2024 33rd IEEE International Conference on
Robot and Human Interactive Communication (RO-MAN)
August 26-30, 2024. Pasadena, California, USA.

Gendered and Machine-like Features in Voices Affect Social Judgments

Kelsey L. Neuenswander[1], Gregory A. Bryant[1], and Steven J. Stroessner[1]

Abstract— **People regularly interact with synthetic speech systems that vary in vocal characteristics. Three studies explored whether anthropomorphic (i.e., humanlike) and gendered (i.e., masculine/feminine) features of voices affected fundamental judgments of warmth, competence, and discomfort. Study 1 used vocal recordings with semantic content removed. Natural voices were judged as warmer, more competent, and produced less discomfort than machine-like voices. Feminine voices were judged as warmer and elicited less discomfort than masculine voices. Study 2 used unfiltered versions of the same stimuli and found similar negative impacts of machine-like features of social judgments, but the effects of voice gender disappeared. Study 3 examined whether gendered effects returned when gender-ambiguous semantic content was labeled as stereotypically feminine (braiding hair) or stereotypically masculine (tying rope). Stereotype consistency between task label and voice gender mattered for discomfort ratings of feminine voices. In sum, semantic content and the consistency between content and stereotypical expectations affected judgments of gendered voices. Additional considerations regarding gendered voices and speech content should be explored in future research.**

I. Introduction

People engage daily with technology through synthetic speech interfaces. Systems' communicative capabilities affect user experience, requiring research into how users perceive different interactive features. Synthetic voices range from artificial and robotic-sounding (e.g., Samsung Bixby) to natural and human-sounding (e.g., ChatGPT Voice). What are the impacts of different voice interfaces, and to what extent do social cognitive processes drive our judgments about these interfaces?

We present findings from three experiments investigating the impact of gendered (masculine vs feminine) and anthropomorphic (natural vs machine-like) vocal traits on social perception. We assessed the influence of these manipulations on fundamental dimensions of social judgment using the Robotic Social Attributes Scale (RoSAS), a measure for studying human interactions with technologies [1]. These studies add to our understanding of the social perception of various technologies.

II. Background

A. Anthropomorphism in Human-Technology Interactions

People regularly anthropomorphize the nonsocial world by attributing human traits and motivations to animals, objects, and shapes [2-4]. Technologies are also anthropomorphized. Robots often possess physical features that resemble humans (e.g., arms, legs, torsos, and detailed faces), leading people to think about robots as humans [5]. Importantly, anthropomorphism has been shown to affect various phenomena, including consumer behavior [6-10], moral judgment [11-14], and support for policies [15-17].

One timely yet understudied question concerns the social perception of auditory stimuli prevalent in human-computer interaction (HCI), human-robotic interaction (HRI), and interactions with vocal assistants (VAs). Vocal interfaces often involve artificial intelligence programs that simulate human conversation to conduct searches and complete simple commands. People regularly communicate with computers [18-19] and robots [20-21] through voice. VAs like Amazon Alexa, Google Assistant, and Apple Siri utilize voice for communication between digital devices and users, and the global frequency of VAs is projected to reach 8.4 billion by 2024 [22]. Given the escalating presence of vocal interfaces across various technologies in daily life, it is crucial to understand how voices are perceived socially.

Research also indicates that objects and technologies can become associated with human social categories like gender, ethnicity, and age [23-24]. When nonhuman entities implicate social categories based on their physical properties or actions, subsequent interpretations, evaluations, and inferences can be affected [25-26]. Our research adds to the developing literature regarding how vocal communication affects the social perception of nonhuman agents, focusing specifically on anthropomorphism and gender.

B. The Social Perception of Human Voices

Understanding how vocal features affect assessments of human speakers provides a foundation for comprehending how vocalized technologies might be perceived socially. Voice pitch, for instance, significantly influences the social perception of humans, accounting for over 40% of the variance in gender-related judgments, including stereotypes [27-28]. Speakers with higher-pitched voices are judged more approachable but less competent [29], while lower vocal pitch is associated with higher judgments of dominance and strength [30]. Additionally, acoustic cues such as higher pitch, greater loudness, and loudness variability reliably signal social rank, with speakers possessing these traits accurately perceived as higher in social standing [31].

Human vocal features associated with gender categories elicit judgments consistent with gender stereotypes. One study exploring judgments of job applicants found that perceived competence was solely influenced by vocal masculinity. More

[1] Department of Communication, University of California Los Angeles 2225 Rolfe Hall, Los Angeles, CA 90095 (corresponding author K.L.N. email: klneuenswander@ucla.edu). Materials, data, and code are available from the corresponding author upon reasonable request.

masculine-voiced applicants were rated more competent than feminine-voiced applicants [28]. Another study found that higher vocal pitch led to more robust attribution of feminine traits and greater likability but lower perceived competence [32]. Additionally, high-pitched male speakers were perceived as less masculine than low-pitched speakers. Overall, pitch plays a crucial role in the social perception of human speakers, leading to stereotypical judgments of gender-typical voices.

C. The Social Perception of Synthetic Voices

Several research areas have examined the perception of voices from supposed nonhuman communicators, including their role in anthropomorphism and the impact of gender associations on voice perception. We provide a brief overview of findings in these domains involving synthetic voices.

1) Synthetic vocal communication and anthropomorphism

Incorporating voices into existing technologies enhances anthropomorphism. For example, adding a voice to a computer-generated script increased the likelihood of perceiving the script's creator as human [33]. In another study, participants who watched videos showing a robot approaching a person attributed more human characteristics to robots imbued with voices [34].

For VAs, natural (i.e., humanlike) voices heighten anthropomorphism. VAs with natural voices increase relational investment and trust among users toward the organization represented by the VA [35]. Similarly, participants interacting with an autonomous vehicle with anthropomorphic features such as a name, gender, and voice judge it as competent and trustworthy [36]. Adding humor and voice to VAs also positively impacts users' perceptions of anthropomorphism, increasing trust and intention to use the VA [37]. Natural voices tend to be viewed as more approachable [38] and likable [39] than machine-like voices.

2) Synthetic vocal communication and gender

While developers are advised to avoid gendering VAs [40], evidence indicates that listeners often assign gender to synthesized speech [41-44]. Listeners appear to use perceptual features (e.g., pitch) and contextual cues (e.g., names) to gender categorize voiced technologies [45-47].

Importantly, the perceived gender of synthesized speech influences subsequent judgments. Decades of research demonstrate strong gender stereotypes that characterize men as agentic and competitive and women as warm and communal [48]. Similar effects are evident in HCI and HRI. Female-voiced robots are perceived as warmer and generally preferred over male robots [49-50]. Consequently, this preference has resulted in most VAs being designed with a default female gender [51]. However, female-voiced technologies often receive negative judgments when in dominant roles [47] or evaluated for competence [49-50, 52]. Conversely, lower-pitched, more masculine-voiced technologies are taken more seriously [47] and perceived as more knowledgeable and skilled than their female-voiced counterparts [53]. In sum, vocal characteristics related to gender (e.g., pitch) can affect judgments of various crucial social attributes.

D. Fundamental Dimensions of the Social Perception of Technology

Voiced technologies are judged socially based on anthropomorphic and gendered vocal features. However, greater investigation into how voice affects fundamental aspects of judgment in HRI and HCI is needed. Recently, a psychometrically validated scale of robot social perception, the Robotic Social Attributes Scale (RoSAS), has been developed [1]. This scale reflects people's tendency to focus on three fundamental aspects of robots: their warmth, competence, and the degree to which they produce feelings of discomfort. These judgments strongly predict evaluations of and willingness to engage with robots [54-55]. Although initially designed to assess responses to robots, the RoSAS has been effectively used to study human interactions with various types of technology [50, 56-57].

Regarding visual perception, humanlike robots are typically perceived as warmer, more competent, and less discomfort-inducing than machine-like robots. Feminine-appearing robots are generally considered warmer and evoke less discomfort than masculine robots [1, 54-55]. However, studies have produced inconsistent gender effects regarding competence. Some show equal competence between male and female robots [54], whereas other studies suggest higher competence ratings for perceived male robots [25, 58] or female robots [1, 55]. These inconsistent findings parallel variable results seen in judgments of human competence, ranging from presumed male superiority [29] to gender equality [58] to even female advantage [59]. The inconsistent findings may reflect differences in the interpretation of "competence," which can vary substantially across contexts [48] and individuals [61]. It is important to note that competence, as measured by the RoSAS emphasizing reliability, differs from measures of human competence, which often focus on agency [62]. One study found that when using the RoSAS, female-appearing robots were judged equally and sometimes higher than male-appearing robots on competence [54]. However, female-appearing robots were considered less competent using a scale reflecting human agency. Therefore, the lack of differences in RoSAS competence judgments does not preclude the possibility of gender stereotyping involving agency, independence, and self-confidence.

While existing studies have primarily explored social judgments of robots based on their visual appearance, more research is needed concerning the perception of voices along fundamental RoSAS dimensions. Our work aims to contribute to the growing body of knowledge in HRI and HCI research by manipulating gendered and anthropomorphic vocal features and assessing their impact on fundamental social judgments captured by the RoSAS.

III. Overview of Unique Contributions of Current Studies

The current studies add to the understanding of the social perception of voices in several respects. First, they assess how vocal features and content variations affect social judgments using a psychometrically validated scale, the RoSAS, frequently used in studying human interactions with technology. This ensures a focus on fundamental aspects of social perception and facilitates comparisons across studies.

Second, we use vocal stimuli systematically varied to reflect different levels of anthropomorphism and gender-typicality. Instead of using different speakers in different conditions (e.g., gender might be varied by using separate male and female speakers), we manipulate recordings of single speakers to eliminate numerous confounds. Third, several variables (i.e., gender-typicality, anthropomorphism, task instructions) are examined across the studies to pose novel questions within a single empirical framework. For instance, does varying the presence of semantic content impact how a voice is perceived in gendered or anthropomorphic terms? Do vocal features interact with speech content to affect how a voice is judged? Many of these questions have been examined in isolation, but they are posed here within a unified set of studies using consistent materials and rigorous controls.

IV. Study 1

Study 1 explored how variations in vocal features affect social judgments of machine-like voices. Participants listened to four voices devoid of semantic content, systematically varying in gender-typicality (masculine or feminine) and anthropomorphism (natural or machine-like). Subsequently, participants rated each voice on the RoSAS. All audio samples were derived from the same base voice recording to control for prosodic differences across stimuli. Furthermore, factors like background noise and context were controlled. We expected listeners' judgments to reflect category-level beliefs based on vocal features (i.e., gender stereotypes and judgments of humans vs. nonhumans), akin to those observed for robots differing in gendered and anthropomorphic appearance [1, 54]. Specifically, we predicted that i) feminine and natural voices would be perceived as warmer than masculine and machine-like voices, ii) natural voices would be perceived as more competent than machine-like voices, and iii) masculine and machine-like voices would elicit more discomfort than feminine and natural voices.

A. Method

1) Participants

We recruited 48 undergraduate students (Gender: 42 women, 6 men; Age: M = 19.79, range = 18-27) who participated in the study in exchange for course credit, exceeding sample size recommendations to achieve 80% power [63].

2) Stimuli

To create the vocal stimuli, an audio recording was made of an adult male speaker reading instructions with an average fundamental frequency (f_0) of approximately 150 Hz (SD = 30 Hz). A male voice was chosen because the transformation of male-to-female voices using PSOLA (Pitch Synchronous Overlap Add) results in higher quality tokens with fewer artifacts than female-to-male transformations. The reading was recorded digitally (16-bit, 44.1 kHz) using a Logitech USB desktop microphone in a quiet room with no background noise.

To manipulate gender typicality, we utilized the VTChange script (C. Darwin) in Praat (version 6.0.43) [64]. f_0 and formant frequencies were independently manipulated using PSOLA resynthesis. For the male-sounding version, f_0 was shifted down 20% (~30 Hz), and apparent vocal tract length (VTL) was increased by 10%, resulting in versions with f_0 values around 120 Hz (SD = 20 Hz). Female-sounding versions were created by shifting f_0 up 50% (~65 Hz) and lowering apparent VTL by 10%, resulting in f_0 values of approximately 215 Hz. These adjustments align with average f_0 and formant values for men and women, respectively [65].

The altered audio files were further transformed to emulate machine-like voices by creating two additional versions of each file. The first versions were pitch-shifted using the Adobe Audition 3.0 stretch function (Ratio: 80, Stretching Mode: Pitch Shift). The second versions were generated from stereo conversions of the audio that were imported into Audacity software [66] and transformed using the vocoder effect (Distance: 20, Vocoder bands: 168; Amplitude of Radar Needles (%): 84, Frequency of Radar Needles (Hz): 77). The three audio files (original, pitch-shifted, and vocoder) were imported into a multitrack session in Adobe Audition 3.0 for each sample and then mixed with the vocoder track raised 6 dB relative to the other two tracks into a single mono .wav file.

For Study 1, all semantic content was eliminated from the recordings by low-pass filtering the uncompressed audio samples at 0.5 kHz using a Butterworth filter (24 dB/octave roll-off, 100 Hz transition bandwidth) in Adobe Audition 3.0 with the scientific filter function. This level of filtering reduces lexical identification to near zero while retaining f_0 and most F1 information, as well as amplitude and speech rhythm dynamics [67].

These manipulations produced four audio samples devoid of discernable verbal content but differing in gender-typicality and anthropomorphism: two resembling masculine and feminine human voices and two resembling masculine and feminine mechanical voices that might be associated with VAs or robots.

3) Procedure

Participants were told the study aimed "to examine how vocal qualities affect trait judgments of different voices." They sat individually before a computer, which played all four audio samples counterbalanced. After each sample, participants rated the degree to which they associated the voice with each of the 18 RoSAS attributes, covering warmth (happy, feeling, sociable, organic, compassionate, emotional), competence (capable, responsive, interactive, reliable, competent, knowledgeable), and discomfort (scary, strange, awkward, dangerous, awful, aggressive) in random order.

B. Results

Responses to each RoSAS subscale were analyzed using repeated measures ANOVAs, with Gender-Typicality (masculine, feminine) and Anthropomorphism (natural, machine-like) as factors. The means for each voice on each RoSAS dimension are displayed in Table 1.

- **Warmth:** Natural voices (M = 3.22, SD = 1.18) were judged as warmer than machine-like voices (M = 2.29, SD = 0.98), $F(1, 188) = 35.61$, $p < .001$, partial $\eta 2 = .159$, and feminine voices (M = 2.92, SD = 1.18) were judged as warmer than masculine voices (M = 2.59, SD = 1.17), $F(1, 188) = 4.54$, $p = .034$, partial $\eta 2 = .024$.

- **Competence:** Natural voices (M = 4.19, SD = 1.35) were judged as more competent than machine-like voices (M = 2.99, SD = 1.14), $F(1, 188) = 44.45$, $p < .001$, partial $\eta 2 = .19$. There was no significant main effect of gender on perceived competence, $F(1, 188) = 0.03$, $p = .870$, partial $\eta 2 = .00$.
- **Discomfort:** Machine-like voices (M = 4.21, SD = 1.33) were associated with higher levels of discomfort than natural voices (M = 2.51, SD = 1.07), $F(1, 188) = 98.24$, $p < .001$, partial $\eta 2 = .34$, and masculine voices (M = 3.54, SD = 1.64) produced more discomfort than feminine voices (M = 3.19, SD = 1.27), $F(1, 188) = 4.27$, $p = .040$, partial $\eta 2 = .02$. Gender-Typicality and Anthropomorphism variables interacted, $F(1, 188) = 4.62$, p = .033, partial $\eta 2 = .024$. Simple effect analyses revealed that within natural stimuli, perceived discomfort did not differ between masculine (M = 2.51, SD = 1.20) and feminine (M = 2.52, SD = 0.94) voices, $F(1, 188) = .003$, $p = .954$. However, for the machine-like voices, the masculine version evoked higher levels of discomfort (M = 4.57, SD = 1.36) than the feminine one (M = 3.85, SD = 1.21), $F(1, 188) = 8.89$, $p = .003$.

TABLE I. RoSAS Judgments of Vocal Stimuli (M (SD))

Voice Type	Study	Trait Dimension		
		warmth	*competence*	*discomfort*
Natural Masculine	1	3.14 (1.14)	4.26 (1.22)	2.51 (1.20)
Natural Feminine	1	3.31 (1.23)	4.13 (1.48)	2.52 (0.94)
Machine-like Masculine	1	2.05 (0.91)	2.89 (1.16)	4.57 (1.36)
Machine-like Feminine	1	2.54 (1.00)	3.08 (1.11)	3.85 (1.21)
Natural Masculine	2	3.17 (1.62)	5.03 (1.02)	2.37 (1.63)
Natural Feminine	2	3.20 (1.54)	4.92 (1.13)	2.46 (1.43)
Machine-like Masculine	2	2.07 (1.44)	3.81 (1.30)	4.11 (1.14)
Machine-like Feminine	2	2.33 (1.44)	3.95 (1.36)	4.04 (1.37)

C. Discussion

Our manipulation of a voice to influence gender-typicality (masculine or feminine) and anthropomorphism (natural or machine-like) generally resulted in the expected pattern of trait judgments. Natural voices were perceived as warmer and more competent, generating less discomfort than machine-like voices. Additionally, feminine voices were judged as warmer than masculine voices. While there was no difference in discomfort elicited by masculine and feminine natural voices, masculine machine-like voices produced more discomfort than feminine machine-like voices.

These findings underscore the significance of nonlinguistic vocal features in social perception, as observed in assessments of spoken vocal stimuli devoid of verbal information. Many studies use synthetic voices with no semantic content with evidence of perceptual effects of nonlinguistic features [68]. However, real-world voices typically contain words with meaning. Hence, different social judgment effects might stem from responses to vocal properties, content, or a combination [69-70]. This possibility was addressed in Study 2 using a broader sample of participants.

V. Study 2

Study 2 reintroduced semantic content to the voice recordings from Study 1 to determine whether the effects of nonlinguistic cues might interact with speech content to affect social judgments. This change necessitated a design ensuring listeners did not encounter repeated verbal content. Two potential outcomes regarding the introduction of semantic content were considered: adding verbal information might contextualize vocal feature judgment, possibly strengthening the effects of these features on judgments. Conversely, including semantic content might shift listeners' attention towards the content rather than vocal qualities. In this case, we would anticipate reducing the effects observed in Study 1 by adding semantic content.

In Study 2, participants' expectations (human or robot source) were established before they listened to one of four voices delivering task instructions. These voices varied in gender-typicality (masculine or feminine) and anthropomorphism (natural or machine-like). Participants then rated the speaker on RoSAS dimensions (warmth, competence, and discomfort). Measures of enjoyment, engagement, and task performance on a multiple-choice attention check exam were collected for exploratory purposes.

A. Method

1) Participants

A total of 363 online workers (https://www.mturk.com/) completed the study in exchange for $1.00. After removing 23 participants who scored no better than chance-level (25%) on the multiple-choice attention check, the final sample comprised 340 participants (Gender: 188 men, 151 women, 1 unidentified; Age: M = 36.34, range = 19-74). This exceeded our recommended sample size for 80% power [63].

1) Stimuli and Procedure

Study 2 followed a similar procedure to Study 1, with some exceptions. The same four audio clips were used, but the semantic content was unfiltered. This allowed participants to hear and respond to table-setting instructions delivered by voices varying in gender-typicality and anthropomorphism. Instructions describing arranging plates and silverware were chosen because they emulated those that a VA might provide (full instructions available from the corresponding author). Given identical instructions across voices, Study 2 used a between-subjects design, with each participant hearing only one voice. To manipulate expectations, half of the participants were told before listening to the instructions that the voice had initially come from a human. In contrast, the other half were told that it was machine-generated.

The same dependent measures as in Study 1 were collected, along with additional exploratory measures to assess the effects of expectation, gender-typicality, and anthropomorphism on task performance, engagement, and enjoyment. These findings are not reported in the current manuscript. After listening to verbal instructions, participants completed a surprise multiple-choice test with four options per question (e.g., "At what angle is the bread plate placed in relation to the presentation plate?") and rated their perceived

engagement with the task (1 = not at all, 7 = very much). Finally, participants responded to three questions about their enjoyment of the task (1 = not at all, 7 = very much). Following data collection, participants were debriefed.

B. Results

Responses to each RoSAS subscale were initially analyzed using a 2 (Expectation: human, machine) x 2 (Gender-Typicality: masculine, feminine) x 2 (Anthropomorphism: natural, machine-like) ANOVA. The manipulation of expectations did not significantly influence the results. Therefore, only the effects of anthropomorphism and gender-typicality are reported (Table 1). The lack of influence from expectations bolsters confidence that vocal features and semantic content, rather than expectations regarding the communication source, underlie the obtained effects.

- **Warmth:** Natural voices (M = 3.19, SD = 1.57) were considered warmer than machine-like voices (M = 2.21, SD = 1.44), $F(1, 336) = 35.85$, $p < .001$, partial $\eta 2 = .096$. The difference in perceived warmth between masculine (M = 2.63, SD = 1.62) and feminine (M = 2.76, SD = 1.55) voices was not significant, $F(1, 336) = 0.59$, $p = .442$, partial $\eta 2 = .002$.
- **Competence:** Natural voices (M = 4.98, SD = 1.08) were judged more competent than machine-like voices (M = 3.88, SD = 1.33), $F(1, 336) = 69.52$, $p < .001$, partial $\eta 2 = .171$. There was no difference in perceived competence between masculine (M = 4.44, SD = 1.31) and feminine (M = 4.43, SD = 1.34) voices, $F(1, 336) = 0.00$, $p = .953$, partial $\eta 2 = .000$.
- **Discomfort:** Machine-like voices (M = 4.07, SD = 1.26) produced higher levels of discomfort than natural voices (M = 2.42, SD = 1.53), $F(1, 336) = 117.62$, $p < .001$, partial $\eta 2 = .259$. Discomfort in response to masculine (M = 3.22, SD = 1.66) and feminine (M = 3.26, SD = 1.61) voices was similar, $F(1, 336) = 0.07$, $p = .790$, partial $\eta 2 = .000$.

C. Discussion

Study 2 presented listeners with unfiltered stimuli used in Study 1, including verbal content. This led to a different pattern of RoSAS judgments. While anthropomorphism effects replicated—natural voices were perceived as warmer, more competent, and led to lower discomfort ratings than machine-like voices—gender effects wholly disappeared. Thus, when speech content was included, listeners remained sensitive to some perceptual features of the voice (anthropomorphism) but not others (gender-typicality).

One limitation of this study is the potentially gendered nature of the task instructions (setting a table), which might be more strongly associated with women than men. Study 3 addressed this issue by manipulating the nature of the instructions (i.e., stereotypically masculine vs. feminine instructions) without altering vocal features. This tested whether gendered vocal features interacted with the stereotypicality of the task being described.

VI. Study 3

Study 3 investigated whether the stereotypic consistency (e.g., a feminine vs masculine voice discussing a gendered task) of semantic content in task instructions would influence RoSAS judgments independently or in conjunction with vocal characteristics. Previous research suggests that users prefer gendered robots that fulfill stereotypical roles, such as a female caretaker or a male protector [25, 71-72], and listeners process voices differently when they describe gender-incongruent statements (e.g., male voice uttering 'I like to wear lipstick') versus gender-congruent statements [70]. Participants in Study 3 were given instructions for a task from voices varying in gender-typicality and anthropomorphism. The gendered nature of the task was manipulated by altering the described activity (braiding hair vs. tying rope) [73].

A. Method

1) Participants

We recruited 550 online workers (www.prolific.co) and compensated them $1.50 for their participation. After removing 46 participants who scored less than 25% on the multiple-choice attention check, we retained a final sample of 527 participants (Gender: 297 men, 222 women, 8 unidentified; Age: M = 38.20, range = 18-84). This exceeded our sample recommendation [63].

1) Stimuli and Procedure

As in Study 2, participants heard task instructions delivered by a voice varying in gender-typicality and anthropomorphism. However, we used ambiguous instructions that could be labeled as describing a stereotypically masculine or feminine activity. Participants in the masculine condition were told, "You will be asked to listen to a set of task instructions on performing a key outdoor skill: tying a knot." Those in the feminine condition were told, "You will be asked to listen to a set of task instructions on performing a key domestic skill: braiding hair" (full instructions available from the corresponding author). This study utilized a between-subjects design with three factors: Topic (masculine, feminine), Gender-Typicality (masculine, feminine), and Anthropomorphism (natural, machine-like).

After listening to instructions from one of the four voices, participants completed the same dependent measures as in Study 2 and an updated multiple-choice test to assess task performance based on the new instructions. They then provided demographic information and were debriefed.

B. Results

Responses to each RoSAS subscale were analyzed using a 2 (Topic: masculine, feminine) x 2 (Gender-Typicality: masculine, feminine) x 2 (Anthropomorphism: natural, machine-like) ANOVA.

- **Warmth:** Natural voices (M = 2.91, SD = 1.14) were judged warmer than machine-like voices (M = 1.73, SD = 0.89), $F(1, 519) = 175.90$, $p < .001$, partial $\eta 2 = .251$. However, there were no main effects of Topic, $F(1, 519) = 0.35$, $p = .553$, partial $\eta 2 = .000$, or gender-typicality, $F(1, 519) = 0.18$, $p = .669$, partial $\eta 2 = .000$, on warmth judgments.

- **Competence:** Natural voices (M = 4.52, SD = 1.16) were judged more competent than machine-like voices (M = 3.42, SD = 1.26), $F(1, 519) = 109.31$, $p <$ .001, partial η2 = .173. There were no main effects of Topic, $F(1, 519) = 0.41$, p = .524, partial η2 = .001, or Gender-Typicality, $F(1, 519) = 0.34$, p = .558, partial η2 = .001, on competence judgments.
- **Discomfort:** Machine-like voices (M = 3.96, SD = 1.25) were associated with higher discomfort levels compared to natural voices (M = 1.80, SD = 0.90), $F(1, 519) = 523.16$, $p <$ .001, partial η2 = .500. There were no main effects of Topic, $F(1, 519) = 0.99$, p = .320, partial η2 = .002, or Gender-Typicality, $F(1, 519) = 0.06$, p = .815, partial η2 = .000, on discomfort judgments. However, a significant Gender-Typicality x Topic interaction emerged, $F(1, 519) = 6.54$, p = .011, partial η2 = .012. Simple effects analysis showed that feminine voices paired with a masculine topic label (M = 3.09, SD = 1.54) elicited higher levels of discomfort than with a feminine topic label (M = 2.71, SD = 1.36), $F(1, 519) = 6.64$, p = .010. For masculine voices, discomfort levels did not significantly differ between masculine (M = 2.74, SD = 1.58) and feminine topic labels (M = 2.99, SD = 1.61), $F(1, 519) = 1.06$, p = .305. Additionally, a significant Gender-Typicality x Anthropomorphism interaction emerged, $F(1, 519) = 5.37$, p = .021, partial η2 = .010. When voices were natural, feminine voices (M = 1.93, SD = 0.95) led to marginally higher levels of discomfort than masculine voices (M = 1.69, SD = 0.85), $F(1, 519) = 3.01$, p = .084, although this did not reach traditional levels of significance. When voices were machine-like, the difference in discomfort ratings between masculine voices (M = 4.05, SD = 1.26) and feminine voices (M = 3.86, SD = 1.23) was not significant, $F(1, 519) = 2.34$, p = .126 (Figure 1).

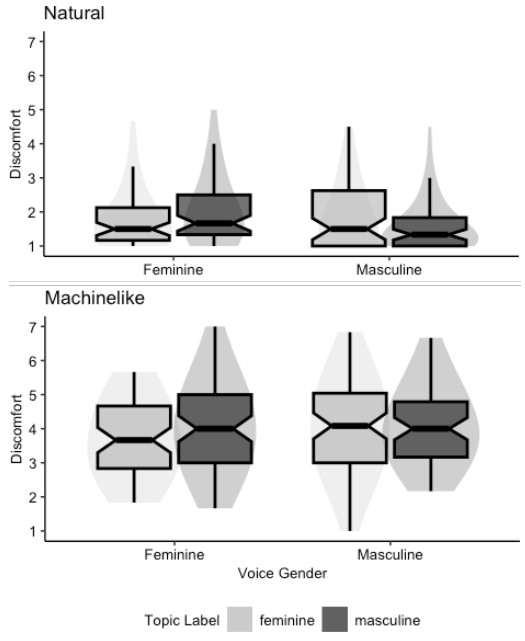

Figure 1. Discomfort Ratings by Voice Gender-Typicality and Topic Label for Natural Voices (top) and Machine-like Voices (bottom).

C. Discussion

Manipulations of task labels (masculine, feminine) and vocal features (gender-typicality, anthropomorphism) affected trait judgments of voices. Natural voices were perceived as warmer and more competent than machine-like voices, replicating findings from Study 2. Once again, no gendered effects emerged on warmth or competence. However, a notable result was observed: feminine voices discussing stereotypically masculine topics (tying rope) generated more discomfort than those discussing stereotypically feminine topics (braiding hair). This effect was not observed for masculine voices, suggesting that congruence between voice gender and task stereotypicality may impact discomfort judgments more prominently when the voice is feminine.

VII. General Discussion

The gender-typicality and anthropomorphism of voices influenced trait judgments of warmth, competence, and discomfort. When voices were stripped of semantic content (Study 1), perceptual features of both gender-typicality and anthropomorphism influenced judgments. Natural voices were rated warmer, more competent, and less discomfort-inducing than machine-like voices. Feminine voices were rated as warmer and less discomfort-inducing than masculine voices. When semantic content was included (Study 2), the effects of anthropomorphism persisted, but the impact of gender-typicality disappeared. Some gendered effects resurfaced when the semantic content was explicitly gendered (Study 3). Specifically, feminine voices describing a stereotypically masculine task (tying rope) elicited more discomfort than those describing a stereotypically feminine task (braiding hair). This was not the case for masculine voices describing stereotypically congruent or incongruent tasks. Additionally, natural feminine voices elicited more discomfort than natural masculine voices. This gender difference was not observed for machine-like voices.

These findings suggest that while the influence of anthropomorphism on trait judgments remains consistent, gender cues still play a subtle role in shaping perceptions of voices. Specifically, when voices are feminine, the stereotype consistency between voice gender and task matters for feelings of discomfort but not warmth or competence. This poses a potential problem given the prevalent "female by default" paradigm for vocal assistants. Indeed, most users depend on VAs for quick information and directions [74], and there are prevalent stereotypes that men possess superior navigation skills compared to women [75]. This mismatch between voice gender (female by default) and stereotypical semantic content (i.e., directions) may lead to user discomfort. However, this doesn't imply that VAs should conform strictly to stereotypical roles. Instead, the findings suggest an alternate strategy—leveraging variations in anthropomorphism to mitigate gender-related disparities in discomfort.

VIII. Limitations and Future Directions

These studies suggest several directions for future research. First, further exploration of the interaction between gendered and anthropomorphic vocal features could assist in optimizing the benefits of machine-like voices, such as reducing gender-related disparities, while minimizing potential drawbacks (e.g., negative evaluations overall). Second, investigating the

impact of congruence between content and voice remains crucial. The finding that a mismatch between a female voice and stereotypically male content can lead to discomfort has significant implications for design considerations. There may be other topics where a mismatch between a male voice and stereotypically female content could similarly lead to discomfort (e.g., a hypermasculine voice providing care for elderly adults). Third, while using a single manipulated human voice maximized control, future studies could explore how these manipulations translate to authentic computer-generated voices. Fourth, there is evidence of evaluative differences of voices within gender categories. For instance, high-pitched female robots are perceived as more attractive and rated as more enjoyable than low-pitched female robots [76]. These evaluations also vary by context [77]. Future research might investigate how pitch variations within gender categories interact with anthropomorphism to affect judgments.

In sum, voices are judged differently depending on their gendered and anthropomorphic qualities and the semantic content they convey. These findings offer novel insights into the implications of vocal variations in HRI and HCI.

REFERENCES

[1] Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017). The robotic social attributes scale (RoSAS): Development and validation. *ACM/IEEE International Conference on Human-Robot Interaction*, *Part F127194*, 254–262.

[2] Guthrie, S. E. (1993). *Faces in the clouds: A new theory of religion*. Oxford University Press.

[3] Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.

[4] Stroessner, S. J., & Koya, P. D. (2024). Thinking socially about the nonsocial world. In D. Carlston, K. Hugenberg, & K. L. Johnson (Eds.), *Oxford Handbook of Social Cognition, Volume 2* (pp. 616-643). New York, NY: Oxford University Press.

[5] Mathur, M. B., & Reichling, D. B. (2009). An uncanny game of trust: Social trustworthiness of robots inferred from subtle anthropomorphic facial cues. *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, 313–314.

[6] Aggarwal, P., & McGill, A. L. (2007). Is that car smiling at me? Schema congruity as a basis for evaluating anthropomorphized products. *Journal of Consumer Research, 34*(4), 468–479.

[7] Aggarwal, P., & McGill, A. L. (2012). When brands seem human, do humans act like brands? Automatic behavioral priming effects of brand anthropomorphism. *Journal of Consumer Research*, *39*(2), 307–323.

[8] Hur, J. D., Koo, M., & Hofmann, W. (2015). When temptations come alive: How anthropomorphism undermines self-control. *Journal of Consumer Research, ucv017*.

[9] Puzakova, M., Rocereto, J. F., & Kwak, H. (2013). Ads are watching me. *International Journal of Advertising, 32*(4), 513–538.

[10] Riva, P., Sacchi, S., & Brambilla, M. (2015). Humanizing machines: Anthropomorphization of slot machines increases gambling. *Journal of Experimental Psychology: Applied*, *21*(4), 313–325.

[11] Brown, C. M., & McLean, J. L. (2015). Anthropomorphizing dogs: Projecting one's own personality and consequences for supporting animal rights. *Anthrozoös*, *28*(1), 73–86.

[12] Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619–619.

[13] Wang, F., & Basso, F. (2019). "Animals are friends, not food": Anthropomorphism leads to less favorable attitudes toward meat consumption by inducing feelings of anticipatory guilt. *Appetite*, *138*, 153–173.

[14] Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232.

[15] Ahn, H.-K., Kim, H. J., & Aggarwal, P. (2014). Helping fellow beings. *Psychological Science*, *25*(1), 224–229.

[16] Martín-Forés, I., Martín-López, B., & Montes, C. (2013). Anthropomorphic factors influencing Spanish conservation policies of vertebrates. *International Journal of Biodiversity*, *142670*, 1–9.

[17] Sacchi, S., Riva, P., & Brambilla, M. (2013). When mother earth rises up. *Social Psychology*, *44*(4), 271–277.

[18] Garg, R., Cui, H., Seligson, S., Zhang, B., Porcheron, M., Clark, L., Cowan, B., & Beneteau, E. (2022). The last decade of HCI research on children and voice-based conversational agents. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19.

[19] Sutton, S. J., Foulkes, P., Kirk, D., & Lawson, S. (2019). Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

[20] Schreibelmayr, S., & Mara, M. (2022). Robot voices in daily life: Vocal human-likeness and application context as determinants of user acceptance. *Frontiers in Psychology*, *13*(787499), 1–17.

[21] Wuth, J., Correa, P., Núñez, T., Saavedra, M., & Yoma, N. B. (2021). The role of speech technology in user perception and context acquisition in HRI. *International Journal of Social Robotics*, *13*, 949–968.

[22] Juniper Research (2023). Voice assistants: Market forecasts, monetisation strategies, and competitive landscape 2021-2026. https://www.juniperresearch.com/researchstore/innovation-disruption/voice-assistants-market-research-report

[23] Kuchenbrandt, D., Eyssel, F., Bobinger, S., & Neufeld, M. (2013). When a robot's group membership matters. *International Journal of Social Robotics*, *5*, 409–417.

[24] Stroessner, S. J., & Koya, P. D. (in press). Thinking socially about the nonsocial world. To appear in D. Carlston, K. Hugenberg, & K. L. Johnson (Eds). *Oxford Handbook of Social Cognition*, *Volume 2*. New York, NY: Oxford University Press.

[25] Eyssel, F., & Hegel, F. (2012). (S)he's got the look: Gender stereotyping of robots. *Journal of Applied Social Psychology*, *42*(9), 2213–2230.

[26] Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, *51*(4), 724–731.

[27] Leung, Y., Oates, J., & Chan, S. P. (2018). Voice, articulation, and prosody contribute to listener perceptions of speaker gender: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, *61*(2), 266–297.

[28] Ko, S. J., Judd, C. M., Stapel, D. A. (2009). Stereotyping based on voice in the presence of individuating information: Vocal femininity affects perceived competence but not warmth. *Personality and Social Psychology Bulletin*, *35*(2), 198–211.

[29] Oleszkiewicz, A., Pisanski, K., Lachowicz-Tabaczek, K., & Sorokowska, A. (2017). Voice-based assessments of trustworthiness, competence, and warmth in blind and sighted adults. *Psychonomic Bulletin & Review*, *24*, 856–862.

[30] Bryant, G. A. (2022). Vocal communication across cultures: Theoretical and methodological issues. *Philosophical Transactions of the Royal Society B*, *377*(1841).

[31] Ko, S. J., Sadler, M. S., Galinsky, A. D. (2015). The sound of power: Conveying and detecting hierarchical rank through voice. *Psychological Science, 26(1)*, 3–14.

[32] Krahé, B., Uhlmann, A., & Herzberg, M. (2021). The voice gives it away: Male and female pitch as a cue for gender stereotyping. *Social Psychology*, *52*(2), 101–113.

[33] Schroeder, J., & Epley, N. (2016). Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology: General*, *145*(11), 1427–1437.

[34] Miwa, H., Takanishi, A., & Takanobu, H. (2001). Experimental study on robot personality for humanoid head robot. *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the Next Millennium*, 1183–1188.

[35] Lu, L., McDonald, C., Kelleher, T., Lee, S., Chung, Y. J., Mueller, S., Vielledent, M., & Yue, C. A. (2022). Measuring consumer-perceived

humanness of online organizational agents. *Computers in Human Behavior*, *128*(107092).

[36] Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117.

[37] Moussawi, S., Benbunan-Fich, R. (2021). The effect of voice and humour on users' perceptions of personal intelligent agents, *Behaviour & Information Technology*, *40*(15), 1603–1626.

[38] Walters, M. L., Syrdal, D. S., Koay, K. L., Dautenhahn, K., & te Boekhorst, R. (2008). Human approach distances to a mechanical-looking robot with different robot voice styles. *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN*, 707–712.

[39] Kühne, K., Fischer, M. H., & Zhou, Y. (2020). The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. Evidence from a subjective ratings study. *Frontiers in Neurorobotics*, *14*.

[40] Amazon (2023). *The Alexa personality*. https://developer.amazon.com/en-US/alexa/branding/alexa-guidelines/communication-guidelines/brand-voice

[41] Abercrombie, G., Cercas Curry, A., Pandya, M., & Rieser, V. (2021). Alexa, Google, Siri: What are your pronouns? Gender and anthropomorphism in the design and perception of conversational assistants. *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, 24–33.

[42] Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103.

[43] Seaborn, K., & Pennefather, P. (2022). Neither "hear" nor "their": Interrogating gender neutrality in robots. *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, 1030–1034.

[44] Søraa, R. A. (2017). Mechanical genders: How do humans gender robots? *Gender, Technology and Development*, *21(*1–2), 99–115.

[45] Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, *132*, 138–161.

[46] Loideain, N. N., & Adams, R. (2020). From Alexa to Siri and the GDPR: The gendering of virtual personal assistants and the role of data protection impact assessments. *Computer Law & Security Review*, *36*(105366).

[47] 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.

[48] Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

[49] Dou, X., Wu, C.-F., Niu, J., & Pan, K.-R. (2022). Effect of voice type and head-light color in social robots for different applications. *International Journal of Social Robotics*, *14*, 229–244.

[50] Dou, X., Yan, L., Wu, K., & Niu, J. (2022). Effects of voice and lighting color on the social perception of home healthcare robots. *Applied Sciences*, *12*(23), 12191.

[51] Tolmeijer, S., Zierau, N., Janson, A., Wahdatehagh, J., Leimesiter, J. M., & Bernstein, A. (2021). Female by default? – Exploring the effect of voice assistant gender and pitch on trait and trust attribution. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7.

[52] Ernst, C. P., & Herm-Stapelberg, N. (2020). Gender stereotyping's influence on the perceived competence of Siri and Co. *Hawaii International Conference on System Sciences*, 4448–4453.

[53] Powers, A., & Kiesler, S. (2006). The advisor robot: Tracing people's mental model from a robot's physical attributes. *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, 218–225.

[54] Benitez, J., Wyman, A. B., Carpinella, C. M., & Stroessner, S. J. (2017). The authority of appearance: How robot features influence trait inferences and evaluative responses. *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 397–404.

[55] Stroessner, S. J., & Benitez, J. (2019). The social perception of humanoid and nonhumanoid robots: Effects of gendered and machinelike features. *International Journal of Social Robotics*, *11*(2), 305–315.

[56] Pan, M. K. X. J., Croft, E. A., Niemeyer, G. (2018). Evaluating social perception of human-to-robot handovers using the robot social attributes scale (RoSAS). *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 443–451.

[57] Harris-Watson, A. M., Larson, L. E., Lauharatanahirun, N., DeChurch, L. A., & Contractor, N. S. (2023). Social perception in human-AI teams: Warmth and competence predict receptivity to AI teammates. *Computers in Human Behavior*, *145*.

[58] Fleischmann, A., Sieverding, M., Hespenheide, U., Weiß, M., & Koch, S. C. (2016). See feminine—think incompetent? The effects of a feminine outfit on the evaluation of women's computer competence. *Computers & Education*, *95*, 63–74.

[59] Eagly, A. H., & Mladinic, A. (1994). Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. *European Review of Social Psychology*, *5*(1), 1–35.

[60] Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., & Sczesny, S. (2020). Gender stereotypes have changes: a cross-temporal meta-analysis of U.S. public opinion polls from 1946 to 2018. *American Psychologist*, *75*(3), 301–315.

[61] Hehman, E., Sutherland, C. A. M., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology*, *113*(4), 513–529.

[62] Haines, E. L., Deaux, K., & Lofaro, N. (2016). The times they are a-changing … or are they not? A comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly*, *40*(3), 353–363.

[63] Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191

[64] Boersma, P., & Weenink, D. (2021). *Praat: doing phonetics by computer* (6.1.56).

[65] Titze, I. R. (1994). *Principles of voice production*. Prentice Hall.

[66] Audacity Team. (2021). *Audacity(R): Free audio editor and recorder* (3.0.0).

[67] Bryant, G. A., & Fox Tree, J. E. (2005). Is there an ironic tone of voice? *Language and Speech*, *48*(3), 257–277.

[68] Yilmazyildiz, S., Read, R., Belpeame, T., Verhelst, W. (2016). Review of semantic-free utterances in social human-robot interaction. *International Journal of Human-Computer Interaction*, *32*(1), 63–85.

[69] Belin, P., Bestelmeyer, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology*, *102*(4), 711–725.

[70] Lattner, S., & Friederici, A. D. (2003). Talker's voice and gender stereotype in human auditory sentence processing - Evidence from event-related brain potentials. *Neuroscience Letters*, *339*(3), 191–194.

[71] Carpenter, J., Davis, J. M., Erwin-Stewart, N., Lee, T. R., Bransford, J. D., & Vye, N. (2009). Gender representation and humanoid robots designed for domestic use. *International Journal of Social Robotics*, *1*(3), 261–265.

[72] 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.

[73] Bosson, J. K., Prewitt-Freilino, J. L., & Taylor, J. N. (2005). Role rigidity: A problem of identity misclassification? *Journal of Personality and Social Psychology*, *89*(4), 552–565.

[74] Olson, C., & Kemery, K. (2019). *2019 voice report: Consumer adoption of voice technology and digital assistants*. Microsoft. https://about.ads.microsoft.com/en-us/insights/2019-voice-report

[75] Caplan, P. J., MacPherson, G. M., & Tobin, P. (1985). Do sex-related differences in spatial abilities exist? A multilevel critique with new data. *American Psychologist*, *40*(7), 786–799.

[76] Niculescu, A., Van Dijk, B., Nijholt, A., & See, S. L. (2011). The influence of voice pitch on the evaluation of a social robot receptionist. *Proceedings of the 2011 International Conference on User Science and Engineering*, 18–23.

[77] Wu, H. X., Li, Y., Ching, H. H. B., & Chen, T. T. (2023). You are how you speak: The roles of vocal pitch and semantic cues in shaping social perceptions. *Perception*, *52*(1), 40–55.