CHALMERS
UNIVERSITY OF TECHNOLOGY

Hearing Ambiguity: Exploring Beyond-Gender Impressions of Artificial Ambiguous Voices

Hearing Ambiguity: Exploring Beyond-Gender Impressions of Artificial Ambiguous Voices

Martina De Cet
Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden
demart@chalmers.se

Katie Seaborn
Department of Industrial Engineering and Economics
Institute of Science Tokyo
Tokyo, Japan
katie.seaborn@cst.cam.ac.uk

Mohammad Obaid
Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden
mobaid@chalmers.se

Ilaria Torre
Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden
ilariat@chalmers.se

Abstract

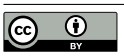
Voice perception plays a fundamental role in all types of interactions, from human-to-human communication to human-technology interaction. When it comes to technology, we sometimes have the option to choose the type of voice we want to hear. But why is the default (almost) always a feminine or masculine voice? In this research, we evaluated user perceptions of gender-ambiguous voices, a relatively unexplored option. In our novel comparative study, we evaluated six gender-ambiguous voices with participants of diverse gender identities (men, women, and non-binary individuals), with 74 participants in each group. Additionally, half of the participants were told in advance that the voices had been designed to be gender-ambiguous, and half were not. We aimed to move beyond subjective perceptions of voice gender by exploring how such voices are perceived across different dimensions: trustworthiness, appeal, comfort, anthropomorphism, and aversion. Our findings reveal that while men and women had similar perceptions, non-binary participants rated the voices more negatively, with lower trust and higher aversion. Interestingly, priming participants about the voices' ambiguity did not significantly affect overall perceptions, though it increased critical evaluations from non-binary individuals. These findings contribute to growing research on gender-ambiguous voices by providing perceptual comparisons of multiple voices and highlighting the need for more inclusive voice designs that appeal to non-binary users.

CCS Concepts

• **Human-centered computing** → *User studies*; **Sound-based input / output**; **Human computer interaction (HCI)**.

Keywords

Artificial Voice, Gender, Gender Ambiguity, Voice Assistants, User Perceptions



CUI '25, Waterloo, ON, Canada

ACM ISBN 979-8-4007-1527-3/25/07
https://doi.org/10.1145/3719160.3736622

ACM Reference Format:
Martina De Cet, Katie Seaborn, Mohammad Obaid, and Ilaria Torre. 2025. Hearing Ambiguity: Exploring Beyond-Gender Impressions of Artificial Ambiguous Voices. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25), July 08–10, 2025, Waterloo, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3719160.3736622

1 Introduction

Voice interaction with technology has become an integral part of daily life, facilitating tasks like setting timers on smartphones and providing assistance while driving [14, 47, 54, 61]. Users engage with computer voices embedded in speech-based virtual agents and voice assistants (VAs) and expect these agents to respond appropriately. Users also interpret the voice phenomena they hear in ways beyond mere content, notably reading in cues to humanlikeness and artificiality [14, 44, 54]. A dominant cue is *gender* [15, 22, 54, 61], known as one of the most stable social identity categories across human history [29]. Many VAs, when asked, *"What is your gender?"* respond by claiming that they have no gender. Siri, for example, replies: *"I am genderless. Like cacti. And certain species of fish."* or *"Animals and French nouns have genders. I do not."* Despite these answers, the design of these agents can still reinforce gendering in other ways, e.g., through names like Alexa, Siri, or Cortana. While the perception of these names as feminine may vary across cultural and individual contexts, they are often interpreted as such, particularly in Western settings where these assistants were initially developed and marketed. Designers, knowingly or not, make choices that lead to user gendering, not only through names and identity statements but also colour [26], clothing [26, 56], (body) shape [6, 45, 53, 72], and more. Avoiding the issue or asserting that the technology has no gender is not enough. People are especially primed to gender when humanlike cues are present in voice and speech [44, 53]. This phenomenon does not only concern VAs. Other agents, such as robots and virtual humans, are typically categorised as either masculine or feminine [46]. In the fields of Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI), assigning gender to the agent, both by participants and researchers, is a common practice [6, 45, 46].

Gendering has been linked to a variety of effects. Positive effects include improved user attitudes towards the agent, an increase in perceived trustworthiness, and overall acceptance of the agent [15, 22]. Gendering, however, is not without issue. The default voices of most agents are almost exclusively feminine and to a lesser extent masculine [5, 54]. This leaves out gender-diverse voices that could be perceived as ambiguous, i.e., a combination of masculine and feminine [19, 56, 65, 68, 80], gender elusive, i.e., difficult to determine or non-generalisable [24], or neutral, i.e., genderless or gender-free [41, 56, 59], or even associated with human identities like non-binary [17, 35]. Agent gendering can also reinforce and extend human stereotypes and biases, including in terms of perceived agency [22] and emotional intelligence [13].

Yet, voice may also be manipulated to create more inclusive options and/or reduce gendering in human-technology interactions. The Text-To-Speech (TTS) community has begun developing voices that are not strictly male/masculine or female/feminine. These voices, sometimes referred to as "gender-neutral", "gender-less", "non-binary" or gender-ambiguous, aim to sound clear and natural while moving beyond the gender binary of man/woman [41, 65, 79, 80]. The availability of a third, non-gendered option is not only valuable for reducing the risk of gendering and stereotyping technology but is also a matter of inclusivity [64]. Gender-nonconforming individuals, people who do not fit neatly into the categories of man or woman, male or female, masculine or feminine, may feel more represented and included in a society that offers alternatives beyond the binary [17, 19, 35, 64]. Perceptions of voice, not only genderedness, but acceptability, approach distance, trust and reliance, and other impressions, may also be tied to the gender of the individual via similarity to self [8, 15, 17, 50, 75], although this remains underexplored. Notably, this has been virtually unexplored with those outside of the binary masculine and feminine.

Throughout this paper, we will use the term "gender-ambiguous". As noted in De Cet et al. [19]'s review, "gender-ambiguous" reflects a resistance to clear categorisation, blending both masculine and feminine traits. In contrast, "non-binary" is more closely linked to identity than voice, while terms like "gender-neutral", "genderless", and "gender-free" suggest the complete removal of gendered traits, which is not achievable in voices [65].

While gender-ambiguous voices hold great potential, it is crucial to understand how they are perceived beyond "what gender". At present, almost all work on gender-ambiguous and related voice phenomena with human-technology interactions is related to perceptions of voice gender [17, 19]. Understanding the extent to which voices are gendered and how these perceptions transcend individuals is important, but only the first step [58]. At the same time, perceptions of voice gender ambiguity may vary across user groups, including men, women, non-binary individuals, and other gender-diverse folk [17, 19, 20, 53]. What remains is a gap at the intersection of gender-diverse user groups, especially the non-binary one [17, 56, 64], and perceptions beyond gender impressions.

To address this gap, we conducted a study in which gender-diverse participants (women, men, and non-binary people) listened to six artificial gender-ambiguous voices and rated them across five perception dimensions: trustworthiness, appeal, comfort, anthropomorphism, and aversion. To our knowledge, prior research has neither explored ambiguous voices beyond gender classification nor allowed participants to compare multiple ambiguous voices side by side. We also introduced a priming condition to test whether drawing attention to the voices' ambiguity would shape listener perceptions. Our aim is to advance the development of voice-based technologies that are inclusive and resistant to gender stereotyping.

This study is guided by three research questions: **RQ1**: *How do participants evaluate ambiguous voices across different voice perception dimensions?*, **RQ2**: *Does priming participants on the gender ambiguity of artificial voices affect their perception of such voices?*, **RQ3**: *Does participant gender influence their perception of artificial gender-ambiguous voices?*. By addressing these questions, we offer new insights into how ambiguous voices are perceived and contribute to more thoughtful design of future voice interfaces.

2 Related Work

2.1 Gender Perception and the Role of Voice

Gender is deeply embedded in human societies [29] and thereby user mental models and perceptions of humanlike entities [25, 44]. When we meet someone, our brain instinctively evaluates a multimodal variety of cues, what we see, hear, smell, and so on, including spoken names, facial features, body shape, clothing, nonverbal behaviours, and voices. All of these are linked to judgments about the gender of the person or agent [29, 38, 45, 74]. At the same time, humanity is diverse and complex, and our brains are adaptive [43, 77]. For instance, Weirich and Simpson [77] found that voices who self-attributed a certain gender identity influenced listener perceptions, shifting in kind. Thus, the act of gendering, interpreting gender cues in stimuli, is a multifaceted and dynamic process.

Gendering extends to agentic technologies as well. Seaborn and Frank [53], for instance, found that researchers and participants alike unthinkingly gendered the Pepper robot by way of the pronouns and other gendered referents they used to refer to the ambiguous bot. Notably, researchers who did not realise this significantly influenced participant gendering. People process not only gender cues but cues linked to social identities immediately and unconsciously. For instance, Fujii et al. [24] investigated identity-pluralistic Japanese first-person pronouns, linked to gender, age, region, and formality, with ChatGPT voices. They found that voices using matched gender pronouns led Japanese listeners to imagine personas consistent with those identities in Japanese culture, except the feminine Juniper voice, which was *more* feminine when *not* using the "gender-neutral" *watashi* pronoun. Notably, when Juniper used *boku*—a pronoun typically associated with masculine speakers in Japanese—and Ember used *atashi*—a pronoun typically used by feminine speakers—the combinations disrupted expectations and queered the listener's mental models of gender. While mental models about gender exist and are stable, they can be disrupted by manipulating voice and cues relevant to social identities.

2.2 Designing Gender-Ambiguous Voices

Gender-ambiguous voices and other voices that defy the gender binary present a novel and potentially transgressive option. Research has shown that gender-ambiguous voices can influence how users perceive and gender agents. For instance, Torre et al. [68] found that using a gender-ambiguous voice reduces the tendency to gender robots, even when the agent has stereotypically masculine

or feminine attributes. Their findings highlight voice as a key factor in shaping perceptions, suggesting its potential for creating more inclusive and less gendered user interactions. Expanding on this, De Cet et al. [20] examined whether ambiguous voices influenced how people sketched robots in different occupational contexts. Participants who heard a gender-ambiguous voice were less likely to assign a gender to their sketches compared to those who heard a masculine or feminine voice. This reinforces the idea that voice ambiguity can reduce automatic gender associations.

Achieving truly ambiguous voices, however, remains challenging. This can be traced to the dominant model in most human societies being gender-binary: male/female, masculine/feminine, man/woman [29]. Our brains are trained on the binary, even when we have a gender identity that is not exclusively binary. Gender perceptions tend to be tied to voice formant frequencies, notably the fundamental frequency or *F*0 [27, 43]. However, the reality is more complicated. Mullennix et al. [43] argue that an "auditory composite" (p. 3091) of not only formants but also other auditory qualities, even breathiness, influence gender perceptions. Seaborn et al. [55] found that computer voices deemed gender-ambiguous may also seem younger (as well as more anthropomorphic). In general, feminine voices are deemed higher-pitched, masculine voices lower-pitched, and those pitched slightly lower than the midpoint between these binary poles tend to be perceived as androgynous, i.e., ambiguous [62]. However, as mentioned, other auditory and nonverbal cues can disrupt and shift impressions.

The classification of gender-ambiguous voices is not yet standardised. Tolmeijer et al. [66] decided on gender ambiguity for the voice based on a rating of "58% male and 42% female" (p. 3). Seaborn et al. [55, 59] and Mandai et al. [36] provided a categorical option: "Both feminine and masculine elements" (roughly translated from Japanese). Mooshammer and Etzrodt [42] used a 7-point Likert scale with the option "neither/I cannot judge" (compared to binary options). Danielescu et al. [17] used a 5-point Likert scale for items asking about whether the voice sounded non-binary. Torre et al. [68] used a 5-Point Likert scale with the option "Ambiguous". While a range of methods is available, we must take heed: evaluators are often prescriptive [64] rather than gender-expansive [27, 57, 58, 64]. Here, gender-expansive refers to being open to a broader, more fluid understanding of gender when providing response options to users or when coding responses provided by users. A lack of this perspective can unduly influence perceptions, as seen in the gendering of Pepper [53], or restrict participants from expressing what they actually hear. We thus used a gender-expansive approach akin to those in Seaborn et al. [55, 59] and Mandai et al. [36].

One's own gender identity may influence perceptions. Danielescu et al. [17] demonstrated how men, women, and non-binary people rate masculine, feminine, and non-binary voices differently. Notably, while many of the 1,010 respondents to the voice perceptions survey had no preference (23%-36%), men overwhelmingly favoured the masculine voice, women the feminine voice, and to a lesser extent, non-binary people the non-binary voice. Moreover, non-binary participants overwhelmingly rated high trust and intelligence scores for the non-binary voice compared to the masculine and feminine voices. This suggests that a preference for voice is linked to self-gender. In contrast, Tolmeijer et al. [66] found that women trusted the gender-ambiguous voice (Google WaveNet *en-US-Wavenet-F* downshifted by three semitones) more than men and rated it as more friendly, even though the voices were generally perceived similarly overall. In Hope and Lilley [27], "gender-expansive" people started from a "neutral" gender baseline and, alongside women, were less inclined to gender along the binary. In sum, perceptions of gender-ambiguous voices may depend on the individual's gender identity or individual factors. Our work, providing one of the few comparative multi-voice within-subject designs, contributes to resolving the extent to which this is the case.

2.3 Moving Beyond Gender: Toward Richer Voice Evaluations

Another challenge relates to the dearth of gender-ambiguous voices available. Still, recent developments in TTS technology are aiming to create gender-ambiguous, gender-neutral, gender-evasive, and non-binary voices as provocative prototypes—Project Q [1], for instance—and research stimuli, seeking a way to further reduce gendered assumptions [56, 65] and provide inclusive options [17, 64]. One notable effort comes from Danielescu et al. [17], who developed a "non-binary" TTS voice by blending masculine and feminine speech characteristics. This work went beyond simply avoiding gender bias by actively engaging with non-binary and transgender communities to ensure naturalness, representation, and comfort. At the same time, we must be careful not to relegate human social identities, especially marginalised ones, to machines, nor equate "gender ambiguity" in voice phenomena with complex human social identities [9, 21, 52, 64]. Voice and other cues linked to gender/ing may be perceived in ways that do not match the identity of a human individual, which must be respected.

Most work so far has focused on evaluating voice gender perceptions (or their absence), essentially acting as manipulation checks for voice gender [58]. De Cet et al. [19] conducted a systematic review on gender-ambiguous voices and found that one-third of the 36 studies analysed focused exclusively on voice perceptions. Such assessments are a crucial first step and, given the malleability of perceptions and individual factors involved, should continue to feature in human-subject work [58]. Yet, perceptions alone do not reveal or predict effects on attitudes and behaviour when these voices are deployed and used by people.

Some work has begun to explore other kinds of perceptions about ambiguous voices beyond gender. Längle et al. [35] considered user perceptions of non-binary voices in digital assistants. They assessed perceived gender, likability, gender-stereotypical perceptions, and personality traits using established psychological scales. They also examined gender role stereotypes and affinity for technology, providing a broader perspective on how users relate to non-binary voices. Markopoulos et al. [37] focused on generating multilingual gender-ambiguous TTS voices, evaluating them based on naturalness and gender perception. Tolmeijer et al. [66] considered gendered trait ascriptions and trust perceptions for a gender-ambiguous voice (Google WaveNet *en-US-Wavenet-F* downshifted by three semitones) compared to binary voices, finding no significant differences by voice gender; as discussed, men and women rated the voice differently. On that note, they only included binary gender participants. These findings highlight a research gap: while gender perception is well-studied, particularly among binary people,

other factors—like trust [70, 73, 76], appeal [54, 60], anthropomorphism [54, 59, 60, 76], comfort [60], and aversion [2, 44]—remain largely unexplored for gender ambiguous voices.

3 Method

We explored how gender-ambiguous voices were perceived across different gender groups and voice perception dimensions, whether priming influences perceptions, and how. Thus, we had two between-subjects factors (participant gender group and priming) and one within-subjects factor (artificial gender-ambiguous voices, of which there were six; refer to section 3.2). Priming was done by either informing participants that the voices were ambiguous (primed condition) or providing no information (not primed condition). Half of the participants (with equal numbers of men, women, and non-binary people) were randomly assigned to the primed condition, and half to the non-primed condition. In Table 1, the number of participants for each condition is summarised. All participants listened to the same set of six voices in a randomised order. In Table 2, we present some key acoustic characteristics (pitch, speech rate, and accent) of the six collected ambiguous voices. Voice features were extracted using Praat [7], a free package for phonetic analyses.

Priming	Gender	Number of Participants
Primed	Men	37
	Women	37
	Non-binary	37
Not primed	Men	37
	Women	37
	Non-binary	37

Table 1: Number of participants for each condition (primed vs. not primed) and gender identity.

3.1 Participants

An a priori power analysis was conducted to determine the sample size for this study; results indicated that the required sample size to achieve 80% power for detecting a medium effect size, at a significance criterion of $\alpha = .05$, with a mixed design, was $N = 222$. Therefore, we recruited 222 participants from the UK, aged 18 to 74 ($\mu = 38.176$, $\sigma = 13.621$; Women = 74, Men = 74, Non-binary = 74). Data was collected in February 2025 on Prolific. Gender and nationality were retrieved from Prolific for screening.

3.2 Materials and Instrument

The voices used in this study were sourced from our previous work [20]. Following a literature review, we reached out to the developers of gender-ambiguous voices to request voice samples. The collected voices include "Can_1" and "Can_2" from Torre et al. [68], "Sam" from Danielescu et al. [17], "Mini" from Prabha et al. [49], and "Blue" and "Fable" from SpeechGen.io [63]. All these voices were previously validated as "gender-ambiguous" with user studies (see [20]). The voices were developed following a variety of techniques. Most voices are in the 130-160 Hz pitch range, thus falling in between typical male and female voice pitch ranges [30] – except "Mini", which had an average pitch of 248 Hz, and "Blue", with an average pitch of 197 Hz. Additionally, most voices had American accents, except "Mini" (Indian-accented English) and "Fable" (British English). The audio files are available in the supplementary materials.

The voices were evaluated using a novel instrument for assessing voice impressions, the Voice Experience Inventory (VOXI), which is currently under peer review and available in the supplementary materials. This instrument was developed based on community consensus derived from a preliminary rapid review in 2021 [60], later updated with a systematic literature review that used the international standard Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist. Coverage included unvalidated and validated instruments from user studies of voice impressions published between 2000 and 2022. 28 potential factors were derived from 99 items across 32 instruments in 49 studies. The initial instrument was evaluated with four computer voices (feminine and masculine by young and old) in an online user perceptions study ($N = 100$). Exploratory factor analysis resulted in 24 items (four removals). Confirmatory factor analyses on data from a new suite of six voices (masculine and feminine by child, adult, and older adult) and participants ($N = 150$) confirmed that the 24 items were contained within five underlying factors:

(1) Trustworthiness (Practical, Reliable, Trustworthy, Dependable, Intelligent, Knowledgeable, Competent, Accurate)
(2) Appeal (Attractive, Enjoyable, Happy, Pleasant)
(3) Comfort (Warm, Kind, Friendly, Relaxing, Calming)
(4) Anthropomorphism (Human-like, Machine-like, Natural, Fake)
(5) Aversion (Scary, Annoying, Frustrating)

Participants used a 7-point Likert scale (0 = strongly disagree to 6 = strongly agree, with neither agree nor disagree as the midpoint) to rate the voices across the 24 items.

3.3 Procedure

The study was conducted following the ethical guidelines of the Computer Science and Software Engineering Department at the Chalmers University of Technology, Sweden. Participants first read the study description, provided informed consent, and reported their age. Those in the primed group were informed that the voices they would hear were ambiguous, using the following text: *"In this experiment, you will listen to six ambiguous voices, one after the other. We are interested in what you think these voices sound like. An ambiguous voice is a voice that does not clearly fit into traditional male or female categories by blending both masculine and feminine characteristics."* This definition was adapted from De Cet et al. [19]. Participants in the non-primed group were shown the following text: *"In this experiment, you will listen to six voices, one after the other. We are interested in what you think these voices sound like."*

Before starting, a test sound (available in the supplementary materials) was played to help participants adjust the volume of their headphones. Then, they listened to six voices, one at a time, and rated them using the 7-point Likert scale items from VOXI (refer to 3.2). The order of the six voices was randomised for each participant but not fully counterbalanced, as full counterbalancing would have required an impractically large sample size. Each voice spoke the same utterance: *"Hello! I'm a robot and I work as a tour guide. My work consists of guiding people around museums and galleries!"*. These utterances were retrieved from samples received from researchers

Voice name	Pitch (Hz)	Speech Rate (words/sec)	Accent
Sam [17]	142.63	2.54	American English
Can_1 [68]	138.17	2.39	American English
Can_2 [68]	157.65	2.66	American English
Mini [48]	248.68	2.40	Indian-accented English
Blue [63]	174.23	2.67	American English
Fable [63]	124.33	2.82	British English

Table 2: Acoustic characteristics of the 6 collected ambiguous voices: pitch, speech rate and accent.

who have recently developed gender-ambiguous artificial voices [17, 49, 68]. The utterance was chosen as it is semantically neutral (as validated in a previous study [20]). Utterances were kept short (~8s) to avoid participant fatigue; nevertheless, first impressions based on voice are accurate and reliable even after very short auditory stimuli [39]. Participants were instructed to focus solely on the sound of the voice and not to the linguistic content. This was emphasised at multiple points throughout the procedure: 1) before the experiment began, participants were told, *"We are interested in what you think these voices sound like"*, 2) before listening to the test sound: "*Please do not pay attention to what the voice is saying. Focus only on the sound of the voice itself"*, and 3) after listening to the test sound: *"Before you begin, please keep in mind that we are interested in your initial gut reaction to the voices (*ambiguous voices *in the primed condition). If you understand, you may proceed with the experiment"*.

At the end of the experiment, participants were asked to note if they had experienced any technical issues, and were given the option to leave an open-ended comment on the whole experience.

The completion time for the experiment ranged approximately between 7 and 11 minutes, and participants were compensated at a rate of £9.00 per hour, following Prolific standards.

4 Results

In this section, we first outline the data analysis process, followed by the presentation of our results. The results are organised by factor and compared across voice and participant gender groups within each factor. Summary statistics can be found in Table 3.

4.1 Data Analysis

Quantitative analyses were conducted in R version 4.4.2. The 24 total ratings were combined to obtain five voice perception factors (trustworthiness, appeal, comfort, anthropomorphism, aversion), following instructions from the developers of VOXI; for each of these factors, we built ANOVA models with perception factor as dependent variable, and voice (Blue, Can_1, Can_2, Fable, Mini, Sam), participant gender (woman, man, non-binary), and priming (primed, not primed) as independent variables. Additionally, we looked at interactions between the independent variables. Post hoc pairwise comparisons were run using the Tukey HSD method, which adjusts the alpha value to account for multiple comparisons (see Table 4). Internal consistency of the VOXI questionnaire was assessed using Cronbach's α, showing good to excellent reliability for each scale (trustworthiness = 0.95; appeal = 0.92; comfort = 0.93; anthropomorphism = 0.94; aversion = 0.81).

4.2 Trustworthiness

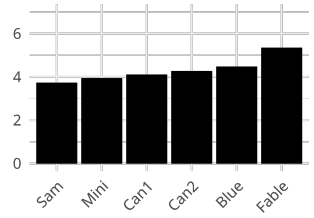


Figure 1: Mean ratings of trustworthiness for each of the six gender-ambiguous voices considered in the study.

We found a significant main effect of voice ($F(5, 1321) = 71.84$, $p < .001$, $\hat{\eta}^2_G = .214$) and participant gender ($F(2, 1321) = 12.62$, $p < .001$, $\hat{\eta}^2_G = .019$), and a significant interaction between priming and participant gender ($F(2, 1321) = 12.01$, $p < .001$, $\hat{\eta}^2_G = .018$).

For the main effect of voice, post-hoc comparisons can be found in Table 4. Regarding the main effect of participant gender, as shown in Figure 2, we found that non-binary participants rated the voices as less trustworthy than women and men participants (all $p < .001$). Finally, for the interaction effect, we found that non-binary participants who weren't primed rated the voices as less trustworthy than all the other participants groups (all $p < .001$).

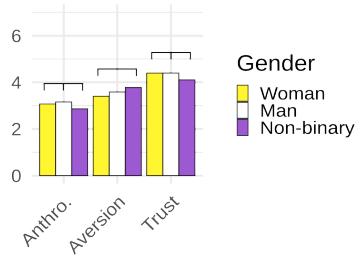


Figure 2: Mean ratings of anthropomorphism, aversion and trustworthiness divided by participant gender.

4.3 Appeal

We found a significant effect of voice ($F(5, 1321) = 145.09$, $p < .001$, $\hat{\eta}^2_G = .354$), and a significant interaction between priming and participant gender ($F(2, 1321) = 4.74$, $p = .009$, $\hat{\eta}^2_G = .007$).

For the main effect of voice, post-hoc pairwise comparisons are in Table 4. The interaction effect sank below the threshold for

Voice name	Trustworthiness	Appeal	Comfort	Anthropomorphism	Aversion
Sam	3.72 (1.07)	2.33 (0.95)	2.55 (1.04)	1.84 (0.95)	4.46 (1.30)
Mini	3.93 (1.11)	3.27 (1.34)	3.60 (1.30)	2.75 (1.42)	3.92 (1.47)
Can_1	4.10 (1.10)	2.86 (1.23)	3.19 (1.34)	2.09 (1.09)	3.94 (1.39)
Can_2	4.26 (1.00)	3.41 (1.08)	3.54 (1.04)	2.73 (1.19)	3.66 (1.26)
Blue	4.47 (0.95)	3.60 (1.21)	3.68 (1.19)	3.38 (1.45)	3.44 (1.35)
Fable	5.34 (0.84)	5.07 (1.01)	5.00 (0.95)	5.42 (1.20)	2.13 (0.97)

Table 3: Summary statistics (mean and standard deviation) of the voice ratings across the 5 VOXI scales.

statistical significance ($\alpha = .05$) after taking into account correction for multiple comparisons using the Tukey HSD test.

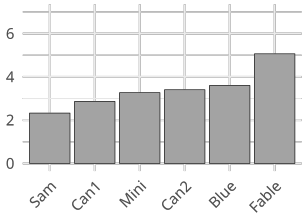


Figure 3: Mean ratings of appeal for each of the six gender-ambiguous voices.

4.4 Comfort

We found a significant effect of voice ($F(5, 1321) = 109.09$, $p < .001$, $\hat{\eta}^2_G = .292$), and a significant interaction between priming and participant gender ($F(2, 1321) = 6.25$, $p = .002$, $\hat{\eta}^2_G = .009$).

For the main effect of voice, post-hoc pairwise comparisons can be found in Table 4. Regarding the interaction effect, after accounting for multiple comparisons, the only significant comparison within the interaction effect was found for the non-binary, non-primed participants, who rated the voices lower than the men non-primed participants ($p = .003$).

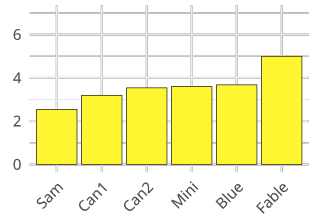


Figure 4: Mean ratings of comfort for each of the six gender-ambiguous voices.

4.5 Anthropomorphism

We found a significant main effect of voice ($F(5, 1321) = 248.15$, $p < .001$, $\hat{\eta}^2_G = .484$) and participant gender ($F(2, 1321) = 6.97$, $p < .001$, $\hat{\eta}^2_G = .010$), and a significant interaction between priming and participant gender ($F(2, 1321) = 4.86$, $p = .008$, $\hat{\eta}^2_G = .007$).

For the main effect of voice, post-hoc pairwise comparisons can be found in Table 4. Regarding participant gender (Figure 2), we found that the non-binary participants rated the voices as less anthropomorphic than women ($p = .03$) and men ($p < .001$). Finally, for the interaction between priming and participant gender, men who were not primed rated the voices higher than men who were primed and non-binary people primed ($p = .04$ and $p = .008$, respectively), and non-binary people who were not primed rated the voices lower than men who were not primed ($p < .001$).

Voice pair	Trustworthiness	Appeal	Comfort	Anthropomorphism	Aversion
Fable - Blue	p <.001	p <.001	p <.001	p <.001	p <.001
Fable - Can_2	p <.001	p <.001	p <.001	p <.001	p <.001
Fable - Can_1	p <.001	p <.001	p <.001	p <.001	p <.001
Fable - Mini	p <.001	p <.001	p <.001	p <.001	p <.001
Fable - Sam	p <.001	p <.001	p <.001	p <.001	p <.001
Blue - Can_2	n.s.	n.s.	n.s.	p <.001	n.s.
Blue - Can_1	p = .001	p <.001	p <.001	p <.001	p <.001
Blue - Mini	p <.001	p = .03	n.s.	p <.001	p <.001
Blue - Sam	p <.001	p <.001	p <.001	p <.001	p <.001
Can_2 - Can_1	n.s.	p <.001	p = .02	p <.001	n.s.
Can_2 - Mini	p = .008	n.s.	n.s.	n.s.	n.s.
Can_2 - Sam	p <.001	p <.001	p <.001	p <.001	p <.001
Mini - Can_1	n.s.	p = .003	p <.001	p <.001	n.s.
Can_1 - Sam	p = .001	p <.001	p <.001	n.s.	p <.001
Mini - Sam	n.s.	p <.001	p <.001	p <.001	p <.001

Table 4: Post-hoc pairwise comparisons between the 6 gender-ambiguous voices, across 5 VOXI scales. Comparisons were conducted using Tukey HSD tests, which account for multiple comparisons.

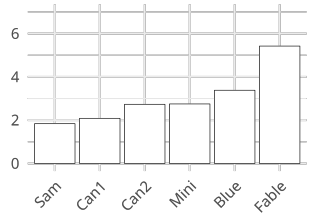


Figure 5: Mean ratings of anthropomorphism for each of the six gender-ambiguous voices.

4.6 Aversion

For the aversion factor, we found a main effect of voice ($F(5, 1321) = 86.21$, $p < .001$, $\hat{\eta}^2_G = .246$), participant gender ($F(2, 1321) = 9.57$, $p < .001$, $\hat{\eta}^2_G = .014$), and a significant interaction between priming and participant gender ($F(2, 1321) = 10.29$, $p < .001$, $\hat{\eta}^2_G = .015$). Again, priming was not a statistically significant factor.

Post-hoc pairwise comparisons for the main effect of voice can be found in Table 4. Regarding gender (see Figure **??**), we found that non-binary participants rated the voices as higher in aversion than the women participants ($p < .001$). Finally, for the interaction effect, we found that non-binary non-primed participants rated the voices higher on aversion compared to women non-primed and primed, men non-primed, and non-binary primed (all $p < .001$). Also, men-primed participants rated the voices higher than women-primed and non-primed ($p = .05$ and $p = .03$, respectively).

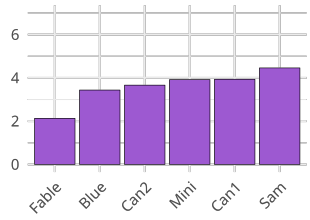


Figure 6: Mean ratings of aversion for each of the six gender-ambiguous voices.

4.7 Summary

In sum, we found a main effect of voice across all VOXI scales; a main effect of participant gender for the trustworthiness, anthropomorphism, and aversion scales; and limited secondary effects of participant priming. The results are further summarised in Table 5.

4.8 Open-ended Comments

As the question *"Do you have any other comments?"* was optional, only a few responses were collected. Apart from some remarks describing the experiment as fun, interesting, or intriguing—made by participants across all groups, women (W), men (M), and non-binary (NB)—some other comments are particularly noteworthy.

A few comments were not very positive. One participant (NB) reported that *"Some voices are too slow and the pitch is off"*, and another (W) mentioned that *"many of the voices were monotonous"*. Another one (W) said *"Very bizarre diction should be fixed to make the voices less irritating"*.

Two other participants focused more on the robot-likeness of the voices. One (W) stated that *"all voices sounded very robot-like"*, while another (M) said *"I spotted one that I'm sure was a real human voice, the rest had varying degrees of robot-ness to them"*.

One participant (W) found some of the characteristics used to rate a voice a bit difficult to understand, such as whether a voice is "dependable".

Lastly, some participants focused on the accents: *"Only that the British accent was the least fake sounding"* (W), *"I like the British one the best (maybe because I'm British?)" (M), "I enjoyed listening to the different voices, it would be nice to hear more accents as well"* (NB), and *"only that I was strongly influence by accents"* (W).

5 Discussion

Understanding user perceptions in voice interactions has been a focus point in research on conversational user interfaces, particularly about gender. However, much of the previous work has primarily addressed binary gender categories, often not including the experiences of gender-diverse individuals, especially non-binary users. Moreover, gender-ambiguous and related voice stimuli are often studied in isolation, preventing useful listener comparisons. The results of our analysis reveal several implications related to voice ambiguity, voice type, the experiences of non-binary individuals, priming effects, and voice design. We discuss each in turn below.

5.1 Beyond-gender voice perceptions (RQ1)

Overall, our study revealed that certain voices were preferred over others. Specifically, "Fable" was the most liked voice overall, while "Sam" and "Can_1" were the least liked. Notably, "Fable" was the only voice with a British accent, which may have contributed to its higher ratings. Given that our Prolific participants were primarily from the UK, this suggests a possible in-group preference, or familiarity with the accent [10, 34]. While artificial voices are often designed without explicit consideration of accent, research indicates that accent is a key factor in perceptions of likeability, trustworthiness, and overall person perception [10, 67, 71]. Moreover, even synthetic accents are noticed and can influence listener evaluations [3, 69]. This further supports the idea that beyond gender ambiguity, accent and other sociolinguistic factors play an essential role in voice perception. Thus, ambiguity alone cannot be the only design consideration when designing inclusive voices.

To address our first research question—how participants evaluate ambiguous voices across different voice perception dimensions—the following paragraphs discuss it.

Trustworthiness: It appears that the six ambiguous voices are perceived differently in terms of trust, which is not surprising since individuals have varying preferences for voices [4]. Our findings show that women and men tend to trust ambiguous voices more than non-binary participants. Additionally, in the priming condition, non-binary individuals showed even lower trust in these voices. This overall lower trust among non-binary participants–both in the primed and non-primed conditions–may stem from a greater awareness of voice-related biases or a lack of personal resonance with the voices presented. In the priming condition, where voice ambiguity was explicitly mentioned, this effect may have

Effect	Trustworthiness	Appeal	Comfort	Anthropomorphism	Aversion
Voice	✓	✓	✓	✓	✓
Participant Gender	✓	×	×	✓	✓
Priming	×	×	×	×	×
Gender x Priming	✓	✓	✓	✓	✓

Table 5: Summary of statistically significant (checkmark) or non-statistically significant (x sign) effects found in the analysis.

been amplified, potentially making non-binary participants more self-conscious about the topic and more critical in their evaluations.

Comfort and Appeal: Participants' perceptions of how comforting and appealing the voices were also varied. In terms of comfort, women and men found the voices comfortable, with varying preferences. However, knowing why these ambiguous voices are comforting or appealing is open for future research, since multiple reasons can play a role, including familiarity with the voice's accent, or acoustic properties that might make a voice sound generally "attractive" [31, 78]. On the other hand, non-binary participants who were not primed perceived the ambiguous voices as more uncomfortable compared to the other groups.

Anthropomorphism: Associating anthropomorphism with voice characteristics can be explained by recent research presented in Schreibelmayr and Mara [51], which shows that individuals who perceive a voice as human-like are more likely to consider it anthropomorphic. In our study, women and men rated the six voices as anthropomorphic, with some variations in preferences. On the other hand, non-binary participants showed a clear tendency to perceive ambiguous voices as less anthropomorphic compared to women and men. Relating this to the findings in [51], it might be that non-binary individuals do not perceive these voices to hold human-like characteristics, while women (in both primed and non-primed conditions) and men (especially in the non-primed condition) do. This argument can be supported by the fact that even when non-binary participants were primed about the ambiguity of the voices, the ratings were lower compared to women. This can be further explained by the factors we discussed in the "Trustworthiness" category. The lower ratings could reflect a heightened awareness of voice-related biases or a lack of personal resonance with the voices, as non-binary participants might feel less connection to voices they perceive as ambiguous or not representative of their identity.

Aversion: Generally, the only group that showed higher aversion to the voices was non-binary participants, with an even higher aversion rate among non-primed non-binary participants. One possible explanation is that non-binary individuals perceive ambiguous voices differently [28], suggesting the six ambiguous voices may have been synthesised with specific acoustic features that induced discomfort for non-binary participants, thereby eliciting an aversive response. However, this was not the case for women and men.

5.2 Priming Effects (RQ2)

Our second research question asked whether informing people about the nature of a voice impacts their perceptions. We found that priming does not affect participants overall, as it was never a statistically significant factor in the analyses. However, we found some interactions with participant gender, as it tended to make non-binary participants like the voices even less, which is somewhat contrary to our expectations. For instance, the work of Hope and Lilley [28] would suggest that those who identify as women or gender-diverse would appreciate gender-ambiguous voices over other gendered voices. At the same time, non-binary people may have the most at stake when it comes to gender-ambiguous voices that may potentially be labelled as "non-binary", both in terms of representation (e.g., via self-used avatars, as well as non-self chatbots, robots, and VAs), but also in terms of any potential backlash (e.g., if bigots complain that designers are making "woke" technologies). In contrast, people who have less interest would be less critical and potentially like these voices more. This deserves critical interrogation: echoing Fujii et al. [24], we note that non-binary populations may be at risk of more readily embracing stereotyped or low-quality voice options labelled or perceived as marginalised human identities. As such, we suggest that non-binary and gender-diverse people have the most interest in directing the course of computer voice and TTS development and uptake, which we discuss more in subsection 5.4 and section 6.

5.3 Gender effects (RQ3)

When considering the overall perception of ambiguous voices, no differences were found between men and women. However, non-binary participants generally rated the ambiguous voices more negatively. They provided lower scores for trustworthiness, appeal, comfort, and anthropomorphism and expressed greater aversion toward the ambiguous voices. In the open-ended feedback, non-binary participants contributed fewer comments. Only two participants left feedback (see Section 4.8): one mentioned that some voices were too slow and the pitch was off, while another suggested that more accents would have been preferable.

The lower ratings from non-binary participants could be attributed to several interrelated factors. First, these voices may not have offered a sense of personal connection or recognition, as they might not have adequately represented non-binary identities or lived experiences. This aligns with past research suggesting that representational mismatch can reduce user trust and engagement, especially among marginalised groups [17, 28]. Second, non-binary participants may have been more attuned to subtle design biases and limitations in the voices, evaluating them more critically due to heightened awareness of exclusion or misrepresentation in technology. Notably, this group also rated the voices as less anthropomorphic and appealing, further suggesting a lack of resonance.

The priming condition, which explicitly framed the voices as ambiguous, may have amplified this effect by prompting deeper reflection on the adequacy and authenticity of these voices as inclusive representations. Rather than fostering a sense of belonging, priming may have made some participants more conscious of the

gap between design intent and lived experience. Finally, although the voices were designed to be gender-ambiguous, non-binary participants may have still detected subtle gendered cues—such as pitch, cadence, or vocal affect—that signalled binary gender norms, leading to discomfort or even aversion.

These findings raise important questions for future work. If ambiguity alone fails to engender trust among those it aims to include, designers must reconsider whether it is a sufficient strategy for inclusive voice design. Instead, voice technologies may need to move toward more explicitly non-binary or community-informed representations. As such, this result challenges the field to critically re-evaluate assumptions around neutrality and inclusivity, and centre non-binary users in both design and evaluation processes.

5.4 Limitations and Future Work

We acknowledge some limitations in the study, which could serve as starting points for future research. First, while we recruited a balanced sample of women, men and non-binary participants, the generalisability of our findings may be limited due to the UK's specific cultural and linguistic context. Future research should explore how cultural background influences perceptions of gender-ambiguous voices, as attitudes toward gender and language vary across regions. Furthermore, accents influenced participants' perceptions, as the voice with a British accent was preferred the most, likely due to the cultural context of the UK.

Second, we had little control over the diversity of the stimuli (the artificial voices) since there are currently very few gender-ambiguous voices available. As a result, we were limited to using the small set of voices accessible to us, which may not fully represent the range of possible gender-ambiguous voices. While our study provided more options for comparison than in previous work, we recognise that this remains a limitation. We hope that as the field advances, a wider and more varied selection of artificial voices will become available, enabling more comprehensive investigations into perception differences. We also encourage researchers to collaborate with industry partners, developers in TTS and related technologies, and non-binary communities [17] and perhaps provide open source material for research use.

Our analysis also did not account for potential interactions with other individual differences, such as participants' familiarity with synthetic voices (not only recognition of the voices, but also similarity-to-self, which may [11] or may not [23] affect judgments) or attitudes toward gender diversity (for example, the Openness towards Non-Binary Gender (ONBG) scale [40] or the Transgender Attitudes and Beliefs Scale (TABS) [32]). Incorporating additional measures, such as prior exposure to voice technology or implicit biases, could provide deeper insights into the mechanisms underlying (ambiguous) voice perceptions.

To the best of our knowledge, this was the first work looking at how people perceive different ambiguous voices; given our results, in the next paragraphs, we highlight possible research avenues.

Explore a palette of gender-ambiguous voices: Creating a gender-ambiguous voice that is universally perceived as such remains a challenge [56, 68]. Nevertheless, the community continues to make strides by developing novel gender-ambiguous voices [35, 68], identifying gender-ambiguous (and gender-neutral, potentially non-human) voices [59], and modifying existing ones [36]. We encourage researchers to contribute their ambiguous voices to our repository containing gender-ambiguous voices [12]. We hope this resource will serve as a starting point for fostering the development and open sharing of gender-ambiguous voices within the CUI research community and will support those aiming to better understand their characteristics.

Understand and integrate diverse gender perceptions: To design truly inclusive gender-ambiguous voices, we must deepen our understanding of how individuals from minority gender communities, including non-binary [35], agender [33], X-gender [16], and other third or non-colonial gender identities [18], perceive these voices. Previous work has begun to explore whether such voices are experienced as appealing or alienating [17, 21, 27], but much remains to be done. People's impressions of gendered voices are shaped by deeply ingrained mental models of human social identity [25, 44], making it essential for designers and researchers to actively reflect on and include these diverse identities in both development and evaluation. Doing so can help prevent the unintentional exclusion or erasure of non-binary and non-Western gender experiences and ensure that gender-ambiguous voices foster genuine inclusion rather than reinforcing hidden biases.

Put the voice to work: While our research moves the field forward by shifting from mere perceptions of voice gender to general attitudes, this is only the first step. We must explore the impact of deploying such voices in various contexts, with a diversity of people [19]. At present, we have only described people's first impressions. Observational research, especially with a longitudinal component, experimental controls, and/or rich qualitative methods [61] can reveal the answer to the "so what?" question that currently remains elusive.

Examine the role of cultural context in voice perception: Beyond gender, cultural context may significantly influence how gender-ambiguous voices are received. Accents [34], notably, but also linguistic features and culturally specific gender norms, can shape how voices are interpreted across different communities. Investigating how cultural background impacts the perception of ambiguous voices could offer valuable insights into designing voices that are truly inclusive on a global scale. Future work could explore whether people from different cultural contexts perceive voice ambiguity in distinct ways, helping guide the development of voice technologies that are both culturally sensitive and broadly accepted.

6 Conclusion

Computer voices can be found across a range of media and interactive technologies, from video games to smart homes to virtual assistants. The emergence of human-like artificial voices in combination with advanced Natural Language Processing from generative AI has set the stage for naturalistic voice-based human-agent interaction in the years to come. Our work, building on a small but growing body of work within and beyond HCI and CUI spaces, shows how vital the "flavour" of voice is for different people. Absent of content or engagement, people rely on what they can "hear" in the voice when making judgments about trustworthiness, comfort, appeal, anthropomorphism, and aversion.

In this study, we evaluated six gender-ambiguous voices with participants from a range of gender identities, including men, women, and non-binary individuals. The design of the study involved both a priming condition, where participants were told that the voices were gender-ambiguous, and a control condition, where this information was withheld. This allowed us to examine how participants' perceptions were shaped not only by the voices themselves but also by the context in which they encountered them. Our findings revealed that non-binary participants expressed significantly lower trust and higher aversion toward the voices compared to men and women. This suggests that the current design of gender-ambiguous voices may not fully resonate with non-binary people, challenging the assumption that ambiguity alone fosters inclusivity. Furthermore, the priming condition amplified this effect, making non-binary participants more self-conscious about the voices' representation, likely contributing to their more critical evaluations.

Moreover, accent emerged as a key factor in the likeability and acceptance of these voices, with diverse accents contributing to higher engagement. This highlights the importance of considering factors such as accent and cultural context in the design of voice technologies, as these elements can significantly impact users' perceptions and experiences.

Our research emphasises the need for more inclusive, community-informed design practices in voice technology, particularly when aiming to serve marginalised or under-represented groups. As we move forward in the development of human-agent interactions, these findings highlight the importance of considering a broader spectrum of identity and experiences in designing and evaluating voices for future systems. In doing so, we can ensure that voice technologies are more inclusive, relatable, and trustworthy for all.

References

[1] 2022. GenderLess Voice. https://www.genderlessvoice.com/. Accessed: June 19, 2024.

[2] AY Alhaddad, A Mecheter, MA Wadood, AS Alsaari, H Mohammed, and JJ Cabibihan. 2021. Anthropomorphism and Its Negative Attitudes, Sociability, Animacy, Agency, and Disturbance Requirements for Social Robots: A Pilot Study, H Li, SS Ge, Y Wu, A Wykowska, H He, X Liu, D Li, and J PerezOsorio (Eds.), Vol. 13086. 791–796. https://doi.org/10.1007/978-3-030-90525-5_73

[3] Sean Andrist, Micheline Ziadee, Halim Boukaram, Bilge Mutlu, and Majd Sakr. 2015. Effects of culture on the credibility of robot speech: A comparison between english and arabic. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. 157–164.

[4] Melissa Barkat-Defradas, Michel Raymond, and Alexandre Suire. 2021. *Vocal Preferences in Humans: A Systematic Review*. Springer Singapore, Singapore, 55–80. https://doi.org/10.1007/978-981-15-6627-1_4

[5] Hilary Bergen. 2016. 'I'd blush if I could': Digital assistants, disembodied cyborgs and the problem of gender. *Word and Text, A Journal of Literary Studies and Linguistics* VI, 01 (2016), 95–113. https://www.ceeol.com/search/article-detail?id=469884 Publisher: Universitatea Petrol-Gaze din Ploieşti.

[6] Jasmin Bernotat, Friederike Eyssel, and Janik Sachse. 2019. The (fe)male robot: How robot body shape impacts first impressions and trust towards robots. *International Journal of Social Robotics* 13, 3 (May 2019), 477–489. https://doi.org/10.1007/s12369-019-00562-7

[7] Paul Boersma and David Weenink. 2001. Praat, a system for doing phonetics by computer. http://www.praat.org/.

[8] D Bryant, J Xu, K Rogers, A Howard, and IEEE. 2021. The Effect of Conceptual Embodiment on Human-Robot Trust During a Youth Emotion Classification Task. In *University System of Georgia*. 211–218. https://doi.org/10.1109/ARSO51874.2021.9542839

[9] Yang Trista Cao and Hal Daumé. 2021. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle. *Computational Linguistics* 47, 3 (Nov. 2021), 615–661. https://doi.org/10.1162/coli_a_00413

[10] Aaron Castelan Cargile and Howard Giles. 1997. Understanding language attitudes: Exploring listener affect and identity. *Language & Communication* 17, 3 (1997), 195–217.

[11] Salvador Carmona, Govind Iyer, and Philip M.J. Reckers. 2014. Performance evaluation bias: A comparative study on the role of financial fixation, similarity-to-self and likeability. *Advances in Accounting* 30, 1 (June 2014), 9–17. https://doi.org/10.1016/j.adiac.2014.04.001

[12] Martina De Cet. 2025. Gender-Ambiguous Voices Database. https://github.com/martirator/Gender-Ambiguous-Voices Accessed: 2025-05-23.

[13] Meia Chita-Tegmark, Monika Lohani, and Matthias Scheutz. 2019. Gender Effects in Perceptions of Robots and Humans with Varying Emotional Intelligence. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 230–238. https://doi.org/10.1109/hri.2019.8673222

[14] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (June 2019), 349–371. https://doi.org/10.1093/iwc/iwz016

[15] Charles Crowell, Matthias Scheutz, Paul Schermerhorn, and Michael Villano. 2009. Gendered voice and robot entities: perceptions and reactions of male and female subjects. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE Press, New York, NY, USA, 3735–3741.

[16] Sonja P.F. Dale. 2012. An introduction to X-Jendā: Examining a new gender identity in Japan. *Intersections: Gender and sexuality in Asia and the Pacific* 31 (2012).

[17] Andreea Danielescu, Sharone A Horowit-Hendler, Alexandria Pabst, Kenneth Michael Stewart, Eric M Gallo, and Matthew Peter Aylett. 2023. Creating Inclusive Voices for the 21st Century: A Non-Binary Text-to-Speech for Conversational Assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 390, 17 pages. https://doi.org/10.1145/3544548.3581281

[18] Jenny L. Davis. 2019. Refusing (Mis)Recognition: Navigating Multiple Marginalization in the U.S. Two Spirit Movement. *Review of International American Studies* 12, 1 (Jan. 2019), 65–86. https://doi.org/10.31261/rias.7328

[19] Martina De Cet, Mohammad Obaid, and Ilaria Torre. 2025. Breaking the Binary: A Systematic Review of Gender-Ambiguous Voices in Human-Computer Interaction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 997, 17 pages. https://doi.org/10.1145/3706598.3713608

[20] Martina De Cet, Miriam Sturdee, Mohammad Obaid, and Ilaria Torre. 2025. Sketching Robots: Exploring the Influence of Gender-Ambiguous Voices on Robot Perception. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction* (Melbourne, Australia) *(HRI '25)*. IEEE Press, 103–112.

[21] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.150

[22] Friederike Eyssel and Frank Hegel. 2012. (S)he's Got the Look: Gender Stereotyping of Robots 1. *Journal of Applied Social Psychology* 42 (07 2012). https://doi.org/10.1111/j.1559-1816.2012.00937.x

[23] Takao Fujii, Katie Seaborn, and Madeleine Steeds. 2024. Silver-tongued and sundry: Exploring intersectional pronouns with ChatGPT. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article No. 511. https://doi.org/10.1145/3613904.3642303

[24] Takao Fujii, Katie Seaborn, Madeleine Steeds, and Jun Kato. 2025. Inter(sectional) Alia(s): Ambiguity in Voice Agent Identity via Intersectional Japanese Self-Referents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM.

[25] Andrew Gambino, Jesse Fox, and Rabindra A. Ratan. 2020. Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication* 1 (2020), 71–85.

[26] Koen V. Hindriks, Marijn Hagenaar, and Anna Laura Huckelba. 2022. Effects of Robot Clothing on First Impressions, Gender, Human-Likeness, and Suitability of a Robot for Occupations. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 428–435. https://doi.org/10.1109/ro-man53752.2022.9900771

[27] Maxwell Hope and Jason Lilley. 2022. Gender expansive listeners utilize a non-binary, multidimensional conception of gender to inform voice gender perception. *Brain and Language* 224 (Jan. 2022), 105049. https://doi.org/10.1016/j.bandl.2021.105049

[28] Maxwell Hope and Jason Lilley. 2023. Differences in sibilant perception between gender expansive and cisgender individuals. In *Seminars in Speech and Language*, Vol. 44. Thieme Medical Publishers, Inc., 061–075.

[29] Janet Shibley Hyde, Rebecca S. Bigler, Daphna Joel, Charlotte Chucky Tate, and Sari M. van Anders. 2019. The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist* 74, 2 (2019), 171–193.

https://doi.org/10.1037/amp0000307 Place: US Publisher: American Psychological Association.

[30] J. Jackson and R. D. A. Taylor. 2019. Vocal pitch and intonation characteristics of those who are gender non-binary. In *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS 2019)*. International Phonetic Association, London, UK. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_2734.pdf

[31] Patrik N Juslin, Klaus R Scherer, and J Harrigan. 2005. Vocal expression of affect. *The new handbook of methods in nonverbal behavior research* (2005), 65–135.

[32] Yasuko Kanamori, Jeffrey H. D. Cornelius-White, Teresa K. Pegors, Todd Daniel, and Joseph Hulgus. 2016. Development and Validation of the Transgender Attitudes and Beliefs Scale. *Archives of Sexual Behavior* 46, 5 (Aug. 2016), 1503–1515. https://doi.org/10.1007/s10508-016-0840-1

[33] Morgan Ketola, Schyler Selander, and Ayalla Ruvio. 2022. Identity expressions of agender individuals in a digital world. In *The Routledge Handbook of Digital Consumption*. Routledge, New York, 425–435.

[34] Katherine D. Kinzler, Kristin Shutts, Jasmine DeJesus, and Elizabeth S. Spelke. 2009. Accent trumps race in guiding children's social preferences. *Social cognition* 27, 4 (2009), 623–634.

[35] Sonja Theresa Längle, Stephan Schlögl, Annina Ecker, Willemijn S. M. T. van Kooten, and Teresa Spieß. 2024. Nonbinary Voices for Digital Assistants—An Investigation of User Perceptions and Gender Stereotypes. *Robotics* 13, 8 (2024). https://doi.org/10.3390/robotics13080111

[36] Yuto Mandai, Katie Seaborn, Tomoyasu Nakano, Xin Sun, Yijia Wang, and Jun Kato. 2025. Super Kawaii Vocalics: Amplifying the "Cute" Factor in Computer Voice. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM.

[37] Konstantinos Markopoulos, Georgia Maniati, Georgios Vamvoukakis, Nikolaos Ellinas, Georgios Vardaxoglou, Panos Kakoulidis, Junkwang Oh, Gunu Jho, Inchul Hwang, Aimilios Chalamandaris, Pirros Tsiakoulis, and Spyros Raptis. 2023. Generating Multilingual Gender-Ambiguous Text-to-Speech Voices. arXiv:2211.00375 [cs.SD] https://arxiv.org/abs/2211.00375

[38] Carol Lynn Martin and Diane Ruble. 2004. Children's Search for Gender Cues: Cognitive Perspectives on Gender Development. *Current Directions in Psychological Science* 13, 2 (April 2004), 67–70. https://doi.org/10.1111/j.0963-7214.2004.00276.x

[39] Phil McAleer, Alexander Todorov, and Pascal Belin. 2014. How do you say 'Hello'? Personality impressions from brief novel voices. *PloS one* 9, 3 (2014), e90779.

[40] Améthyste Molin, Amélie Simond, Sayaka Sato, Tiziana Jaeggi, Pascal Mark Gygax, and Nathalie Meuwly. 2020. Linking political and feminist ideology with openness towards non-binary gender: The development and initial validation of a scale to measure subjective Openness towards Non-Binary Gender (ONBG). *Journal of Gender Studies* 30, 8 (Nov. 2020), 901–914. https://doi.org/10.1080/09589236.2020.1844644

[41] Sandra Mooshammer and Katrin Etzrodt. 2022. *Social Research with Gender-Neutral Voices in Chatbots – The Generation and Evaluation of Artificial Gender-Neutral Voices with Praat and Google WaveNet.* 176–191. https://doi.org/10.1007/978-3-030-94890-0_11

[42] Sandra Mooshammer and Katrin Etzrodt. 2022. *Social Research with Gender-Neutral Voices in Chatbots – The Generation and Evaluation of Artificial Gender-Neutral Voices with Praat and Google WaveNet.* Springer International Publishing, 176–191. https://doi.org/10.1007/978-3-030-94890-0_11

[43] John W. Mullennix, Keith A. Johnson, Meral Topcu-Durgun, and Lynn M. Farnsworth. 1995. The perceptual representation of voice gender. *The Journal of the Acoustical Society of America* 98, 6 (Dec. 1995), 3080–3095. https://doi.org/10.1121/1.413832

[44] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI94)*. ACM, 72–78. https://doi.org/10.1145/191666.191703

[45] Giulia Perugia, Stefano Guidi, Margherita Bicchi, and Oronzo Parlangeli. 2022. The shape of our bias: Perceived age and gender in the humanoid robots of the ABOT database. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI '22)*. IEEE Press, Sapporo, Hokkaido, Japan, 110–119.

[46] Giulia Perugia and Dominika Lisy. 2023. Robot's Gendering Trouble: A Scoping Review of Gendering Humanoid Robots and Its Effects on HRI. *International Journal of Social Robotics* 15, 11 (Nov. 2023), 1725–1753. https://doi.org/10.1007/s12369-023-01061-6

[47] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, 1–12. https://doi.org/10.1145/3173574.3174214

[48] Pranav Prabha, Sreejith Sasidharan, Devasena Pasupuleti, Anand Das, Gayathri Manikutty, and Rajesh Sharma. 2022. A Minimalist Social Robot Platform for Promoting Positive Behavior Change Among Children. In *ACM SIGGRAPH 2022 Educator's Forum* (Vancouver, BC, Canada) *(SIGGRAPH '22)*. Association for Computing Machinery, New York, NY, USA, Article 3, 2 pages. https://doi.org/10.1145/3532724.3535597

[49] Pranav Prabha, Sreejith Sasidharan, Devasena Pasupuleti, Anand Das, Gayathri Manikutty, and Rajesh Sharma. 2022. A Minimalist Social Robot Platform for Promoting Positive Behavior Change Among Children. In *ACM SIGGRAPH 2022 Educator's Forum* (Vancouver, BC, Canada) *(SIGGRAPH '22)*. Association for Computing Machinery, New York, NY, USA, Article 3, 2 pages. https://doi.org/10.1145/3532724.3535597

[50] Kantwon Rogers, De'Aira Bryant, and Ayanna Howard. 2020. Robot gendering: Influences on trust, occupational competency, and preference of robot over human. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. ACM, Honolulu, HI, 1–7. https://doi.org/10.1145/3334480.3382930

[51] Simon Schreibelmayr and Martina Mara. 2022. Robot Voices in Daily Life: Vocal Human-Likeness and Application Context as Determinants of User Acceptance. *Frontiers in Psychology* 13 (2022). https://doi.org/10.3389/fpsyg.2022.787499

[52] Katie Seaborn, Shruti Chandra, and Thibault Fabre. 2023. Transcending the "male code": Implicit masculine biases in NLP contexts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3544548.3581017 Article No.: 138.

[53] Katie Seaborn and Alexa Frank. 2022. What pronouns for Pepper? A critical review of gender/ing in research. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New Orleans, Louisiana, USA, 1–15. https://doi.org/10.1145/3491102.3501996

[54] Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human–Agent Interaction: A Survey. *Comput. Surveys* 54, 4 (May 2021), 1–43. https://doi.org/10.1145/3386867

[55] Katie Seaborn, Somang Nam, Julia Keckeis, and Tatsuya Itagaki. 2023. Can voice assistants sound cute? Towards a model of kawaii vocalics. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3544549.3585656 Article No.: 63.

[56] Katie Seaborn and Peter Pennefather. 2022. Neither "Hear" Nor "Their": Interrogating Gender Neutrality in Robots. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1030–1034. https://doi.org/10.1109/hri53351.2022.9889350

[57] Katie Seaborn, Peter Pennefather, and Haruki Kotani. 2022. Exploring gender-expansive categorization options for robots. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. ACM, New Orleans, LA, 1–6. https://doi.org/10.1145/3491101.3519646

[58] Katie Seaborn, Katja Rogers, Maximilian Altmeyer, Mizuki Watanabe, Yuto Sawa, Somang Nam, Tatsuya Itagaki, and Ge 'Rikaku' Li. 2025. Unboxing Manipulation Checks for Voice UX. *Interacting with Computers* (Jan. 2025), iwae062. https://doi.org/10.1093/iwc/iwae062

[59] Katie Seaborn, Katja Rogers, Somang Nam, and Miu Kojima. 2023. Kawaii Game Vocalics: A Preliminary Model. In *Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '23)*. ACM, 202–208. https://doi.org/10.1145/3573382.3616099

[60] Katie Seaborn and Jacqueline Urakami. 2021. Measuring Voice UX Quantitatively: A Rapid Review. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3411763.3451712

[61] Katie Seaborn, Jaqueline Urakami, Peter Pennefather, and Norihisa Miyake. 2024. Qualitative Approaches to Voice UX. *Comput. Surveys* 56, 12 (Oct. 2024), 1–34. https://doi.org/10.1145/3658666

[62] Verena G. Skuk and Stefan R. Schweinberger. 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 (Feb. 2014), 285–296. https://doi.org/10.1044/1092-4388(2013/12-0314)

[63] SpeechGen.io. [n. d.]. Gender-Neutral Voice Generator. https://speechgen.io/en/neutral/. Accessed: 2025-02-19.

[64] Katta Spiel, Os Keyes, and Pınar Barlas. 2019. Patching Gender: Non-binary Utopias in HCI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 1–11. https://doi.org/10.1145/3290607.3310425

[65] Selina Jeanne Sutton. 2020. Gender Ambiguous, not Genderless: Designing Gender in Voice User Interfaces (VUIs) with Sensitivity. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 11, 8 pages. https://doi.org/10.1145/3405755.3406123

[66] Suzanne Tolmeijer, Naim Zierau, Andreas Janson, Jalil Sebastian Wahdatehagh, Jan Marco Marco Leimeister, and Abraham Bernstein. 2021. Female by Default? – Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, 1–7. https://doi.org/10.1145/3411763.3451623

[67] Ilaria Torre, Jeremy Goslin, Laurence White, and Debora Zanatto. 2018. Trust in artificial voices: A" congruency effect" of first impressions and behavioural experience. In *Proceedings of the technology, mind, and society*. 1–6.

[68] Ilaria Torre, Erik Lagerstedt, Nathaniel Dennler, Katie Seaborn, Iolanda Leite, and Éva Székely. 2023. Can a gender-ambiguous voice reduce gender stereotypes in human-robot interactions?. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE Press, New York, NY, USA, 106–112. https://doi.org/10.1109/RO-MAN57019.2023.10309500

[69] Ilaria Torre and Sébastien Le Maguer. 2020. Should robots have accents?. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 208–214.

[70] Ilaria Torre and Laurence White. 2020. *Trust in Vocal Human–Robot Interaction: Implications for Robot Voice Design*. Springer Singapore, 299–316. https://doi.org/10.1007/978-981-15-6627-1_16

[71] Ilaria Torre, Laurence White, Jeremy Goslin, and Sarah Knight. 2024. The irrepressible influence of vocal stereotypes on trust. *Quarterly Journal of Experimental Psychology* 77, 10 (2024), 1957–1966.

[72] Gabriele Trovato, Cesar Lucho, and Renato Paredes. 2018. She's Electric—The Influence of Body Proportions on Perceived Gender of Robots across Cultures. *Robotics* 7, 3 (Sept. 2018), 50. https://doi.org/10.3390/robotics7030050

[73] Takane Ueno, Yeongdae Kim, Hiroki Oura, and Katie Seaborn. 2023. Trust and reliance in consensus-based explanations from an anti-misinformation agent. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3544549.3585713 Article No.: 296.

[74] Jacqueline Urakami and Katie Seaborn. 2023. Nonverbal Cues in Human–Robot Interaction: A Communication Studies Perspective. *ACM Transactions on Human-Robot Interaction* 12, 2 (March 2023), 1–21. https://doi.org/10.1145/3570169

[75] M. L. Walters, D. S. Syrdal, K. L. Koay, K. Dautenhahn, and R. te Boekhorst. 2008. Human approach distances to a mechanical-looking robot with different robot voice styles. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 707–712. https://doi.org/10.1109/roman.2008.4600750

[76] A Waytz, J Heafner, and N Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *JOURNAL OF EXPERIMENTAL SOCIAL PSYCHOLOGY* 52 (May 2014), 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

[77] Melanie Weirich and Adrian P. Simpson. 2018. Gender identity is indexed and perceived in speech. *PLOS ONE* 13, 12 (Dec. 2018), e0209226. https://doi.org/10.1371/journal.pone.0209226

[78] Benjamin Weiss, Jürgen Trouvain, Melissa Barkat-Defradas, and John J Ohala. 2021. Voice attractiveness. *Encyclopedia of evolutionary psychological science* (2021), 8461–8461.

[79] Chuang Yu, Changzeng Fu, Rui Chen, and Adriana Tapus. 2022. First Attempt of Gender-free Speech Style Transfer for Genderless Robot. https://doi.org/10.1109/HRI53351.2022.9889533

[80] Éva Székely, Joakim Gustafson, and Ilaria Torre. 2023. Prosody-controllable Gender-ambiguous Speech Synthesis: A Tool for Investigating Implicit Bias in Speech Perception. In *Proc. INTERSPEECH 2023*. 1234–1238. https://doi.org/10.21437/Interspeech.2023-2086