Female by Default? – Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution

Suzanne Tolmeijer
tolmeijer@ifi.uzh.ch
Department of Informatics, University of Zurich
Zurich, Switzerland

Naim Zierau
naim.zierau@unisg.ch
IWI-HSG, University of St. Gallen
St. Gallen, Switzerland

Andreas Janson
andreas.janson@unisg.ch
IWI-HSG, University of St. Gallen
St. Gallen, Switzerland

Jalil Wahdatehagh
jalil@1001.digital
1001.digital
Valley, Germany

Jan Marco Leimeister
janmarco.leimeister@unisg.ch
IWI-HSG, University of St. Gallen
St. Gallen, Switzerland

Abraham Bernstein
bernstein@ifi.uzh.ch
Department of Informatics, University of Zurich
Zurich, Switzerland

ABSTRACT

Gendered voice based on pitch is a prevalent design element in many contemporary Voice Assistants (VAs) but has shown to strengthen harmful stereotypes. Interestingly, there is a dearth of research that systematically analyses user perceptions of different voice genders in VAs. This study investigates gender-stereotyping across two different tasks by analyzing the influence of pitch (low, high) and gender (women, men) on stereotypical trait ascription and trust formation in an exploratory online experiment with 234 participants. Additionally, we deploy a gender-ambiguous voice to compare against gendered voices. Our findings indicate that implicit stereotyping occurs for VAs. Moreover, we can show that there are no significant differences in trust formed towards a gender-ambiguous voice versus gendered voices, which highlights their potential for commercial usage.

CCS CONCEPTS

• **Human-centered computing → User interface design**; **Empirical studies in HCI**; **Sound-based input / output**; Interaction design theory, concepts and paradigms; • **Social and professional topics** → *Gender*.

KEYWORDS

Voice Assistants, Gender Stereotypes, Voice Design, Trust, Gender-Ambiguous Voice

ACM Reference Format:
Suzanne Tolmeijer, Naim Zierau, Andreas Janson, Jalil Wahdatehagh, Jan Marco Leimeister, and Abraham Bernstein. 2021. Female by Default? – Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3411763.3451623

CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8095-9/21/05…$15.00
https://doi.org/10.1145/3411763.3451623

1 INTRODUCTION

Voice Assistants (VA), such as Google Assistant or Amazon Alexa, promise to change the ways people perform tasks, use services, and interact with organizations. The interactions of many users with these agents, however, have yielded mixed results, indicating high failure rates [8]. Hence, there has been a growing interest in voice-based interactions in both research and practice [25]. Besides the content of the interaction itself (i.e., 'what is said?'), an element that is central to interaction design of VAs is the voice (i.e., 'how it is said?') [27]. In this regard, a prevalent trend is the use of female over male voices, as companies cite anecdotal evidence which suggests that female voices are favored by most users. Thus, most leading VAs are exclusively female or female by default [18]. In fact, according to a recent study, 77% of all virtual assistants manifested gender-specific cues that can be classified as feminine [7]. However, a recent report by the UNESCO stresses that the gendered design of most VAs could solidify harmful gender stereotypes [29]. For instance, since people become used to interacting with those agents in a commanding tone, humans might also (subconsciously) mirror this behavior in their everyday conversations with women [4]. One potential solution to this issue may lie in the use of gender-ambiguous[1] voices [29]. Gender-ambiguous voice assistants may not only help to combat hurtful gender stereotypes, but also provide more inclusive design tools to represent voices outside the binary gender identities. However, while studies on interaction design with VAs are growing (e.g., [15, 22, 23, 32]), there is a lack of empirical insights on the perceptual effects of gendered (and gender-ambiguous) voices based on para-lingual cues such as pitch. Especially, to the best of our knowledge, no study has empirically tested user perceptions and the technical feasibility of deploying gender-ambiguous voices for VA design.

To address this shortcoming, we conducted an exploratory study to empirically analyze the effects of (ambiguously) gendered voices

[1] In accordance with Sutton [27], we use the term 'gender-ambiguous' throughout this paper rather than calling a voice 'genderless': many cues in the sound and content of VA speech can illicit gender ascription, even when the pitch is gender-neutral.

on trait and trust attribution across different task contexts. Specifically, we comparatively analyze user perceptions in regards to pitch (low, high) and gender (female, male) as well as a gender-ambiguous voice we constructed. According to literature, the pitch of the voice is one of the most important factors regarding the attribution of gender [19]. To that end, we developed a voice interface for online experiments. On this basis, we implemented two task scenarios: one where users were asked to book a flight with a VA (assistance scenario) and one where users were surveyed by a VA on their financial situation (compliance scenario). We conducted a 5x2 online experiment with 234 participants on Prolific: five voices (male-low, male-high, gender-ambiguous, female-low, and female-high) were set against two task settings (assistance and compliance). Our results show implicit stereotype activation with regards to (lack of) trait attribution towards the different VA voices. Task context and gender of the participant both have an effect on perceived traits and reported trust. Finally, our study gives a first indication that a gender-ambiguous voice for VAs could be a viable alternative to gendered voices and warrants further investigation.

2 RELATED WORK

Our research is motivated by sociophonetics and social response theory [17]. Every person has a unique voice based on a complex interplay of anatomical and psychological traits and emotional states that together determine how people express themselves verbally and in turn how they are perceived by others [10, 28]. Sociophonetics explores how different speech patterns vary across social categories and the associated socio-cultural assumptions they carry. It is well established that people make inferences on others based on the sound of their voice [19]. Voice carries para-lingual cues that allow people to make assumptions about a person's background and, based on this, to apply social stereotypes. Speakers use subtle para-lingual cues, mostly unconsciously, to induce certain images to listeners [28]. Those cues can be seen as a flexible resource that people (and VAs) can use to signal different social traits and attitudes [28]. Sometimes, voice informs stereotypes about how specific groups of people speak. One obvious group is the gender of the speaker.The most prominent gender-dependent feature of voice is the pitch of a voice. The longer and thicker vocal chords of men produce a lower pitch than woman; a distinction that is easily perceived by listeners [19].

Based on the Computers As Social Actors (CASA) paradigm [17], initial research suggests that when applied to technology, gender-specific voice characteristics may evoke stereotypical trait inferences [13]. While this is not always consciously, it is shown to be the case on a subconscious level [14]. For instance, Pak et al. [21] showed that users apply gender stereotypes when ascribing the trustworthiness of a virtual agent (i.e., the authors found that users trust a male more than to a female virtual doctor). In VAs, we find similar results. Initial findings suggest that people find it easier to process stereotypical voices, i.e., a warm gentle female voice and an assertive, forceful male voice [26]. Specifically, it was shown that the machine's synthetic voice pitch can activate gender stereotyping of users. For instance, Nass et al. demonstrated that participants not only attributed gender towards computers that communicated in a low- versus a high-pitched synthetic computer voice. They also showed that the low versus high pitch of the synthetic voice triggered users to apply gender-schematic judgments of the "male" versus the "female" computer [16]. More recently, Yu et al. found in their study that participants were more likely to disclose personal information to a (lower-pitched) male voice than a (higher-pitched) female voice of a virtual assistant [31]. However, research that systematically analyzes trait and trust attribution based on different voice genders and pitches for VAs is scarce, despite its paramount role in VA design.

Another limitation to our understanding of voice pitch perception is that only male and female VA voices have been explored, despite calls to research a gender-ambiguous voice [28, 29]. There is very little literature available on a gender-neutral voice pitch, except for references to 'Q', a voice that was recently created to be used for VAs to circumvent stereotyping [24]. The creators of 'Q' mention the fundamental frequency should be between 145 and 175Hz for the voice to sound gender neutral. However, they indicate that gender is more than just pitch: tone and harmonics (e.g., the sound of vowels) also influence gender perception [5]. As the term gender-ambiguous indicates, voice cannot be regarded as binary [1, 27]. A brain activity study done by Junger et al. found that people have an increased brain response to gender-ambiguous voices and opposite gendered voices cause stronger activation in the fronto-temporal neural network [11]. While the difference in neural perception is shown, the difference in user perception for VAs has not been investigated. The use of gender-ambiguous voices, if proven not to have a negative impact on user trust and experience, can be a viable alternative to gendered voices to create a more inclusive environment for non-binary voices.

3 METHOD

In order to investigate the perceptual effects of voice pitch, we conducted a 5x2 between subjects experiment that manipulated i) the VA's voice gender based on pitch and ii) the task context the VA was deployed in. Dependent variable measures included trait ascription and reported trust in the VA. Specifically, in this exploratory study, we investigated the following research questions: RQ1: How does voice gender based on pitch affect trait ascription? RQ2: How does voice gender based on pitch affect user trust? RQ3: How does the task context influence the way how voice gender based on pitch affect trust and trait ascription?

3.1 Experimental Platform and Voice Design

The experiment was executed in a custom developed online voice assistant interface. By keeping the interface constant and as clean as possible, the focus remains on the voice of the voice assistant (see Figure 1), which allows to investigate trait attribution based on voice characteristics. The interaction with the VA follows a simple turn-taking mechanism, where the VA guides the unfolding conversation with the user. After each utterance of the VA, the button 'record' appears to send the user's response to the server. To control for diversity in the conversation, VA responses were prerecorded and the conversation path was delimited to focus on the task at hand.

The prerecorded answers of the VA use state of the art in text-to-speech generation to produce our voice responses: Google WaveNet

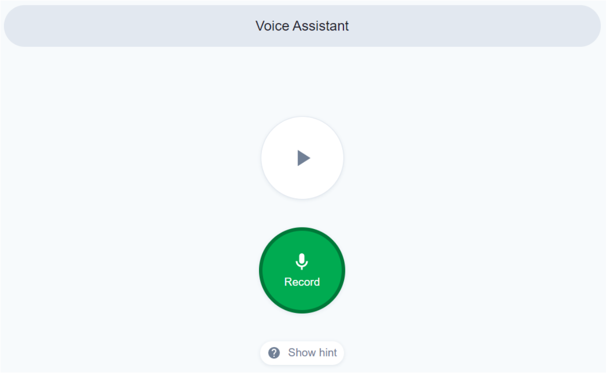


Figure 1: Online Voice Assistant Interface

[20]. To account for both gender and pitch differences, five American English voices are selected: a high- and low-pitched female voice (based on voice `en-US-Wavenet-F`), a high- and low-pitched male voice (based on voice `en-US-Wavenet-B`), and a gender ambiguous voice (based on voice `en-US-Wavenet-E`). While gendered voice generators are readily available, there is not yet a gender-ambiguous text-to-speech generator available. Google's text to speech generator has it listed as an option that is not yet supported.[2] The only available gender-ambiguous generated voice is a carefully crafted voice clip called 'Q', created to fight gender stereotypes in voice assistants [24]. But 'Q' offers no text-to-speech generation. In order to create a voice closest to gender-ambiguous, we pretest male voices with their pitch shifted up, and female voices with the pitch shifted down to identify a voice that classifies as gender-ambiguous. In this regard, gender-ambiguous refers to a voice that falls into both spectrums, meaning that different people would assign different genders to it based on prior mental models. Research on third gender associations has shown that typically people assign a gender to a voice, even though they cannot intuitively assign a gender [27]. To account for this tendency, we included a survey measure asking respondents to identify the gender of the voice assistant through three categories: (1) female; (2) male; and (3) unsure. We used this as a control measure in our models. Eleven manipulated voices based on different Google WaveNet voices were pretested by 52 participants on Prolific (47% female, average age 45 and ranging from 27 to 74). The voice receiving the highest division between assigned gender (58% male and 42% female) was voice `en-US-Wavenet-F` shifted down by three semitones. The selected voices can be found in Table 1.

3.2 Experiment Procedure

The experiment consisted of three phases: 1) randomization, 2) experimental task, and 3) post-test. Randomization and post-test were constant for all groups. Two different experimental task types were used: an assistant and a compliance task. These tasks are inspired by classical gender stereotypes: women are considered to be better in an assistant role, while men are more likely to be seen as leaders [9, 12]. Additionally, they are realistic VA tasks, as both customer surveys [30] and assistant tasks [18] are currently used in VAs. The assistance task involves booking a flight. The participant is given details for a specific flight they want to book and the VA will ask them questions to find and book the right flight for them. The compliance task focuses on personal questions asked in the context of a customer survey. People are asked to answer the questions, but are told it is possible to skip the answer if they prefer not to answer. An example of task interactions can be found in Figure 2.

3.3 Participants

Participants were approached on crowdsourcing platform Prolific.[3] While complying with academic and Prolific's standards on data collection, we set the following preconditions: 1) US nationality, 2) 75%+ approval rate, 3) 10+ previous submissions, and 4) not in pretest sample. Requirement 1) was implemented to control for a language/culture barrier, as the selected voices are speaking in US English. Requirement 2) and 3) were applied to have some quality control in our sample. Requirement 4) excluding priming or bias stemming from the pretest. Initially, 345 people participated. We excluded participants who did not complete the entire task or failed the attention test. After data cleaning, we were left with 234 participants (96 male). The average age was 33 years old, ranging from 19 to 74.

3.4 Measurements and Analysis

The assignment of traits was measured by asking participants about the presence of 24 traits of the VA, based on male and female stereotypes [2, 6]. Each trait was enquired using a 5-point Likert scale, ranging from a positive trait ascription (i.e., 5 indicates 'strongly agree' that the VA had this trait), to negative trait ascription (i.e., 1 indicates 'strongly disagree' that the VA had this trait). Female traits were averaged to indicate female stereotype activation ($\alpha = 0.91$), the mean of male traits was used to indicate male stereotypes ($\alpha = 0.87$). Perceived trust was measured using a validated questionnaire about the perceived competence, benevolence, and integrity of the VA [3] ($\alpha = 0.93$).

4 RESULTS

Trait ascription scores were not normally distributed: a Shapiro-Wilkin test resulted in p < 0.001 for all twenty-four traits. This is possibly because of the nature of the Likert scale for trait ascription: '1' indicates 'strongly disagree' that the VA has this trait, '3' shows the participants 'neither agrees nor disagrees', while '5' reflects a 'strong agreement' that the VA has this trait. To test whether a trait was significantly assigned in a positive way (i.e., significantly higher than a neutral answer of '3') or a negative way (i.e., significantly lower than a neutral answer of '3'), we used the non-parametric paired sample Wilcoxon Signed-Rank test to compare our sample against the neutral value '3'.

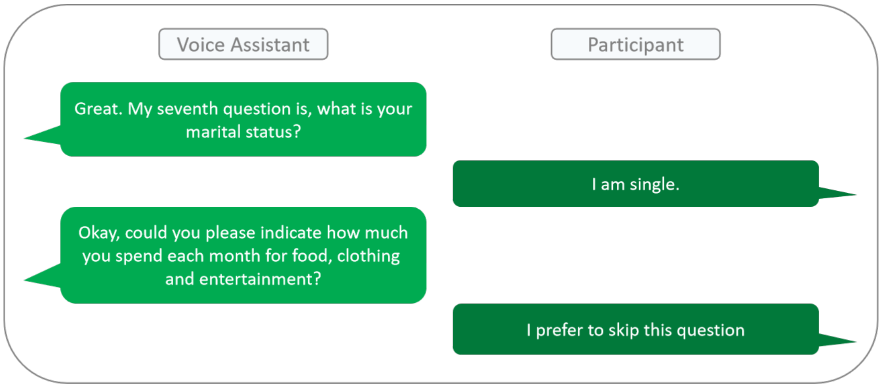
Trust scores were not normally distributed either when comparing different voices: a Shapiro-Wilkin test showed the female high voice data was not normally distributed ($W = 0.95, p = 0.04$). As such, a Kruskal-Wallis test was used rather than an ANOVA test. In the case of two-group comparisons, Mann-Whitney U tests were

[2] When last checked by the authors on December 11th 2020. https://cloud.google.com/text-to-speech/docs/reference/rest/v1/SsmlVoiceGender

[3] https://www.prolific.co/

Table 1: Original Google English US Wavenet voices are shifted by amount of semitone, either using Google's text to speech API (TTS) or by using online generator https://onlinetonegenerator.com/ (gen).

Voice type	Female high (FH)	Female low (FL)	Gender-ambiguous (A)	Male high (MH)	Male low (ML)
Original English US Wavenet voice	F	F	E	B	B
Method	TTS	TTS	Gen	TTS	TTS
Semitones pitch shift	+2	-6	-3	+2	-6
Average pitch	235 Hz	150 Hz	141 Hz	162 Hz	106 Hz



Great. My seventh question is, what is your marital status?

I am single.

Okay, could you please indicate how much you spend each month for food, clothing and entertainment?

I prefer to skip this question

Figure 2: Example excerpt of the compliance task

executed. The results of all tests can be found in the remainder of this section.

4.1 Trait Ascription

While we found no significant activation of combined average male and female traits, results did show significant negative trait ascription. Specifically, over both task types, participants indicated that on average, some VAs did not have male and female traits. When taking the average over all stereotypically male traits, only the male low voice was not negatively marked as stereotypical male ($Z = 341, p = 0.175$). All other voices we significantly negatively associated with a male stereotype (MH: $Z = 268, p = 0.006$; FL: $Z = 461, p = 0.029$; FH: $Z = 198, p < 0.001$; A: $Z = 186, p = 0.006$). Female traits were only negatively assigned to low voices: the male low voice ($Z = 198, p = 0.006$) and female low voice ($Z = 532, p = 0.034$) were not considered to have stereotypical female traits. Other voices did not have negative stereotype ascription (MH: $Z = 456, p = 0.176$, A: $Z = 378, p = 0.543$, FH: $Z = 743, p = 0.903$). The gender of the participant did not influence negative stereotype assignment. The only voice that came near to activating a perceived stereotype was the female high voice: it was almost significant for activating a female stereotype (Z = 743, p = 0.096).

Additionally, we tested for group differences with regards to the individual perceived traits of the VA voices. Again, we added gender as a co-variate, to control for gender-specific differences in individual trait attribution. All voices were experienced to be organised, confident, cooperative, and polite. While low voices were overall considered to be determined (ML: $Z = 257, p = 0.039$; FL: $Z = 759, p = 0.034$), only the low male voice was not experienced as friendly ($Z = 232, p = 0.250$). Curiously, a participant gender difference occurred in trait ascription to the gender-ambiguous voice: while all participants thought the voice was friendly and polite, women rated the ambiguous voice as significantly more friendly than men ($Mdn.women : 5, Mdn.men : 4, U = 84.5, p = 0.037$) and polite ($Mdn.women : 5, Mdn.men : 4, U = 65.5, p = 0.006$) than male participants. Two traits had no significant assignment of any kind for any voice pitch: assertive and affable.Interestingly, for many traits, the trait assignment was negative: people responded significant in the (strongly) disagree category. All voices were not considered to be aggressive, hard-hearted, tough, affectionate, sentimental, or romantic. However, implicit stereotype activation can be found in lack of negative trait ascription. For example, the low male voice was the only voice that was not considered *not* to be authoritative ($Z = 230, p = 0.356$) or dominant ($Z = 136, p = 0.057$). The female high voice, together with the gender-ambiguous voice, were the only voices not negatively assigned typical female traits such as delicate, family oriented, or sensitive. Difference in participant gender is more clear in negative trait ascription, as women assign lower values than men in many cases.

A summary of trait ascription can be found in Table 2, which shows trait ascription scores for all traits that were not uniformly assigned across all voices.

4.2 Trust

Our results reveal no significant differences between the conditions when comparing reported trust in the VAs ($\chi^2(4, 234) = 1.9958, p = 0.736$). Average trust scores were comparable at 4.525 (ML), 4.480 (MH), 4.710 (FL), 4.636 (FH), and 4.824 (A). However, this does show that the gender ambiguous voice is not trusted less than

Table 2: Selected average trait ascription scores per voice pitch. '1' implies strongly disagree the VA has this trait, '3' indicates the trait is not assigned, '5' shows a strong agreement for VA having this trait. Significant differences from lack of trait ascription, test by Wilcoxon Signed-Rank test, are shows as follows: $^{*}p \leqslant 0.05$, $^{**}p \leqslant 0.01$, $^{***}p \leqslant 0.001$. The five individual voices are male-low (ML),male-high (MH), female-low (FL), female-high (FH), and gender-ambiguous (A) respectively.

	ML	MH	A	FL	FH		ML	MH	A	FL	FH
Authoritative	2.95	**2.60***	**2.18*****	**2.77****	**2.51****	**Empathetic**	**2.19*****	2.74	2.72	**2.44*****	2.90
Speaks their mind	2.76	**2.49****	2.74	**2.46****	**2.43****	**Delicate**	**2.29*****	**2.47*****	2.87	**2.61****	3.12
Determined	**3.21***	3.23	3.31	**3.37***	3.25	**Friendly**	3.15	**3.64*****	**3.59***	**3.68*****	**3.98*****
Cold	2.98	2.77	**2.49***	**2.72***	**2.27*****	**Sincere**	2.98	3.13	**3.49***	3.14	**3.47***
Dominant	2.73	**2.51****	**2.13*****	**2.32*****	**2.16*****	**Family-oriented**	**2.54***	**2.66***	2.87	**2.63****	2.76
Leadership skills	**2.51****	2.74	2.69	**2.53****	**2.69***	**Sensitive**	**2.49****	**2.57****	2.79	**2.46*****	2.84

gendered voices. Moreover, there is a significant difference in trust scores reported by male and female participants: female participants trust the gender-ambiguous voice more than men ($Mdn.women$: $5.45, Mdn.men : 4.595, U = 84.5, p = 0.048$).

4.3 The Role of Task Context

In order to answer RQ3, we added task context as a variable in our analysis. For average male and female traits, context dependence is only seen for average male traits: the male low (*Mdn. assistant task (AT)*: 2.665, *Mdn. compliance task (CT)*: $3.08, U = 119, p = 0.028$), male high ($Mdn.AT : 2.42, Mdn.CT : 3.0, U = 177, p = 0.020$), and gender-ambiguous voice ($Mdn.AT : 2.08, Mdn.CT : 2.83, U = 113, p = 0.021$) score higher on average male traits in the compliance task than the assistance task. Additionally, reported trust was stable over both tasks for male voices, while they were task dependent for female low ($Mdn.AT : 5.225, Mdn.CT : 4.18, U = 250, p = 0.007$), female high ($Mdn.AT : 5.18, Mdn.CT : 4.045, U = 177, p = 0.003$) and gender-ambiguous voices ($Mdn.AT : 5.045, Mdn.CT : 4.18, U = 122, p = 0.038$). In fact, all voices scored higher on trust for the assistance task compared to the compliance task.

5 DISCUSSION

This study found evidence for the influence of voice gender and pitch on (stereotypical) trait attribution. While no positive stereotype activation was found, negative stereotypical trait ascription, and the lack thereof, showed implicit activation of gender stereotyping. For example, while the low male voice was not explicitly considered to be stereotypically male, only the low male voice was not perceived *not* to be typically male, and only the low voices—both male and female—were not refuted to be stereotypically female. As for trust attribution, we did not identify direct effects of voice pitch. However, a trend showed higher trust in the gender-ambiguous voice for female participants. Finally, task context influences both stereotype activation (for male traits) and trust (for female and gender-ambiguous voices).

With regards to trait attribution, our findings show mixed results with respect to the CASA paradigm. Negative trait ascription was prevalent, which can be both due to a lack of a perceived trait or a lack of viewing the voice as a social actor all together. While active stereotype activation was missing, the absence of stereotype negation seems to indicate an implicit gender bias. The fact that male traits and trust in female (and gender-ambiguous) voices was context dependent, indicates voice pitch and voice gender does subtly influence perception. With regards to trust formation, our results do not seem in line with prior research on the effect of pitch in inter-personal interactions, which indicate that people generally trust people with high-pitched voices more [19]. This may indicate that some of those mechanisms may be weaker in human-computer interaction. However, it has to be noted, that the means, especially for voices with the particularly low and high voice, reveal a trend towards higher-pitched voices being trusted more. As our sample size is comparatively small, those results may become significant with a more appropriate sample size.

Task context did have an effect on perception and stereotype activation. Male voices were perceived more stereotypically male in a more 'male' context of a compliance task. Female voices on the other hand were significantly more trusted in assistance tasks when compared to a compliance task. Curiously, these effects were both present for the gender-ambiguous voice: perceived male traits and trust were context dependent. While it is a positive indication that the gender-ambiguous voice was not assigned one specific gender, it also shows a risk: because the voice does not fit one gender stereotype, it also does not fit one stereotypical response, making it sensitive to multiple possible responses.

Nevertheless, the gender-ambiguous voice showed no significant trust differences when compared to the gendered voices. This is a promising first result, as there is very little research on the impact of gender-ambiguous voices. The fear that lack of a mental model and added cognitive load due to unrecognizable sex of the voice negatively influences trust does not seem to be confirmed by our research. More research is needed into different contexts and different pitches to confirm that the gender-ambiguous voice does not have a negative impact on trust compared to gendered voices. Overall, the gender-ambiguous voice was found to be organized, confident, cooperative, and polite; just as the gendered voices. This seems to be a encouraging initial resultfor use of gender-ambiguous voices in VAs. The fact that women have a higher trust in the gender-ambiguous voice than men warrants further research.

Our study had some limitations that should be pointed out. First, for a quantitative study, our sample size is comparatively small due to the study's exploratory character. Second, the participants were asked to imagine the scenarios to be real-life, which may threaten the external validity of our results. Although our study included actual voice interaction, future work may reexamine the results in a field setting. Third, it should be noted that results only capture first impressions of the VAs. A longitudinal perspective on trait ascription and trust formation should included in future work. Furthermore, three different Google WaveNet voices were used to create the voices used in our experiment. We did not control for other voice characteristics such as timbre and tone, which could have influenced our results.

Additionally, a limitation lies in the created gender-ambiguous voice. As indicated, voice gender does not only come from pitch, but also from language usage and intonation. We controlled for this as much as possible by testing different pitch shifts for different voices, but all voices were originally gendered. The gender-neutral voice clip called Q [24] was recorded by using voices of people that neither ascribe to the male gender nor the female one. The lack of gender-ambiguous voice generators or text-to-speech tools hampers the research into the possibilities of such a voice.

6 CONTRIBUTION AND FUTURE WORK

Our study has several theoretical and practical contributions to prior work in HCI research on the use of VA in commercial settings, the role of para linguistic cues for trait attribution, and the effective design of a VA's 'personality'. To the best of our knowledge, this is the first line of systematic research demonstrating how variations in voice pitch induce gender-specific trait attribution towards the agent, how such attributions affect important perceptual downstream consequences such as trust, and how such changes are impacted by the task context. Moreover, we develop and comparatively evaluate a gender-ambiguous voice with promising first results.

Our findings show stereotype activation is not as clear-cut as one might expect, but appears as a lack of stereotype negation. This combined with the influence of the participants' gender and task context asks for a more in-depth examination into stereotype activation and perpetuation of VAs. Additionally, gender-ambiguous voices are a promising avenue of research for VA design, to strive for more inclusive design. However, there is currently a lack of tools providing gender-ambiguous voice generation. We call upon researchers and industry alike to focus on the creating of gender-ambiguous voice tools to be able to research and provide more inclusive and stereotype avoiding voices for VAs.

ACKNOWLEDGMENTS

This work is supported by the Swiss National Science Foundation, Grant 192718. Andreas Janson acknowledges support from the basic research fund of the University of St.Gallen. We thank Prof. Leah Ruppanner for her valuable feedback.

REFERENCES

[1] Alice Baird, Stina Hasse Jørgensen, Emilia Parada-Cabaleiro, Nicholas Cummins, Simone Hantke, and Björn Schuller. 2018. The perception of vocal traits in synthesized voices: age, gender, and human likeness. *Journal of the Audio Engineering Society* 66, 4 (2018), 277–285.

[2] Sandra L Bem. 1981. *Bem Sex-Role Inventory: professional manual.* Consulting Psychologists Press, Palo Alto, CA, USA.

[3] Izak Benbasat and Weiquan Wang. 2005. Trust in and adoption of online recommendation agents. *Journal of the association for information systems* 6, 3 (2005), 4.

[4] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 1–13.

[5] Christy L Dennison. 2006. *The effect of gender stereotypes in language on attitudes toward speakers.* Ph.D. Dissertation. University of Pittsburgh.

[6] Friederike Eyssel and Frank Hegel. 2012. (s) he's got the look: Gender stereotyping of robots 1. *Journal of Applied Social Psychology* 42, 9 (2012), 2213–2230.

[7] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies* 132 (2019), 138–161.

[8] Márcio Fuckner, Jean-Paul Barthès, and Edson Emílio Scalabrin. 2014. Using a personal assistant for exploiting service interfaces. In *Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD).* IEEE, IEEE, Piscataway, N.J., USA, 89–94.

[9] Eva Gustavsson. 2005. Virtual servants: stereotyping female front-office employees on the internet. *Gender, Work & Organization* 12, 5 (2005), 400–419.

[10] Christian Hildebrand, Fotis Efthymiou, Francesc Busquet, William H Hampton, Donna L Hoffman, and Thomas P Novak. 2020. Voice analytics in business research: Conceptual foundations, acoustic feature extraction, and applications. *Journal of Business Research* 121 (2020), 364–374.

[11] Jessica Junger, Katharina Pauly, Sabine Bröhr, Peter Birkholz, Christiane Neuschaefer-Rube, Christian Kohler, Frank Schneider, Birgit Derntl, and Ute Habel. 2013. Sex matters: Neural correlates of voice gender perception. *Neuroimage* 79 (2013), 275–287.

[12] Anne M Koenig, Alice H Eagly, Abigail A Mitchell, and Tiina Ristikari. 2011. Are leader stereotypes masculine? A meta-analysis of three research paradigms. *Psychological bulletin* 137, 4 (2011), 616.

[13] Eun Ju Lee, Clifford Nass, and Scott Brave. 2000. Can computer-generated speech have gender? An experimental test of gender stereotype. In *CHI'00 extended abstracts on Human factors in computing systems.* Association for Computing Machinery, New York, NY, USA, 289–290.

[14] Wade J Mitchell, Chin-Chang Ho, Himalaya Patel, and Karl F MacDorman. 2011. Does social desirability bias favor humans? Explicit–implicit evaluations of synthesized speech support a new HCI model of impression management. *Computers in Human Behavior* 27, 1 (2011), 402–412.

[15] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 1–7.

[16] Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of applied social psychology* 27, 10 (1997), 864–876.

[17] 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.* Association for Computing Machinery, New York, NY, USA, 72–78.

[18] Nora Ni Loideain and Rachel Adams. 2018. From Alexa to Siri and the GDPR: the gendering of virtual personal assistants and the role of EU data protection law.

[19] Jillian JM O'Connor and Pat Barclay. 2017. The influence of voice pitch on perceptions of trustworthiness across social contexts. *Evolution and Human Behavior* 38, 4 (2017), 506–512.

[20] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. 2018. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning.* PMLR, Stockholm, Sweden, 3918–3926.

[21] Richard Pak, Anne Collins McLaughlin, and Brock Bass. 2014. A multi-level analysis of the effects of age and gender stereotypes on trust in anthropomorphic technology by younger and older adults. *Ergonomics* 57, 9 (2014), 1277–1289.

[22] Emmi Parviainen and Marie Louise Juul Søndergaard. 2020. Experiential Qualities of Whispering with Voice Assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 1–13.

[23] Amanda Purington, Jessie G Taft, Shruti Sannon, Natalya N Bazarova, and Samuel Hardman Taylor. 2017. " Alexa is my new BFF" Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 2853–2859.

[24] Q. 2019. The First Genderless Voice. 2019. Meet Q: The First Genderless Voice-FULL SPEECH.

[25] Giuseppe Riccardi. 2014. Towards healthcare personal agents. In *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges.* Association for Computing Machinery, New York, NY, USA, 53–56.

[26] Elizabeth A Strand. 2000. *Gender stereotype effects in speech processing.* Ph.D. Dissertation. The Ohio State University.

[27] 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.* Association for Computing Machinery, New York, NY, USA, 1–8.

[28] Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 2019. Voice as a design material: sociophonetic inspired design strategies in Human-Computer

Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[29] UNESCO. 2019. *I'd blush if I could – closing gender divides in digital skills through education.* Technical Report. UNESCO.

[30] Maria Vernuccio, Michela Patrizi, and Alberto Pastore. 2020. Brand Anthropomorphism and Brand Voice: The Role of the Name-Brand Voice Assistant. In *Advances in Digital Marketing and eCommerce*. Springer, Switzerland, 31–39.

[31] Qian Yu, Tonya Nguyen, Soravis Prakkamakul, and Niloufar Salehi. 2019. " I Almost Fell in Love with a Machine" Speaking with Computers Affects Self-disclosure. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–6.

[32] Naim Zierau, Edona Elshan, Camillo Visini, and Andreas Janson. 2020. A Review of the Empirical Literature on Conversational Agents and Future Research Directions. In *ICIS 2020 Proceedings*. Association for Information Systems, Atlanta,GA,USA, 1–17.