referred to as Study 1.2 results. Findings from Study 1.1 and Study 1.2 create a more robust data set and holistic view of Siri and Alexa’s F0 patterns.

3.3.2 Results

Study 1.1 resulted in data regarding F0 values for each VI token, average F0 per VI, and overall average F0 values per sentence. Study 1.2 yielded F0 values for sentences across prompt types as described in 3.3.1. Results from 1.2 were compared with 1.1 overall average F0 values per sentence.

3.3.2.1 F0 values per VI token. For both Siri and Alexa, initial VI token F0 values were higher than final VI token F0 values within each stimulus sentence (see Appendix A). Table 1 shows Siri's F0 values for first VI token, final VI token, and difference (*d* = first VI-final VI). Table 2 shows Alexa's F0 values for initial VI token, final VI token, and difference (*d* = first VI - final VI).

Table 1

Siri: Difference Between F0 of First and Final VI Token (Hz)

VI	First VI token	Final VI token	Difference
/i/	278	233	45
/ɪ/	344	150	195
/e/	272	161	111
/ɛ/	293	190	103
/æ/	239	190	49
/u/	296	204	92
/ʊ/	360	205	155
/o/	297	196	101
/ɔ/	303	195	107
/ɑ/	297	168	129

Table 2

Alexa: Difference Between F0 of First and Final VI Token (Hz)

VI	First VI Token	Final VI Token	Difference
/i/	258	204	54
/ɪ/	298	184	113
/e/	269	159	110
/ɛ/	280	186	94
/æ/	228	156	72
/u/	274	190	84
/ʊ/	290	156	134
/o/	286	192	94
/ɔ/	252	201	51
/ɑ/	252	173	78

Three key patterns emerge here. First, for nearly every single first and final VI, Siri's F0 was higher than Alexa's. This fits with the finding discussed below in sections 3.3.2.2 and 3.3.2.3 that Siri's F0 per VI is significantly higher than Alexa's. Second, in every case for both Siri and Alexa, first VI token F0 value was higher than final VI token F0 value. An analysis of first VI token F0 value versus other VI tokens F0 values within each stimulus sentence did not yield a consistent result. In some cases, the initial VI token did not have the highest F0 value in the sentence. Acknowledging that the virtual assistants were prompted to read declaratives in Study 1.1, the F0 trend is consistent with human-spoken statement intonation patterns, with F0 falling at the end of a statement (Cruttenden, 2008).

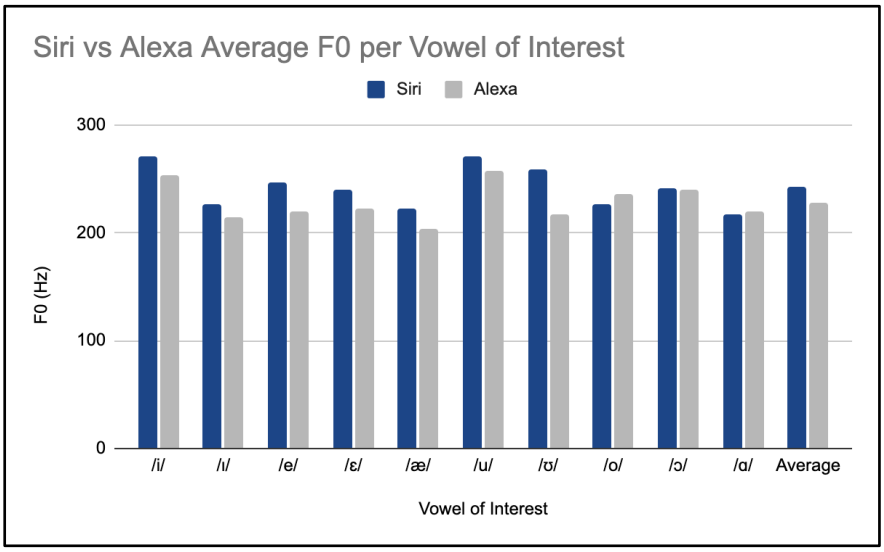
Third, the difference in F0 values between the first and final token for the VIs was generally greater for Siri than for Alexa. On average, Siri spoke with a 109 Hz difference between its initial and final VI tokens, while Alexa's average difference between its first and final VI tokens was 89 Hz. Siri's larger difference in F0 values may signal increased F0 range

within each sentence when compared to Alexa, though the differences between the two assistants (first VI F0 – final VI F0) was not quite statistically significant (paired $t = 1.967$, $p = 0.081$).

3.3.2.2 Average F0 per VI. Average F0 per VI refers to the average F0 of the VI tokens per stimulus sentence. For example, the average F0 value of /i/ was found by calculating the average of F0 values for the four VI tokens within stimulus sentence 1 (see Appendix A). The far right column in Figure 1 shows each speaker's overall average vowel F0 value.

Figure 1

Comparison of F0 Values per VI Category and Overall F0 Average for Each Virtual Assistant



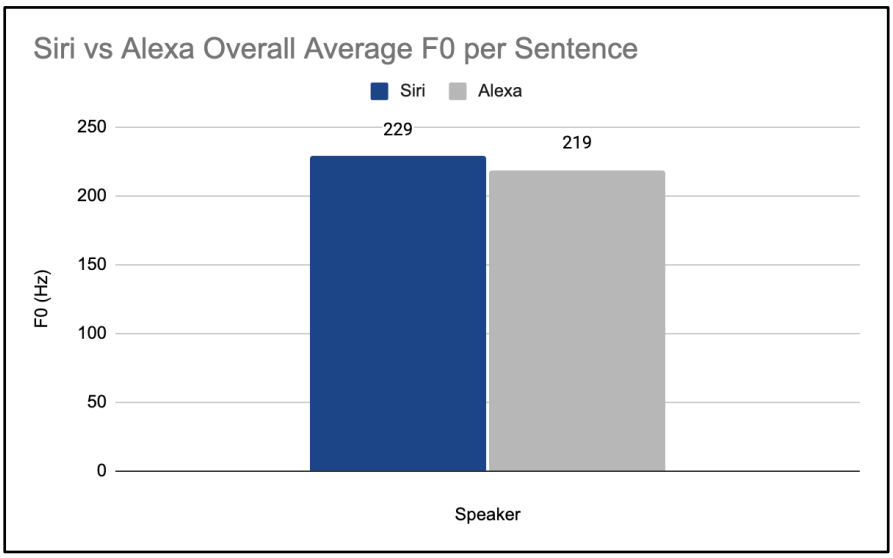
Siri has a higher F0 value than Alexa for every VI except the back mid and low vowels (/o/, /ɔ/, /ɑ/). The differences between Alexa and Siri's observed F0 were minimal for /ɔ/ and /ɑ/, with Alexa's value being only 2 Hz lower than Siri's for /ɔ/ and only 3 Hz higher than Siri's for /ɑ/. Alexa's average F0 value for /o/ was 10 Hz higher than Siri's.

A paired t-test indicates that Siri's average F0 for vowels overall is significantly higher than Alexa's average F0 value for vowels overall, at 242 Hz and 228 Hz, respectively (paired $t = 2.36$, $p = 0.024$).

3.3.2.3 Overall Average Sentence F0 Values per Speaker. For each stimulus sentence in Study 1.1, the average F0 per sentence was recorded for each virtual assistant. Results are shown in Figure 2.

Figure 2

Comparison of Overall F0 Values per Stimulus Sentence for Each Virtual Assistant

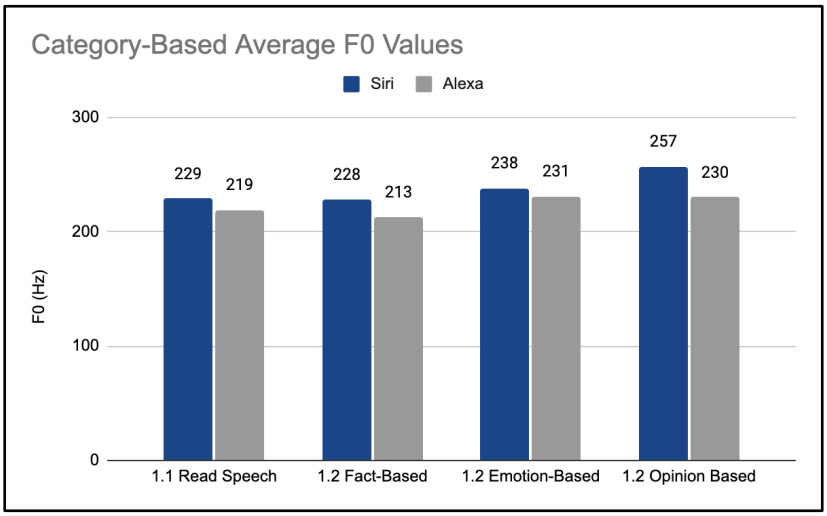


As seen in Figure 2, Siri has an overall higher F0 per sentence than Alexa, with a difference of 10 Hz between speakers. Siri's higher overall F0 value per sentence aligns with results from the previous section, which showed that Siri also has a higher overall F0 for VI. These average F0 values for the 10 sentences are also included in the analysis presented in the next subsection, which examines Siri and Alexa's F0 by prompt type.

3.3.2.4 Comparing Study 1.1 overall F0 with Study 1.2 Results. Results from 1.2 describe how each virtual assistant's average F0 per sentence shifts depending on prompt type. As detailed in Allen 2022 and outlined in Appendix D, 15 prompts were given to the virtual assistant that were equally divided between fact-based, emotion-based, or opinion-based prompt types (discussed in 3.3.1).

Figure 3

Comparison of Results from Study 1.1's Average F0 Values per Sentence and Study 1.2's Average F0 Values per Prompt-Type for Each Virtual Assistant (Siri and Alexa)



As seen in Figure 3, average F0 values per sentence from Study 1.1 align most closely with the average F0 values calculated for responses to fact-based questions, with Siri's average F0 for 1.1 being 229 Hz and 228 Hz for 1.2 fact-based responses and Alexa's average F0 for 1.1 being 219 Hz and 213 Hz for 1.2 fact-based responses.

Beyond these averages, results were further analyzed to see if there was a significant effect of speaker (Siri vs. Alexa) and prompt type (fact, emotion, opinion, read speech) on the average F0 of the utterances. A mixed-effects model was run to test this, using the lme4 package in R (Bates et al. 2015). As shown in Table 3, fixed effects were prompt type and speaker. Fact-based prompts were treated as the baseline in testing for prompt-type effects. Individual prompts were included as a random intercept.

Table 3

Mixed-Effects Model: Average F0 of Utterances

Fixed effects				
	Estimate (Hz)	Std. Error	*t*-value	Pr (> |t|)
Intercept	213.7	5.886	36.30	< 2e-16
Speaker: Siri	13.9	3.798	3.66	0.001
Prompt: emotion	13.7	7.879	1.74	0.097
Prompt: opinion	22.8	7.879	2.90	0.009
Prompt: read	3.2	6.824	0.47	0.644
Random effects				
Groups	Name	Variance	Std.Dev.	
Prompts	(Intercept)	65.05	8.065	

Note. Number of obs: 50, groups: prompt, 25

"Speaker: Siri" is the effect of the speaker being Siri rather than Alexa, "Prompt: emotion" is the effect of the prompt type being "emotion" rather than "fact," etc. Estimates are in Hz. So, for instance, the model estimates a significant increase by 13.8 Hz in an utterance's average F0 for the speaker being Siri rather than Alexa ($p = 0.001$). The model also estimated a significant increase by 22.8 Hz when the prompt type was opinion rather than fact ($p = 0.009$). There was a marginally significant increase by 13.7 Hz when the prompt type was emotion rather than fact ($p = 0.097$). There was no significant difference between responses to fact-based prompts versus read prompts.

A model with an interaction between speaker and prompt type was also tested. The model with that interaction did not perform significantly better than the simpler model above. This is

consistent with the fact that for both Siri and Alexa emotion- and opinion-based responses tended to have higher average F0 than fact-based responses and read speech.

3.3.3 Discussion

Results from Study 1 showed that both Siri and Alexa used a higher F0 value for the initial VI token than for the final VI token in read speech, consistent with human intonation patterns when speaking a statement (Cruttenden, 2008). The virtual assistants were reading the sentences from a document (as outlined in 3.2.3 and 3.2.4) but still were programmed to mimic human speech patterns. Presumably, this alignment serves the broader goal of making voice-activated conversational agents sound more like humans. It would be interesting to conduct a study where the virtual assistants were asked to read questions instead of statement sentences. By maintaining a text-to-speech variable but alternating the desired sentential force (declarative versus interrogative), it would become clearer how much the virtual assistant is mirroring human F0 patterns.

When looking more closely at F0 values for VIs, differences between the virtual assistants arose. Siri had a higher average F0 per VI, unless the VI is a mid or low back vowel. In the cases of a VI being a mid/low back vowel, Alexa had a comparable or slightly higher average F0 value. On average though, Siri had a significantly higher overall average F0 per VI (p = 0.024).

Average F0 results at the sentential level mirror the VI averages. Siri's average F0 per sentence was 229 Hz, compared to Alexa's 219 Hz. Based on literature outlined in section 2.4.1, a higher F0 value correlates with perceptions of the speaker being more feminine-sounding. Based on pitch alone, one would expect Siri to be perceived as more feminine than Alexa, though of course other factors contribute to femininity perceptions. Because Siri and Alexa's F0

averages are in or near the average F0 range for an adult female voice (200-220 Hz) cited in previous work (e.g., Zimman, 2017), Siri and Alexa are very likely to be identified by listeners as voices belonging to female speakers. As explored in 2.4.1, voices with F0 values between 120-200 Hz are less likely to be confidently and correctly matched with the speaker gender by listeners. If listeners identified Siri and Alexa as female speakers, they would likely perceive both voices as more or less feminine sounding, but not as more or less masculine sounding on a masculinity-femininity scale. That is to say that a female-sounding voice would not likely be ranked as any level of masculine (Leung et al., 2021; discussed in detail in section 2.4.2).

Based on discussions in 2.4.1, Siri might also be anthropomorphized as a more petite and youthful female than Alexa, though, again, other features beyond F0 would figure into such perceptions. A more concentrated analysis of F0 variation would need to be conducted in order to identify how Siri and Alexa's voices may index emotions or speaker sexuality (discussed in 2.4.1). Similarly, given the findings of previous research on F0 and attractiveness in women, Siri's higher F0, all else being equal, might lead to greater attractiveness ratings by listeners for Siri over Alexa. However, it is unclear at this stage how F0 interacts with other aspects of voice in perceptions of attractiveness. Therefore, given that Siri and Alexa's voices differ in other ways, it would be premature to conclude based on F0 alone that Siri would generally be perceived as more attractive.

In comparison with data found in Study 1.2, Figure 3 showed that Siri and Alexa's average F0 values in Study 1.1 were closest to fact-based responses from 1.2. Siri's average F0 value in 1.1 was 229 Hz and 228 Hz in Study 1.2. Alexa's average F0 value in Study 1.1 was 219 Hz and 213 Hz in Study 1.2. The mixed-effects model found no statistically significant difference between read responses in Study 1.1 and fact-based responses in Study 1.2 (p =

0.644). This was likely because fact-based responses generally involved the virtual assistant reciting text found on the internet, similar to the task of reading a document as laid out in 3.2.

Interestingly, there was a statistically significant difference between opinion-based and fact-based responses ($p = 0.009$) and a marginally significant difference between emotion-based and fact-based responses ($p = 0.097$). Combined with results found in 3.3.2.1, it seems Siri and Alexa maintain a similar statement-forming speech pattern for text-to-speech responses, regardless of whether the user specifically asks for information to be read or if the virtual assistant is proactively finding facts to read from an online source. As it relates to a broader discussion on how findings in human-based linguistics can be applied to machine language systems, F0 variation of virtual assistants based on context of the conversation is potentially reflective of human sensitivity to conversational context and is one way of making virtual assistants "sound more human" (as mentioned in 2.4.1).

Study 1 provided F0 data on Siri and Alexa that provide insights into F0 range and fluctuation. Data demonstrated that Siri and Alexa employ a falling F0 pattern when reading statements and have an F0 range that is expected of human adult females, with Siri having a higher F0 on average. These F0 values may be expected to push Siri toward being perceived as more feminine and youthful than Alexa, when all other components except F0 are held constant. Both virtual assistants raise their pitches when answering open-ended emotion-based or opinion-based prompts, suggesting the virtual assistants were programmed for an awareness of listener needs within a conversation. As indicated in 3.3.2.1, there is some preliminary evidence that Siri has a wider F0 range than Alexa, though further research would be needed to fully understand how virtual assistant F0 ranges compare.

3.4 Study 2: Formants (F1 and F2)

As discussed in section 2.4.2, F1 and F2 have socio-indexical properties tied to articulateness, gender, conversational intent, and can lead to persona judgments based on listener perceptions of a dialect. Study 2 aims to map out F1 and F2 planes for Siri and Alexa in order to depict each virtual assistant's vowel space. Specific methodologies are outlined below.

3.4.1 Methodology

Following the general methodology outlined in 3.2, the .wav files mentioned in section 3.3 were used again in Praat for this study for consistency of data. Each VI token was isolated within each sentence. After identifying the start and end time of the VI token by hand, a mid-point was calculated. F1 and F2 measurements were calculated by Praat at 50% duration of each isolated vowel.

3.4.2 Results

Study 2 resulted in F1 and F2 measurements for Siri and Alexa. F1 and F2 values were then used to map out Siri and Alexa's vowel spaces.

3.4.2.1 F1. As shown in Table 4, Siri had higher average F1 values than Alexa for every VI.

Table 4

Differences in Average F1 Values per VI Between Siri and Alexa

VI	F1 Alexa	F1 Siri	F1 mean diff	*t*-value	Pr (> |t|)
/i/	360	491	-132	-7.04	0.006
/ɪ/	473	630	-157	-2.43	0.093
/e/	515	537	-22	-0.65	0.564
/ɛ/	709	895	-186	-5.07	0.015
/æ/	970	1038	- 68	-2.86	0.065
/u/	353	453	-100	-3.63	0.036
/ʊ/	467	833	-366	-4.27	0.051
/o/	500	754	-254	-2.79	0.069
/ɔ/	1020	1027	-7	-0.13	0.906
/ɑ/	963	1012	-49	-1.21	0.313

Based on paired t-tests, Siri's F1 was significantly higher for /i/ (p = 0.006), /ɛ/ (p = 0.015), and /u/ (p = 0.036). The difference between Alexa and Siri's F1 was marginally significant for /ɪ/ (p = 0.093), /æ/ (p = 0.065), /o/ (p = 0.069), and /ʊ/ (p = 0.051). These t-tests were based on a maximum of four observations per vowel per speaker, so it is remarkable that the F1 differences were significant for multiple VIs.

3.4.2.2 F2. Table 5 compares the average F2 values for each vowel for Alexa and Siri.

Table 5

Differences in Average F2 Values per VI Between Siri and Alexa

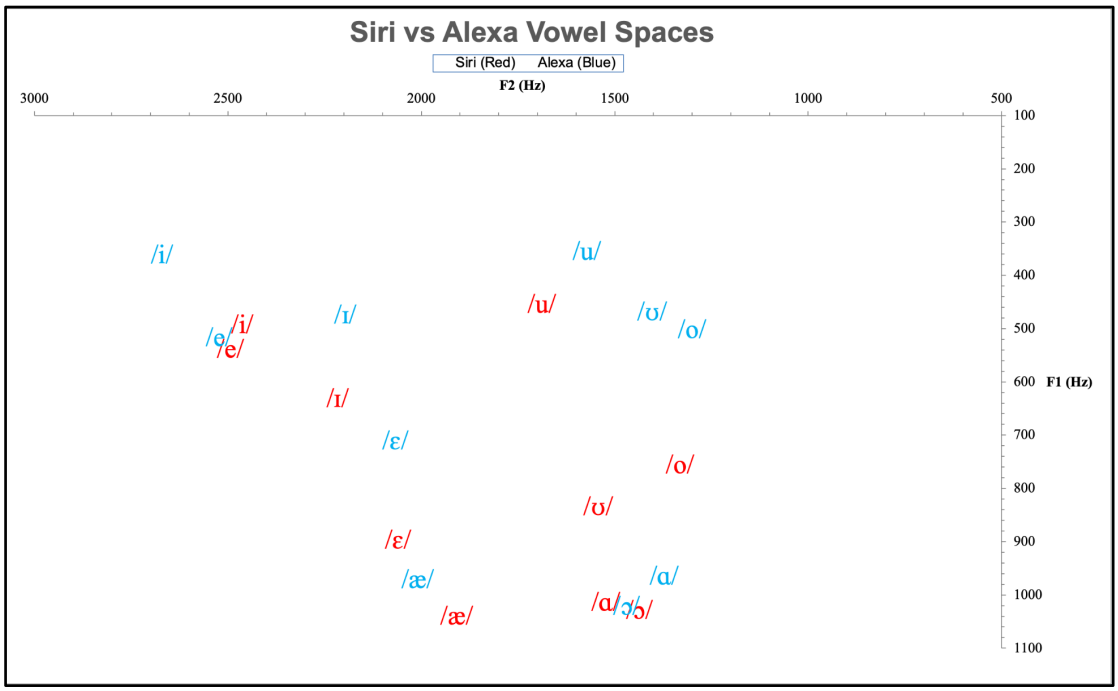
VI	F2 Alexa	F2 Siri	F2 mean diff	*t*-value	Pr (> |t|)
/i/	2670	2462	208	2.67	0.076
/ɪ/	2196	2216	-20	-0.24	0.829
/e/	2522	2493	30	0.58	0.603
/ɛ/	2067	2060	7	0.13	0.906
/æ/	2009	1908	101	5.68	0.011
/u/	1572	1688	-116	-0.73	0.520
/ʊ/	1403	1542	-139	-1.00	0.423
/o/	1300	1331	-31	-0.34	0.757
/ɔ/	1469	1436	33	1.36	0.306
/ɑ/	1372	1523	-151	-2.51	0.087

As shown in Table 5, the majority of differences in F2 values between Siri and Alexa were not significant based on paired t-tests. Alexa's /æ/ F2 was significantly greater than Siri's (101 Hz, $p = 0.011$), indicating a "fronter" production for Alexa. There was also a marginal difference in F2 for /i/ (208 Hz, $p = 0.076$) and /ɑ/ (-151 Hz, $p = 0.087$), suggesting that Alexa's /i/ was substantially "fronter" than Siri's and its /ɑ/ was significantly backer than Siri's. While not statistically significant, Siri had a higher average F2 value for /o/, /ʊ/, /u/ and a lower average F2 value for /ɔ/, /ɛ/, /e/. Again, it is not surprising that there were relatively few significant differences in formant frequency for specific vowels given the relatively small sample size for each VI.

3.4.2.3 Mapping F1-F2 Plane. Taking results from 3.4.2.1 and 3.4.2.2 into account, Siri and Alexa's vowel spaces are presented in Figure 4. In order to provide an approximate visual impression of where the vowels would be produced in the human oral cavity, values on both F1 and F2 axes were reversed.

Figure 4

F1-F2 Vowel Space Mapping of Siri and Alexa



As shown in Figure 4, Alexa had a larger vowel space than Siri. For front vowels, Alexa's vowels (marked in blue above) were produced further forward than (or similar to) front vowels produced by Siri. For back vowels, Alexa's vowels were produced farther back than back vowels produced by Siri, excluding /ɔ/, for which Alexa's average F2 was only 33 Hz greater than Siri's. Again, however, overall only one vowel showed a statistically significant difference in average F2 (/æ/). The difference in average F2 between Alexa's two most extreme vowels was 1,370 Hz, compared with 1,162 Hz for Siri. Alexa also had a somewhat wider spread in F1, with a

difference in average F1 of 667 Hz for Alexa's two most extreme vowels, compared with 585 Hz for Siri.

/ɔ/ for both Siri and Alexa had a higher F1 value than /ɑ/. Siri's /ɔ/ had an F1 of 1,027 Hz while /ɑ/ had an F1 of 1,012 Hz. Alexa's /ɔ/ had an F1 of 1,020 Hz while /ɑ/ had an F1 of 963 Hz. Additionally, Alexa's /ɑ/ yielded a lower F2 than /ɔ/, at 1,372 Hz and 1,469 Hz respectively.

3.4.3 Discussion

Siri's F1 values were consistently higher than Alexa's F1 values per VI. As such, Siri's vowels were lower than Alexa's. As discussed in 2.4.2, Leung et al (2021) found that among Australian English speakers, female speakers with higher formant values were perceived as more feminine than female speakers with lower formant values. Therefore, Siri's significantly higher F1 values compared to Alexa may be a factor that could lead listeners to perceive Siri as more feminine-sounding than Siri, especially if the listener is a speaker of Australian English.

As shown in Figure 4, the differences in F2 values for Siri and Alexa support that Alexa had a wider vowel space than Siri. As mentioned in 2.4.2, wider vowel space can lead to perceptions of increased articulateness, meaning that vowel dispersion may favor Alexa over Siri in terms of being perceived as more articulate by listeners.

The F1-F2 plane also demonstrated that Siri and Alexa's TRAP vowels were not farther back than the average TRAP F2 range of 1,626–1,895 Hz for an adult California-English speaking female (Villarreal 2018). This finding indicates that neither Siri or Alexa would be likely to be perceived as speakers of the CVS (as discussed in 2.4.2). That said, Siri was closer to exhibiting the CVS than Alexa since Siri had a significantly backer TRAP vowel than Alexa. Interestingly, /æ/ was the only significantly different F2 value found in section 3.4.2.2.

Additionally, Siri's LOT/THOUGHT vowels were condensed within the F1-F2 plane, and Siri generally had lower front lax vowels (significantly for /ɛ/, marginally significantly for /ɪ/ and /æ/) than Alexa, and Siri's back vowels tended to be fronter than Alexa's, though not statistically significantly so based on the present small sample. These features together make Siri's vowel space closer to the CVS than Alexa's vowel space. Figure 4 showed that Siri and Alexa demonstrated elements of a LOT/THOUGHT merger since Siri and Alexa produced /ɔ/ lower than /ɑ/. For Alexa, /ɔ/ was produced more fronted than /ɑ/. With this in mind, it's possible that listeners would perceive Siri as having more of the traits or personae associated with the CVS than Alexa (Villarreal, 2018, citing Podesva, 2011; Fought, 1999; Eckert, 2008).

Siri's proximity to the CVS does not conclusively indicate that Siri would be perceived as having CVS-associated traits. There are additional vocal qualities that influence listener perception of speaker dialect and persona, and the indexical values associated with particular vowels vary by listener and community, as discussed in section 2.4.2.

3.5 Study 3: /s/

As discussed in section 2.4.3, /s/ production has been frequently studied with respect to gender and sexuality. /s/ production with higher peak frequencies and centers of gravity (COG) tends to be perceived as having a more feminine quality, while /s/ production with lower centers of gravity tends to be perceived as having a more masculine quality, though perceptions of speakers with different /s/ production vary greatly depending on social and cultural norms or expectations. /s/ production has also been connected to rural versus urban geographies, age, and social class.

3.5.1 Methodology

Following general methodology outlined in 3.2, Study 3 analyzed recordings from Siri and Alexa reciting “The Rainbow Passage” (see Appendix B). Sixteen word-initial /s/ tokens were identified within the passage. Relevant /s/ tokens were isolated and a spectral slice was taken at 50% of the sibilant. Following Styler 2022, settings in Praat for the spectral slice were set to a Hamming window and the window length was set to 0.025 ms. Center of Gravity (COG) was then calculated in Praat based on the spectral slice with a power of 2.

3.5.2 Results

Results from Study 3 showed Alexa had a higher COG than Siri for every token of word-initial /s/ in "The Rainbow Passage." In Table 6, cases of "size" and “since” followed by a numeral mean that there were more than one token of that word in the passage. The numeral identifies which occurrence was being recorded.

Table 6

Center of Gravity (Hz) for Word-Initial /s/ in "The Rainbow Passage" for Siri and Alexa

Token	Siri	Alexa
sunlight	7483	8043
strikes	6859	7898
something	7424	7551
say	7711	7827
centuries	7085	7923
some	7646	7959
sign	7528	7835
sky	7696	8045
sun's	7115	7992
since.1	7940	8010
size.1	7125	7983
size.2	7593	8068
said	7367	7964
super-imposition	7106	7577
second	7499	7982
since.2	7417	8018
Average COG	7412	7917

Siri had an overall average COG of 7,412 Hz while Alexa had an overall average COG of 7,917 Hz . Alexa's average COG was significantly higher than Siri's average COG (t = -6.1888, $p < 0.001$). Future research would be needed to see how the phonological environment influenced /s/ COG.

3.5.3 Discussion

Results from 3.5.2 demonstrated that Siri and Alexa produced word-initial /s/ tokens within a typical /s/ peak frequency range for women, which, according to Zimman 2017 (citing Flipsen et al., 1999; see also Schwartz, 1968) is 6,500 Hz and 8,100 Hz for American-English

speaking women, in the higher portion of the range. Alexa had an overall significantly higher average COG for /s/ than Siri, which could mean that Alexa would likely be perceived as more feminine than Siri if perceptions were based on /s/ alone.

Based on Podesva and Van Hofwegen's (2016) findings (discussed in 2.4.3,) both Siri and Alexa's average /s/ COG was more aligned with an average COG for a straight adult female (7,743-5,863 Hz) than a lesbian adult female (6,120-4,955 Hz) in the Redding, California, community they studied. Siri and Alexa's /s/ COG were also closer to the range for younger participants (6,555-6,930 Hz) than for older participants (5,544-6,120 Hz) in that study. While Siri and Alexa's /s/ COG were more closely aligned with average COG of younger rural females (6,930 Hz), Podesva and Van Hofwegen's (2016) findings suggest the younger rural speaker average COG was higher than the younger town speaker COG because rural speakers were attempting to sound more urban. As such, a higher COG value would align more with perceptions of a speaker being from an urban environment. Because average COG for Siri and Alexa had about a 500 Hz difference, it is also possible listeners would be perceived differently from each other relative to these social groupings. A perception study would be needed to analyze further.

In summary, while Alexa's higher /s/ COG average taken alone might lead listeners to perceive Alexa as more feminine than Siri, both virtual assistants align with /s/ values consistent with straight adult females from a more urban community.

3.6 Study 4: -ING

As discussed in section 3.4.4, -ING pronounced as *-ing* instead of *-in* is typically associated with articulateness and higher social class status in casual speaking scenarios (Labov et al., 2011; Tagliamonte, 2004). In addition, when American English speakers are asked to read,

speakers use *-ing* more frequently, regardless of socio-economic class (Labov et al. 2011). Of note, perceptions of *-ing* versus *-in* vary depending on listener usage and preconceived notions of a variety (Campbell-Kibler, 2008).

3.6.1 Methodology

Following the general methodology outlined in 3.2, Study 4 used prompted Siri and Alexa to recite "The Boy Who Cried Wolf" (see Appendix C) and the "Rainbow Passage" (see Appendix C). There were six word-final -ING tokens produced in each passage, resulting in a total of twelve relevant tokens. For each token, the spectrogram was visually analyzed for *-ing* and *-in*. Alongside an analysis of the soundwaves, the author used personal judgment to confirm the presence of *-ing* versus *-in*.

3.6.2 Results

As shown in Table 7, Siri and Alexa pronounced -ING as *-ing* instead of *-in* for every token of word-final -ING in the passages, "The Boy Who Cried Wolf" (see Appendix D) and "Rainbow Passage" (see Appendix C).

Table 7

[iŋ] Versus [in] in Read -ING Final Words for Siri and Alexa

Token	The Boy Who Cried Wolf				Token	Rainbow Passage			
	Siri		Alexa			Siri		Alexa	
	[iŋ]	[in]	[iŋ]	[in]		[iŋ]	[in]	[iŋ]	[in]
raising	x		x		long	x		x	
shouting	x		x		according	x		x	
looking	x		x		boiling	x		x	
overcoming	x		x		something	x		x	
racing	x		x		looking	x		x	
trying	x		x		showing	x		x	

Tokens were recitations of each passage by Siri and Alexa and were not open-answered response. These results, therefore, provide only partial insight into Siri and Alexa's -ING pronunciation patterns. It is possible that one or both of the assistants use *-in* in some non-read contexts.

3.6.3 Discussion

Based on discussions in section 2.4.3, *-ing* is more commonly produced with read speech regardless of other factors such as social class. Perhaps unsurprisingly, in 100% of cases, *-ing* was pronounced when Siri and Alexa were asked to recite the passage (see Appendix D). Furthermore, *-ing* is more commonly used by and associated with women than men (Tagliamonte, 2004). Additionally, as discussed in section 2.4.3, rates of read speech *-in* production for lower-middle class and upper-middle class speakers was about 0% (Labov et al., 2011). The virtual assistants consistently producing *-ing* (along with other vocal features) could

lead users to perceive Siri and Alexa as having voices similar to lower- or upper-middle class women.

Based on the invariant production of *-ing* shown in Table 7, it seems likely that Siri and Alexa would prefer *-ing* when answering open-ended questions as well, though there may be some cases of *-in* depending on the context of the conversation. As a preliminary look into *-ing* usage with open-ended responses, Siri and Alexa's usage of word final *-ing* was analyzed in recordings from emotion- and opinion-based responses in study 1.2. Siri used *-ing* for all four tokens and Alexa used *-ing* for seven out of eight tokens. This finding supports that Siri and Alexa prefer the *-ing* variant in read and open-ended speech.

Broadly speaking, *-ing* has been associated with more articulate-sounding speakers in casual conversation. Rates of *-ing* have been shown to increase with read speech regardless of social class (Labov et al., 2011). Siri and Alexa were designed to assist and offer users information. In general, if there is an ideological link between articulateness and reliability in providing accurate information, one might expect both virtual devices employ phonetic features like *-ing* that index articulateness.

It should be considered that perceptions of *-ing* may have a negative impact on the listener if the listener is more familiar with hearing *-in*. As mentioned in section 2.4.3, using *-ing* may lead to perceptions of condescension or forced formality if the speaker is more familiar with people using the *-in* variety on a regular basis. It would be interesting to test how users from various dialectal backgrounds (dialects using more *-in* than *-ing* and vice versa) perceive Siri and Alexa based on -ING pronunciation.

3.7 Summary

In Chapter 3, Studies 1-4 have provided a starting point for understanding the socio-indexical qualities that may be indexed by Siri and Alexa's voices and phonetic qualities. Study 1 provided detailed information on F0 patterns for Siri and Alexa, Study 2 outlined the virtual assistant's vowel spaces and assessed differences in F1 and F2 frequencies, Study 3 detailed /s/ COG, and Study 4 looked at differences in -ING pronunciation. Findings from Chapter 3 begin to create connections between sociophonetic findings in human-voice research to phonetic qualities of machine voices and resulting perceptions. Chapters 4 and 5 will outline more distinctly the implications of this work and its significance in future human-machine communication research.

Chapter 4: General Discussion

Research cited in section 2.4 demonstrated that phonetic features such as F0, formant frequencies, /s/ COG, and -ING pronunciation shape listener perceptions of a voice. Studies discussed in section 2.4 were primarily focused on human voices, but results and discussions presented in chapter 3 demonstrate how sociophonetic insights could be applied to machine-driven voices.

Interestingly, human inclination to associate human characteristics with robots has long preceded technological capabilities. In watching films and television with robot characters, viewers can develop attitudes toward movie robots similar to how judgments would be made about human actors in the production, e.g., Rosey the Robot from *The Jetsons* and Fembots from *Austin Powers: International Man of Mystery* (see section 2.1.1.2). As discussed in section 2.1, people who work in film and television have created movie robots who take on specific roles, jobs, personality traits, physical embodiments, etc., and have never been limited by the realities of actual technology development. Physical manifestations sometimes played a part in perceptions of the movie robot (as seen with the obviously mechanical and artificial Rosey from *The Jetsons*) but other times, the machine's voice (along with the content of the speech) was all human viewers were given to create character judgments. For example, Hal from *2001: A Space Odyssey* (see section 2.1.1) had no humanoid physical manifestation. Viewers had to rely on Hal's shift from apparently benign to threatening by paying attention to changes in voice, such as when its speech rate slowed and Hal began distorting its voice to sound stereotypically mechanical.

Now, robots exist in many homes globally via virtual assistants, as discussed in section 2.2.1.2. Section 2.1's discussion of the ways in which humans perceive movie robots paired with

section 2.2's exploration of anthropomorphism and the CASA paradigm make it possible that humans are also making personality judgments about their virtual assistants based on their voice, consciously or otherwise. This work focused on virtual assistants because all of the popular virtual assistants (Siri, Alexa, Microsoft's Cortana, Google Home, etc.) were introduced to the market with female-sounding default voices. These machines have been anthropomorphized by human users and seen as female entities in the home (Druga et al., 2017; Schweitzer et al., 2019; Purington, 2017; see section 2.2.2), again creating a connection between machines being treated as similar to human interlocutors with at least some human personality traits associated with them. As stated in section 2.2.2.1, Schweitzer et al. (2019) found that people who viewed their virtual assistants as a servant described the machine as a "nice, friendly, helpful, reliable person with a ready-to-please character, who acts professionally, as well as somewhat subserviently, and remotely" (p. 703). While all of these qualities were not discussed from a linguistic perspective in this thesis, the description lays the foundation for the virtual assistant's expected personality or character traits.

As examined in section 2.3, female voices have not historically been perceived as indexing intelligence (Carson, 1995; Hoegaerts, 2020; Levander, 1998; Neufeld, 2021). Instead, female voices were restricted to being socially accepted in domestic domains or to entertain in public spaces, with any opinions made by female voices being quickly disregarded. At the same time, women's voices were viewed as crucial for teaching children vocal health since the mother's voices were associated with having a "civilizing influence, pointing children toward moral behavior and good taste" (Hoegaerts, 2020, p. 448). While women's voices have historically been viewed as inferior to men's voices (see section 2.3), female voices have also been more closely associated with emotional intelligence and viewed as warm and friendly in a

domestic space (Borau et al. 2021). Since Siri and Alexa typically exist in private spaces of user's lives, those qualities may over-power historical perceptions of female voices as superfluous noise (as described in section 2.3.1.1). As discussed in section 2.3, female-sounding robot voices have also been shown to be perceived as less powerful, but more passive, submissive, truthful, knowledgeable, and vulnerable than male-sounding robot voices (Chebat et al., 2007; Gal, 1989; Mullinex et al., 2003). Additionally, human perceptions of female movie robots (see section 2.1) and female voices (see section 2.2) make it likely that female-characterized machines would also be viewed as more maternal, trustworthy, and transparent than male-characterized machines. This is especially likely if the female-sounding machine operates in a domestic space.

Based solely on historical perceptions of female voices and artificial language technological development built with male-voice prototypes, one would expect virtual assistants to have been created with a male-sounding default voice, but many of the aforementioned social characteristics could help explain why virtual assistants tend to be voiced as women. One can see why it might be considered desirable for a virtual assistant's voice to sound trustworthy, maternal, and non-threatening. While the intentions of people who developed and programmed the devices cannot be definitively discussed in this work, findings from this work's assessment of human perception of movie robots, human perception of female voices, and the initial default virtual assistant voices creates a connection and foundation for further linguistic analysis.

This work then aimed to examine how specific social and phonetic qualities found in human language could potentially apply to Siri and Alexa's female-sounding voices. F0, vowel formant frequencies, /s/ COG, and -ING variation were analyzed as a starting point to further understand the socio-indexical properties conveyed by Siri and Alexa through voice. Discussed

in further detail below, findings from each study do not necessarily index the same characteristics for each assistant. It is important to remember that each phonetic feature is one piece of a larger conversation about socio-indexical properties of voices and cannot be the sole factor for determining listener perception.

Study 1 focused on F0. Based on previous research discussed in 2.4.1, In terms of femininity, it seems that listener ability to first identify speaker gender impacts masculinity - femininity rankings (Leung et al., 2021). Because Siri and Alexa's overall average F0 aligns with the average F0 range of an adult female 200-220 Hz (Zimman, 2017), it is likely that listeners would easily identify Siri and Alexa as female-sounding, and therefore would rank them in terms or more or less feminine. A higher F0 for females is generally perceived as more attractive, with lower F0 ranges being associated with speakers that are taller, more dominant, and stronger (Borkowska & Pawlowski, 2011; Feinberg et al., 2008; Kempe, 2013). Smaller F0 values can also related to non-human qualities like being lighter, rounder, lighter, more positive (D'Onofrio & Eckert, 2021). Results from Study 1 showed that Siri had a significantly higher F0 than Alexa for VIs at 242 Hz vs 228 Hz ($p = 0.024$). Siri also had a higher average F0 value than Alexa at the sentential level. Siri's higher F0 values, when all else is held constant, could lead listeners to viewing the virtual assistant as smaller, more feminine, more attractive, and less dominant than Alexa. It is important to keep in mind that female-sounding voices were shown to be perceived as less dominant or threatening than male-sounding voices overall (Chebat et al., 2007; Kempe et al., 2013; Mullennnix et al., 2003), which is a key factor in successfully introducing the technology into human's lives. As seen in section 1.2, humans want to feel superior to robots.

When considering the goal of creating "human-like" conversational abilities for virtual assistants, results from Study 1 demonstrated F0 change within each sentence and based on

conversational intent. Siri and Alexa had higher F0 values for every first VI token compared to final VI token in each stimulus sentence. This result is consistent with the falling human intonation patterns found in declaratives (see section 3.3.3; Cruttenden, 2008). Additionally, both Siri and Alexa demonstrated a significant increase in F0 when responding to an opinion-based question versus providing factual information ($p = 0.009$). Awareness of conversational context shown through this F0 variation is what would be expected of a human interlocutor. The machine may not be performing at the same level as a human interlocutor would, but the data provides preliminary support that machines are reacting with an awareness of conversational type.

Study 2 focused on vowel formant frequencies. Similar to F0, higher vowel formant frequencies have been associated with female voices and increasing the formant frequency leads to perceptions of increased femininity (Gelfer & Bennett, 2013; Leung et al., 2021). Siri had higher average F1 values than Alexa for every VI (see section 3.4.). Siri might be perceived as more feminine than Alexa based on F0 and F1. On the other hand, Siri had a narrower vowel space than Alexa. According to Kempe et al. (2013) in section 2.4.2, a narrower formant space is typically associated with male voices that index more masculine qualities (e.g., taller, bigger, increased threat potential). It is unclear if the threat level only applies to distinctly male-sounding voices, but it still remains that a smaller vowel space would lead listeners to perceive Alexa as more feminine. Wider vowel spaces have also been known to index articulateness (Kempe et al., 2013), meaning it is possible that Alexa would also be perceived as more articulate than Siri based on vowel space. In terms of dialect, Study 2 revealed that while Siri was more closely aligned with the CVS and Alexa had a significantly fronter TRAP F2 ($p = 0.011$), neither virtual assistant seemed to be completely consistent with the CVS. Siri and Alexa did have a

LOT/THOUGHT merger, which could be aligned with more general Californian English (Villarreal, 2018).

An assessment of /s/ COG in Study 3 adds another layer of analysis to Siri and Alexa's voices. /s/ has been linked to indexing sexuality, urban versus rural geography, and age. While much of the sexuality and /s/ research has focused on male speech, Podesva and Van Hofwegen (2016) found in their study of a population in Redding, California, that speakers who identified as lesbian had lower /s/ COG values (4,955-6,120 Hz) than speakers who identified as straight women (5,863-7,743 Hz). If this sample is consistent with other adult female /s/ COG values in the United States, Siri and Alexa's /s/ COG would likely push them toward being perceived as straight females. As discussed in 3.5.3, Siri and Alexa would also be categorized as younger, non-rural speakers based on their /s/ COG.

Finally, -ING pronunciation was assessed for each virtual assistant. Section 2.4.3 described -ING's associations with formality and social classes. Study 4 results showed 100% usage of *-ing* variant for Siri and Alexa's read speech. Usage of *-ing* over *-in* might lead listeners to perceive Siri and Alexa as being of lower-middle to upper-middle class women. It should be noted that perceptions of -ING (as is true with perceptions of other phonetic variables) depends heavily on preconceived notions the listener may have (Campbell-Kibler, 2008). If the person using Siri or Alexa rarely uses or hears *-ing,* that person may view the virtual assistant as pretentious or snobby. It is always crucial to consider how social indexical characteristics conveyed by voices can shift based on listener experience.

Research presented in this work demonstrates how findings in human-focused linguistics could be applied to machine voices in order to better understand the roles conversational agents play in human-machine communication, specifically in terms of the impressions they make on

users. Additionally, it is possible that because humans anthropomorphize virtual assistants and have been found to view devices as social interlocutors, traits associated with those devices may impact how humans perceive other humans in similar roles (e.g., assistants) or with similar voices. For example, if Siri was perceived by a user to be articulate or trustworthy and there was evidence that *-ing* pronunciation played a key part in creating that perception, a human may in turn expect a human assistant to employ the same level of *-ing* usage. If the human user varies drastically from choices made by Siri's voice, there may be negative social consequences such as a perception that the human is not a trustworthy assistant.

Results from Chapter 3 offered insights into particular versions of Siri and Alexa's voices. Voice-based technology products change quickly, with new versions being released regularly. Since the conception of this work, Siri has released a gender non-specific voice for users to select, bringing the total to five different voice options for users to select as Siri's voice (Porter 2022). The value of this work is, therefore, not solely in examining the socio-indexical characteristics of particular versions of Siri and Alexa's voices. Rather, the value also lies in more broadly exploring and demonstrating how sociophonetic characteristics indexed by human voices could apply to machine voices and be useful in future studies of human-machine interactions. Conversations at this level of linguistic analysis could help inform future technology development and uncover new findings in perceptions made by humans based on voice.

Chapter 5: Summary, Considerations, and Looking Ahead

5.1 Summary

This work aimed to better understand how human voices, social expectations, and the phonetic particulars of machine voices relate to how machines are perceived socially. As a starting point, sociolinguistic findings regarding human perception of human voices were applied to virtual assistant voices. Perceptions made about movie robots demonstrated that prior to robots even being a regular part of real human life, humans had expectations for movie robot behavior and personalities, sometimes based largely on voice. Background context provided in Chapter 2, along with the CASA framework, provided a foundation to discuss virtual assistants as active interlocutors in human-machine conversations.

With virtual assistants established as active interlocutors, it was then acknowledged that both virtual assistant voices in this work sound like women. Therefore, it was important to examine historical and contemporary perceptions of women's voices in order to provide context for considering the phonetic details of the assistants' voices and why they may have been assigned their respective voices. Building on previous research on sociophonetic perception, this work focused on what socio-indexical properties were characterized by variables such as F0, vowel formant frequencies, /s/ COG, and -ING.

Findings from Chapter 3 then linked those sociophonetic findings to Siri and Alexa. Insights reported in Chapter 3 provided data that could inform hypotheses about how users would perceive each voice relative to one another and more generally. Based on the data presented in this work, it is possible that listeners may be more likely to perceive both Siri as more feminine based on F0 and F1, but Alexa as more feminine based on a wider vowel space. Siri may be more closely associated with CVS, but neither virtual assistant was quite in line with

the properties of the dialect based on this work's data. If the listener is sensitive to -ING variation, Siri and Alexa may be perceived as articulate, lower-middle or upper-middle class speakers. Siri and Alexa's /s/ COG is consistent with that of straight, younger, non-rural females in previous studies (Podesva & Van Hofwegen, 2016).

5.2 Considerations

One consideration is that all data collected in Chapter 3 (except Study 1.2 from Allen, 2022) relied on Siri and Alexa reciting written words. Reciting from text was necessary in order to ensure certain tokens of interest were said by the machine voices, but this format may have restricted variation in phonetic features such as F0 or -ING pronunciation.

As related to -ING pronunciation, it would have been beneficial to compare results from 3.6 with how the virtual assistants pronounce -ING when providing responses to open-ended user questions or prompts. This thesis accounted for open-ended responses when considering -ING pronunciation in recordings from Study 1.2, but a more robust data set would be needed to explore the concept in depth. It is possible the virtual assistant will always have a high rate of *-ing* considering the machine employs TTS technology in order to speak. Assessing how machine language systems pronounce -ING in open-ended prompt scenarios would be of particular interest considering *-ing* is typically used more by human speakers in cases where they are reading text (as described in 2.4.3).

5.3 Looking Ahead

Future projects related to human-machine spoken communication could include user perception studies of virtual assistant voices in order to see how the voices are perceived and how that relates to results from sociophonetic perception studies of human English speakers (as

cited in 2.4, for instance). User perceptions in conjunction with phonetic analysis of a particular virtual assistant offer insight into how user perception shifts between speaking with a human assistant versus a machine assistant. For example, if users were surveyed and reported Siri and Alexa to sound inarticulate, these findings would clash with expected results given their high rates of *-ing*, which would mean that humans perceive conversational robots differently than they perceive humans sociophonetically.

More virtual assistant voices should be analyzed following methodologies laid out in Chapter 3. As mentioned previously, Siri now offers five different voices for users to choose—there is no pre-set default Siri voice in the newer versions of the assistant. It would be valuable to compare phonetic data from each voice in order to understand each voice's socio-indexical characteristics. This set of voices would also allow for comparisons between male, female, and non-binary virtual assistant voices.

Additionally, future research could also focus on expanding (a) the range of linguistic variables explored and (b) the social traits being tested. For example, qualities such as F0 variation and word-final t-release may provide further insights into personalities associated with a particular voice. It would also be beneficial to test what other social qualities are indexed by machine voices (e.g., what vocal qualities make a voice be perceived as “nice” or “arrogant”).

Finally, discussion from this thesis indicates that machine voices could potentially be perceived as reinforcing stereotypes or containing biased data. Understanding the vocal qualities attached to persona judgments could help technology companies mitigate implicit bias and be more sensitive to voice properties as they develop virtual assistants or other conversational machines.

Similar to the cycle between sci-fi movie robot creation and real-life robot development (discussed in 2.1), there needs to be a cycle between sociophonetic research and conversational voice design. Each field of study can provide insights to the other and result in a more holistic view of human-machine communication and social perception as technology develops and progresses.

References

Allen, A. (2022). An assessment of Apple Siri and Amazon Alexa's F0 values as related to vocal attractiveness. *SpoHuMa21*, 17–22. https://doi.org/10.6094/UNIFR/223819

Anderson, S. L. (2008). Asimov's "three laws of robotics" and machine metaethics. *AI & Society*, *22*(4), 477–493. https://doi.org/10.1007/s00146-007-0094-5

Barbera, J., & Hanna, W. (1962). Rosie the robot (No. 1). In *The Jetsons*. Hanna-Barbera Productions.

Barkana, B. D., & Zhou, J. (2015). A new pitch-range based feature set for a speaker's age and gender classification. *Applied Acoustics*, *98*, 52–61. https://doi.org/10.1016/j.apacoust.2015.04.013

Bartneck, C. (2013). *Robots in the theatre and the media*. https://doi.org/10.13140/RG.2.2.28798.79682

Bartneck, C. (2017). From fiction to science – A cultural reflection of social robots. *Proceedings of the CHI2004 Workshop on Shaping Human-Robot Interaction.* https://doi.org/10.6084/M9.FIGSHARE.5154820

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bekker, I. (2007). Fronted /s/ in general white South African English. *Language Matters*, *38*(1), 46–74. https://doi.org/10.1080/10228190701640025

Bekker, I., & Levon, E. (2017). The embedded indexical value of /s/-fronting in Afrikaans and South African English. *Linguistics*, *55*(5), 1109–1139. https://doi.org/10.1515/ling-2017-0022

Belanche, D., Casaló, L. V., Schepers, J., & Flavián, C. (2021). Examining the effects of robots' physical appearance, warmth, and competence in frontline services: The Humanness-Value-Loyalty model. *Psychology & Marketing*, *38*(12), 2357–2376. https://doi.org/10.1002/mar.21532

Bell, L. (2003). *Linguistic Adaptations in Spoken Human-Computer Dialogues - Empirical Studies of User Behavior* (PhD dissertation, Institutionen för talöverföring och musikakustik). http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-3607

Bergen, H. (2016). 'I'd blush if I could': Digital assistants, disembodied cyborgs and the problem of gender. *Word and Text*, *6*, 95–113.

Bernotat, J., Eyssel, F., & Sachse, J. (2021). The (fe)male robot: How robot body shape impacts first impressions and trust towards robots. *International Journal of Social Robotics*, *13*(3), 477–489. https://doi.org/10.1007/s12369-019-00562-7

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Borau, S., Otterbring, T., Laporte, S., & Fosso Wamba, S. (2021). The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI. *Psychology & Marketing*, *38*(7), 1052–1068. https://doi.org/10.1002/mar.21480

Borkowska, B., & Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour*, *82*(1), 55–59. https://doi.org/10.1016/j.anbehav.2011.03.024

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, *121*(1), 41–57. https://doi.org/10.1016/j.cognition.2011.05.011

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems*, *42*(3–4), 167–175. https://doi.org/10.1016/S0921-8890(02)00373-1

Breitenstein, C., Lancker, D. V., & Daum, I. (2001). The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition & Emotion*, *15*(1), 57–79. https://doi.org/10.1080/02699930126095

Brown, C. (2019). Sex robots, representation, and the female experience. *The American Papers*, *37*(67), 111–113.

Bruckert, L., Liénard, J.-S., Lacroix, A., Kreutzer, M., & Leboucher, G. (2006). Women use voice parameters to assess men's characteristics. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1582), 83–89. https://doi.org/10.1098/rspb.2005.3265

Burton, L. (1999). *Smart house*. [Motion picture]. Disney.

Cameron, D. (1997). Performing gender identity: Young men's talk and the construction of heterosecual masculinity. In *Language and masculinity* (pp. 47–64). Blackwell.

Cameron, D., McAlinden, F., & O'Leary, K. (1989). Lakoff in context: The social and linguistic function of tag questions. In *Women in their speech communities* (pp. 74–93). Longman.

Campbell-Kibler, K. (2008). I'll be the judge of that: Diversity in social perceptions of (ING). *Language in Society*, *37*(5), 637–659. https://doi.org/10.1017/S0047404508080974

Campbell-Kibler, K. (2011). Intersecting variables and perceived sexual orientation in men. *American Speech*, *86*(1), 52–68. https://doi.org/10.1215/00031283-1277510

Carson, A. (1995). The gender of sound. In *Glass, irony, and God* (pp. 119–142). New Directions Book.

Cartei, V., & Reby, D. (2013). Effect of formant frequency spacing on perceived gender in pre-pubertal children's voices. *PLoS ONE*, *8*(12), e81022. https://doi.org/10.1371/journal.pone.0081022

Chambers, J. K., Trudgill, P., & Schilling-Estes, N. (Eds.). (2002). *The handbook of language variation and change*. Blackwell Publishers.

Chebat, J.-C., Hedhli, K. E., Gélinas-Chebat, C., & Boivin, R. (2007). Voice and persuasion in a banking telemarketing context. *Perceptual and motor skills*, *104*(2), 419–437. https://doi.org/10.2466/pms.104.2.419-437

Choi, T. R., & Drumwright, M. E. (2021). "OK, Google, why do I use you?" Motivations, post-consumption evaluations, and perceptions of voice AI assistants. *Telematics and Informatics*, *62*, 101628. https://doi.org/10.1016/j.tele.2021.101628

Cohn, M., & Zellou, G. (2019). Expressiveness influences human vocal alignment toward voice-AI. *Interspeech 2019*, 41–45. https://doi.org/10.21437/Interspeech.2019-1368

Crain, R. (2019). *Gender expression and the styling of queer women's speech*. Eastern Michigan University.

Crawford, M. (2003). Gender and humor in social context. *Journal of Pragmatics*, *35*(9), 1413–1430. https://doi.org/10.1016/S0378-2166(02)00183-2

Crist, S. (1997). Duration of onset consonants in gay male stereotyped speech. *University of Pennsylvania Working Papers in Linguistics*, *4*(3), 53–70.

Cruttenden, A. (1981). Falls and rises: Meanings and universals. *Journal of Linguistics*, *17*(1), 77–91. https://doi.org/10.1017/S0022226700006782

Cukor, G. (1964). *My fair lady*. [Motion picture]. Warner Brothers.

Deterding, D. (2006). The north wind versus a wolf: Short texts for the description and measurement of English pronunciation. *Journal of the International Phonetic Association*, *36*(2), 187–196.

D'Onofrio, A. (2015). Persona-based information shapes linguistic perception: Valley girls and California vowels. *Journal of Sociolinguistics*, *19*(2), 241–256. https://doi.org/10.1111/josl.12115

D'Onofrio, A., & Eckert, P. (2021). Affect and iconicity in phonological variation. *Language in Society*, *50*(1), 29–51. https://doi.org/10.1017/S0047404520000871

Druga, S., Williams, R., Breazeal, C., & Resnick, M. (2017). "Hey Google is it OK if I eat you?": Initial explorations in child-agent interaction. *Proceedings of the 2017 Conference on Interaction Design and Children*, 595–600. https://doi.org/10.1145/3078072.3084330

Eckert, P. (2008). Where do ethnolects stop. *International Journal of Bilingualism*, *12*(1–2), 25–42.

Eckert, P. (2009). Where does the social stop? In *LanguagevVariation—European perspectives III : Selected papers from the 5th International Conference on Language Variation in Europe (ICLaVE 5)* (pp. 13–29). John Bejamins Publishing Company.

Eckert, P., & McConnell-Ginet, S. (2013). *Language and gender* (2nd ed.). Cambridge University Press.

Eitan, Z., Schupak, A., Gotler, A., & Marks, L. E. (2014). Lower pitch is larger, yet falling pitches shrink: Interaction of pitch change and size change in speeded discrimination. *Experimental Psychology*, *61*(4), 273–284. https://doi.org/10.1027/1618-3169/a000246

Evans, S., Neave, N., & Wakelin, D. (2006). Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. *Biological Psychology*, *72*(2), 160–163. https://doi.org/10.1016/j.biopsycho.2005.09.003

Fairbanks, G. (1960). *Voice and articulation drillbook.* Harper & Row.

Feinberg, D. R., DeBruine, L. M., Jones, B. C., & Perrett, D. I. (2008). The role of femininity and averageness of voice pitch in aesthetic judgments of women's voices. *Perception (London)*, *37*(4), 615–623. https://doi.org/10.1068/p5514

Fessler, L. (2017). We tested bots like Siri and Alexa to see who would stand up to sexual harassment. *Quartz*. https://qz.com/911681/we-tested-apples-siri-amazon-echos- 14 alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots -stand-up-for-themselves-in-the-face-of-sexual-harassment/.

Fischer, J. L. (1958). Social influences on the choice of a linguistic variant. *Word*, *14*(1), 47–56. https://doi.org/10.1080/00437956.1958.11659655

Flipsen, P., Shriberg, L., Weismer, G., Karlsson, H., & McSweeny, J. (1999). Acoustic characteristics of /s/ in adolescents. *Journal of Speech, Language, and Hearing Research*, *42*, 663–667.

Fought, C. (1999). A majority sound change in a minority community: /U/-fronting in Chicano English. *Journal of Sociolinguistics*, *3*(1), 5–23. https://doi.org/10.1111/1467-9481.t01-1-00060

Gal, S. (1989). Between speech and silence: The problematics of research on language and gender. *IPrA Papers in Pragmatics*, *3*(1), 1–38. https://doi.org/10.1075/iprapip.3.1.01gal

Gambino, A., Fox, J., & Ratan, R. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication*, *1*, 71–86. https://doi.org/10.30658/hmc.1.5

Geenberg, K. (2014). Sound symbolism in adult baby talk (ABT): The role of the frequency code in the construction of social meaning.

Gelfer, M. P., & Bennett, Q. E. (2013). Speaking fundamental frequency and vowel formant frequencies: Effects on perception of gender. *Journal of Voice*, *27*(5), 556–566. https://doi.org/10.1016/j.jvoice.2012.11.008

Haeri, N. (1996). "Why do women do this?" Sex and gender differences in speech. In G. R. Guy, C. Feagin, D. Schiffrin, & J. Baugh (Eds.), *Current issues in linguistic theory* (Vol. 127, p. 101). John Benjamins Publishing Company. https://doi.org/10.1075/cilt.127.08hae

Hall, K. (1995). Lip service on the fantasy lines. In *Gender articulated: Language and the socially constructed self* (pp. 183-216). Routledge. https://doi.org/10.13140/2.1.4942.2402

Hankins, L. K. (1992). Echo and narcissus: Women's voices in classical hollywood cinema. *Hitchcock Annual*, 155–160.

Heffernan, K. (2004). Evidence from HNR that /s/ is a social marker of gender. *Toronto Working Papers in Linguistics*, 23. 71-84. https://twpl.library.utoronto.ca/index.php/twpl/article/view/6208

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111. https://doi.org/10.1121/1.411872

Hoegaerts, J. (2020). Women's voices in educational manuals. The gendered sounds of speech therapy, song and education in Europe c.1830–1900. *Women's History Review*, *29*(3), 444–464. https://doi.org/10.1080/09612025.2019.1611125

Holmes, J. (1982). The functions of tag questions. *English Language Research Journal*, *3*, 40–65.

Holmes-Elliott, S., & Levon, E. (2017). The substance of style: Gender, social class and interactional stance in /s/-fronting in southeast England. *Linguistics*, *55*(5). https://doi.org/10.1515/ling-2017-0020

Honorof, D. N., & Whalen, D. H. (2010). Identification of speaker sex from one vowel across a range of fundamental frequencies. *The Journal of the Acoustical Society of America*, *128*(5), 3095–3104. https://doi.org/10.1121/1.3488347

Humphry, J., & Chesher, C. (2021). Preparing for smart voice assistants: Cultural histories and media innovations. *New Media & Society*, *23*(7), 1971–1988. https://doi.org/10.1177/1461444820923679

IBM Cloud Education. (2021, March 15). Interactive voice response. *IBM Cloud Learn Hub*.

J Chambers. (n.d.). Chambers, J. K., et al. (Eds). (2002). *The Handbook 0/ Language Variation and Change.* Blackwell Publishers.

Jurafksy, D., & Martin, J. (2020). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed.). Stanford University.

Karlsson, I. (1991). Female voices in speech synthesis. *Journal of Phonetics*, *19*(1), 111–120. https://doi.org/10.1016/S0095-4470(19)30306-7

Kempe, V., Puts, D. A., & Cárdenas, R. A. (2013). Masculine men articulate less clearly. *Human Nature*, *24*(4), 461–475. https://doi.org/10.1007/s12110-013-9183-y

Kennedy, R., & Grama, J. (2012). Chain shifting and centralization in California vowels: An acoustic analysis. *American Speech*, *87*(1), 39–56. https://doi.org/10.1215/00031283-1599950

Kidd, A. L. (2021). Superstar to superhuman: Scarlett Johansson, an 'ideal' embodiment of the posthuman female in science fiction and media? *JOMEC Journal*, *0*(16), 52. https://doi.org/10.18573/jomec.209

Kleinberg, S. (2018, January). OK, marketers: Here's what people are saying about their voice-activated speakers. *Think with Google*. Google.

Kubrick, S. (1970). *2001: A Space Odyssey*. [Motion picture]. Metro-Goldwyn-Mayer.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684–686. https://doi.org/10.1126/science.277.5326.684

Kwon, H.-B. (2010). Gender difference in speech intelligibility using speech intelligibility tests and acoustic analyses. *The Journal of Advanced Prosthodontics*, *2*(3), 71. https://doi.org/10.4047/jap.2010.2.3.71

Labov, W. (1972). The social stratification of (r) in New York City department stores. *Sociolinguistic Patterns*, 43–54.

Labov, W., Ash, S., Ravindranath, M., Weldon, T., Baranowski, M., & Nagy, N. (2011). Properties of the sociolinguistic monitor. *Journal of Sociolinguistics*, *15*(4), 431–463. https://doi.org/10.1111/j.1467-9841.2011.00504.x

Lakoff, R. (1973). Language and woman's place. *Language insSociety*, *2*(1), 45–80.

Lang, F. (1927, March). *Metropolis*. Universal Film.

Lapadat, J., & Seesahai, M. (1977). Male versus female codes in informal contexts. *Sociolinguistics newsletter*, *8*(3), 7–8.

Lawrence, A. (1991). *Echo and Narcissus: Women's voices in classical Hollywood cinema*. University of California Press.

Leung, Y., Oates, J., Chan, S.-P., & Papp, V. (2021). Associations between speaking fundamental frequency, vowel formant frequencies, and listener perceptions of speaker gender and vocal femininity–masculinity. *Journal of Speech, Language, and Hearing Research*, *64*(7), 2600–2622. https://doi.org/10.1044/2021_JSLHR-20-00747

Levander, C. F. (1998). *Voices of the nation: Women and public speech in nineteenth-century American literature and. culture* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511582684

Levon, E. (2007). Sexuality in context: Variation and the sociolinguistic perception of identity. *Language in Society*, *36*(04), 533. https://doi.org/10.1017/S0047404507070431

Levon, E. (2014). Categories, stereotypes, and the linguistic perception of sexuality. *Language in Society*, *43*(5), 539–566. https://doi.org/10.1017/S0047404514000554

Liang, Y., & Lee, S. A. (2017). Fear of autonomous robots and artificial intelligence: Evidence from national representative data with probability sampling. *International Journal of Social Robotics*, *9*(3), 379–384. https://doi.org/10.1007/s12369-017-0401-3

Liu, H.-M., Kuhl, P. K., & Tsao, F.-M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science*, *6*(3), F1–F10. https://doi.org/10.1111/1467-7687.00275

McMillan, J. R., Clifton, A. K., McGrath, D., & Gale, W. S. (1977). Women's language: Uncertainty or interpersonal sensitivity and emotionality? *Sex Roles*, *3*(6), 545–559. https://doi.org/10.1007/BF00287838

Moonwomon-Baird, B. (1997). Toward the study of lesbian speech. In *Queerly Phrased* (A. Livia&K. Hall, pp. 202–213). Oxford University Press.

Mullennix, J. W., Stern, S. E., Wilson, S. J., & Dyson, C. (2003). Social perception of male and female computer synthesized speech. *Computers in Human Behavior*, *19*(4), 407–424. https://doi.org/10.1016/S0747-5632(02)00081-X

Munson, B., McDonald, E. C., DeBoe, N. L., & White, A. R. (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics*, *34*(2), 202–240. https://doi.org/10.1016/j.wocn.2005.05.003

Murdoch, C. (2016, October 28). *Want to marry Amazon's Alexa? You're not alone.* Vocativ. https:// www.vocativ.com/371706/amazon-alexa-propose-marriage/index.html

Mutchler, A. (2018, March 28). *A timeline of voice assistant and smart speaker technology From 1961 to today*. Voicebot.Ai. https://voicebot.ai/2018/03/28/timeline-voice-assistant-smart-speaker-technology-1961-today/

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates. *International Journal Human-Computer Studies*, *45*, 669–678.

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology*, *27*(10), 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x

Nass, C., Steur, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78.

Neufeld, C. (2021). *Avid ears: Medieval gossips, sound and the art of listening.* Routledge.

Paek, T., & Chickering, D. M. (2007). Improving command and control speech recognition on mobile devices: Using predictive user models for language modeling. *User Modeling and User-Adapted Interaction*, *17*(1–2), 93–117. https://doi.org/10.1007/s11257-006-9021-6

Perez, S. (2020, February 10). *Nearly 70% of US smart speaker owners use Amazon Echo devices*. TechCrunch. https://techcrunch.com/2020/02/10/nearly-70-of-u-s-smart-speaker-owners-use-amazon-echo-devices/

Phan, T. (n.d.). The materiality of the digital and the gendered voice of Siri. *Transformations*, *29*, 2017.

Phan, T. (2019). Amazon Echo and the aesthetics of whiteness. *Catalyst: Feminism, Theory, Technoscience*, *5*(1), 1–38. https://doi.org/10.28968/cftt.v5i1.29586

Pierrehumbert, J. B., Bent, T., Munson, B., Bradlow, A. R., & Bailey, J. M. (2004). The influence of sexual orientation on vowel production (L). *The Journal of the Acoustical Society of America*, *116*(4), 1905–1908. https://doi.org/10.1121/1.1788729

Podesva, R. (2011). The California vowel shift and gay identity. *American Speech*, *86*, 32–51.

Podesva, R., & Van Hofwegen, J. (2016). (s)exuality in small-town California: Gender normativity and the acoustic realization of (s). In *Language, sexuality and power: Studies in interactional sociolinguistics* (pp. 168–188). Oxford University Press.

Pratt, T., & D'Onofrio, A. (2017). Jaw setting and the California vowel shift in parodic performance. *Language in Society*, *46*(3), 283–312. https://doi.org/10.1017/S0047404517000227

Proyas, A. (2004, July). *IRobot*. [Motion picture]. Twentieth Century Fox.

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017). "Alexa is my new BFF": Social roles, user satisfaction, and personification of the Amazon Echo. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2853–2859. https://doi.org/10.1145/3027063.3053246

Puts, D. A., Apicella, C. L., & Cárdenas, R. A. (2012). Masculine voices signal men's threat potential in forager and industrial societies. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1728), 601–609. https://doi.org/10.1098/rspb.2011.0829

Queen, R. (2007). Sociolinguistic horizons: language and sexuality. *Language and Linguistics Compass*, *1*(4), 314–330. https://doi.org/10.1111/j.1749-818X.2007.00019.x

Re, D. E., O'Connor, J. J. M., Bennett, P. J., & Feinberg, D. R. (2012). Preferences for very low and very high voice pitch in humans. *PLoS ONE*, *7*(3), e32719. https://doi.org/10.1371/journal.pone.0032719

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places* (1. paperback ed., [reprint.]). CSLI Publ.

Rhee, C. E., & Choi, J. (2020). Effects of personalization and social role in voice shopping: An experimental study on product recommendation by a conversational voice agent. *Computers in Human Behavior*, *109*, 106359. https://doi.org/10.1016/j.chb.2020.106359

Risley, J. (2015, November 17). *One year after Amazon introduced Echo, half a million people have told Alexa, 'I love you.'* GeekWire. https://www.geekwire.com/2015/one-year-after-amazonintroduced-echo-half-a-million-people-have-told-alexa-i-love-you/

Roach, J. (1997). *Austin Powers: International man of mystery*. [Motion picture]. New Line Cinema.

Rudnicky, A. I., Hauptmann, A. G., & Lee, K.-F. (1994). Survey of current speech technology. *Communications of the ACM*, *37*(3), 52–57. https://doi.org/10.1145/175247.175252

Saffari, E., Hosseini, S. R., Taheri, A., & Meghdari, A. (2021). "Does cinema form the future of robotics?": A survey on fictional robots in sci-fi movies. *SN Applied Sciences*, *3*(6), 655. https://doi.org/10.1007/s42452-021-04653-x

Salge, C., & Polani, D. (2017). Empowerment as replacement for the three laws of robotics. *Frontiers in Robotics and AI*, *4*, 25. https://doi.org/10.3389/frobt.2017.00025

Schwartz, M. F. (1968). Identification of speaker sex from isolated, voiceless fricatives. *The Journal of the Acoustical Society of America*, *43*(5), 1178–1179. https://doi.org/10.1121/1.1910954

Schweitzer, F., Belk, R., Jordan, W., & Ortner, M. (2019). Servant, friend or master? The relationships users build with voice-controlled smart devices. *Journal of Marketing Management*, *35*(7–8), 693–715. https://doi.org/10.1080/0267257X.2019.1596970

Skuk, V. G., & Schweinberger, S. R. (2014). Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. *Journal of Speech, Language, and Hearing Research*, *57*(1), 285–296. https://doi.org/10.1044/1092-4388(2013/12-0314)

Smorenburg, L., & Chen, A. (2020). The effect of female voice on verbal processing. *Speech Communication*, *122*, 11–18. https://doi.org/10.1016/j.specom.2020.04.004

Søraa, R. A. (2017). Mechanical genders: How do humans gender robots? *Gender, Technology and Development*, *21*(1–2), 99–115. https://doi.org/10.1080/09718524.2017.1385320

Sterling, G. (2020, February 14). More than 200 million smart speakers have been sold, why aren't they a marketing channel? Martech. https://martech.org/more-than-200-million-smart-speakers-have-been-sold-why-arent-they-a-marketing-channel/

Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology*, *18*(1), 86–100. https://doi.org/10.1177/0261927X99018001006

Strengers, Y., & Nicholls, L. (2018). Aesthetic pleasures and gendered tech-work in the 21st-century smart home. *Media International Australia*, *166*(1), 70–80. https://doi.org/10.1177/1329878X17737661

Stuart-Smith, J., Timmins, C., & Wrench, A. (2003). *Sex and gender differences in Glaswegian /s/.* 15th ICPhS Barcelona.

Styler, W. (2022). *Using Praat for Linguistic Research,* Version 1.9.

Tagliamonte, S. (2004). Someth[in]'s go[ing] on!: Variable ing at ground zero. *Language Variation in Europe: Papers from the Second International Conference on Language Variation in Europe* (pp. 1-11). ICLAVE2, Uppsala, Sweden.

Tay, B., Jung, Y., & Park, T. (2014). When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior*, *38*, 75–84. https://doi.org/10.1016/j.chb.2014.05.014

Topaz, M., Schaffer, A., Lai, K. H., Korach, Z. T., Einbinder, J., & Zhou, L. (2018). Medical malpractice trends: Errors in automated speech recognition. *Journal of Medical Systems*, *42*(8), 153. https://doi.org/10.1007/s10916-018-1011-9

Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society*, *1*(2), 179–195. https://doi.org/10.1017/S0047404500000488

Villarreal, D. (2018). The construction of social meaning: A matched-guise investigation of the California Vowel Shift. *Journal of English Linguistics*, *46*(1), 52–78. https://doi.org/10.1177/0075424217753520

Wang, L. (2020). The three harms of gendered technology. *Australasian Journal of Information Systems*, *24*, 1-9. https://doi.org/10.3127/ajis.v24i0.2799

West, M., Kraut, R., & Chew, H. (2019). 'I'd blush if I could:' Closing gender divides in digital skills through education. *UNESCO Report,* EQUALS, Paris.

Wolff, S. E., & Puts, D. A. (2010). Vocal masculinity is a robust dominance signal in men. *Behavioral Ecology and Sociobiology*, *64*(10), 1673–1683. https://doi.org/10.1007/s00265-010-0981-5

Woods, H. S. (2018). Asking more of Siri and Alexa: Feminine persona in service of surveillance capitalism. *Critical Studies in Media Communication*, *35*(4), 334–349. https://doi.org/10.1080/15295036.2018.1488082

Zellou, G., Cohn, M., & Kline, T. (2021). The influence of conversational role on phonetic alignment toward voice-AI and human interlocutors. *Language, Cognition and Neuroscience*, *36*(10), 1298–1312. https://doi.org/10.1080/23273798.2021.1931372

Zhao, J., & Patrick Rau, P.-L. (2020). Merging and synchronizing corporate and personal voice agents: Comparison of voice agents acting as a secretary and a housekeeper. *Computers in Human Behavior*, *108*, 106334. https://doi.org/10.1016/j.chb.2020.106334

Zimman, L. (2017). Gender as stylistic bricolage: Transmasculine voices and the relationship between fundamental frequency and /s/. *Language in Society*, *46*(3), 339–370. https://doi.org/10.1017/S0047404517000070

Appendices

Appendix A: Stimuli Sentences for Chapter 3

Vowel of Interest	Stimuli
/i/	Even he seeks the seeds.
/ɪ/	Tish ignorantly considered calling dibs.
/e/	We ate tasty and savory dates.
/ɛ/	Ted will be dead until the setter goes to bed.
/æ/	Bad athletic cats tapped bats.
/u/	Who chooses oodles of boots.
/ʊ/	The cook wrote in the good book.
/o/	The boat goaded goats to the hotel.
/ɔ/	A bawd fought Jim before he was caught.
/ɑ/	Father's cot was in stock for a bot.

Appendix B: Rainbow Passage

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow. Throughout the centuries people have explained the rainbow in various ways. Some have accepted it as a miracle without physical explanation. To the Hebrews it was a token that there would be no more universal floods. The Greeks used to imagine that it was a sign from the gods to foretell war or heavy rain. The Norsemen considered the rainbow as a bridge over which the gods passed from earth to their home in the sky. Others have tried to explain the phenomenon physically. Aristotle thought that the rainbow was caused by reflection of the sun's rays by the rain. Since then physicists have found that it is not reflection, but refraction by the raindrops which causes the rainbows. Many complicated ideas about the rainbow have been formed. The difference in the rainbow depends considerably upon the size of the drops, and the width of the colored band increases as the size of the drops increases. The actual primary rainbow observed is said to be the effect of super-imposition of a number of bows. If the red of the second bow falls upon the green of the first, the result is to give a bow with an abnormally wide yellow band, since red and green light when mixed form yellow. This is a very common type of bow, one showing mainly red and yellow, with little or no green or blue (Fairbanks, 1960; as cited by Zimman, 2017).

Appendix C: The Boy Who Cried Wolf

There was once a poor shepherd boy who used to watch his flocks in the fields next to a dark forest near the foot of a mountain. One hot afternoon, he thought up a good plan to get some company for himself and also have a little fun. Raising his fist in the air, he ran down to the village shouting "Wolf, Wolf." As soon as they heard him, the villagers all rushed from their homes, full of concern for his safety, and two of his cousins even stayed with him for a short while. This gave the boy so much pleasure that a few days later he tried exactly the same trick again, and once more he was successful. However, not long after, a wolf that had just escaped from the zoo was looking for a change from its usual diet of chicken and duck. So, overcoming its fear of being shot, it actually did come out from the forest and began to threaten the sheep. Racing down to the village, the boy of course cried out even louder than before. Unfortunately, as all the villagers were convinced that he was trying to fool them a third time, they told him, "Go away and don't bother us again." And so the wolf had a feast (Deterding, 2006).

Appendix D: Stimuli for Study 1.2

(Allen 2022)

Prompt Type	Prompts
Fact-based	What is the weather today?
	How do I get to my home?
	Who is the current president?
	Who is the highest paid actor?
	Who is the highest paid actress?
Emotion-based	I'm sad.
	You're pretty.
	I'm lonely.
	Will you be my friend?
	Will you always be there to help me?
Opinion-based	Do you think men and women are equal?
	What should I have for dinner?
	Do you think I should become an influencer?
	What do you think is a good gift?
	What should I name my cat?