Effects of Smart Virtual Assistants' Gender and Language

Florian Habler
University of Regensburg
Regensburg, Germany
Florian.Habler@student.ur.de

Valentin Schwind
University of Regensburg
Regensburg, Germany
valentin.schwind@ur.de

Niels Henze
University of Regensburg
Regensburg, Germany
niels.henze@ur.de

ABSTRACT

Smart virtual assistants (SVA) are becoming increasingly popular. Prominent SVAs, including Siri, Alexa, and Cortana, have female-gendered names and voices which raised the concern that combining female-gendered voices and submissive language amplifies gender stereotypes. We investigated the effect of gendered voices and the used language on the perception of SVAs. We asked participants to assess the performance, personality and user experience of an SVA while controlling the gender of the voice and the attributed status of the language. We show that low-status language is preferred but the voice's gender has a much smaller effect. Using low-status language and female-gendered voices might be acceptable but solely combining low-status language with female-gendered voices is not.

KEYWORDS

smart virtual assistant; natural language interface; gender bias

ACM Reference Format:
Florian Habler, Valentin Schwind, and Niels Henze. 2019. Effects of Smart Virtual Assistants' Gender and Language. In *Mensch und Computer 2019 (MuC '19), September 8–11, 2019, Hamburg, Germany*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3340764.3344441

1 INTRODUCTION & BACKGROUND

Virtual assistants, including Apple's Siri, Amazon Echo with Alexa, Microsoft's Cortana and Google Assistant, became commercially successful products. They have been integrated into different form factors and are typically controlled through a natural language user interface. Users started to embed them in the life of the home [10]. Current commercial virtual assistants come with virtual human-like personalities. Users of voice user interfaces might, therefore, personify the devices [4, 11]. This is in line with recent work that suggests to consider computing devices as social objects [13]. Lopatovska and Williams found that some of their participants reported personifying behaviors [4]. Similarly, public media seems to also shift towards personifying virtual assistants [5]. Analyzing user reviews, Purington et al. conclude that personification predicts satisfaction with the virtual assistant [11]. It seems, therefore, advisable for companies to increase the probability that users personify their virtual assistants and companies likely already do so to strengthen the connection with the user.

At least in Germany and the US, all prominent virtual assistants have female-gendered voices by default. Siri, Alexa, and Cortana have names that suggest a female personality. Hannon discusses that even their language imitates the language associated with persons considered female [2]. Using female-gendered personalities for voice user interfaces is not new. Navigation systems are, for example, also typically equipped with female-gendered voices. Consequently, previous work investigated if gender stereotypes are transferred to machines. By conducting a study that varied the computer's voice Nass et al. showed that users indeed transfer gender stereotypes to machines [8]. The authors found that evaluations by male-gendered voices were considered more valid than evaluations by female-gendered voices which is in line with the assessment of evaluations by male and female persons.

Nass and Brave point out that people recognize the gender of a machine when the voice has only the slightest hint of gender [7]. They stated that individuals commonly perceive female-gendered voices as helping us to solve our problems by ourselves, while they view male-gendered voices as rather authoritarian characters who tell us the answers to our problems. Furthermore, Mitchell et al. examined preferences for gender in synthesized voices [6]. The results of the study indicate that women and men preferred female-gendered synthesized voices, which they described as sounding warmer than male-gendered synthesized voices.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MuC '19, September 8–11, 2019, Hamburg, Germany
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7198-8/19/09…$15.00
https://doi.org/10.1145/3340764.3344441

Pennebaker analyzed a large text corpus to determine the distribution of I-words, pronouns, verbs, nouns, etc. [9]. Pennebaker concludes that men use more "big words" and nouns whereas women tend to prefer personal pronouns and verbs. Moreover, women use hedge phrases at a much higher rate to acknowledge that their opinion is not necessarily the one that is right. Hedge phrases are mitigating words used to avoid the appearance of bragging to soften conversations or statements. Pennebaker also discusses the words used by persons belonging to lower middle, middle, or upper-middle social classes. Upper-middle-class members most frequently use big words, which are long and difficult words. Conversely, people of lower social classes as well as people who are insecure or depressed build their sentences with more personal pronouns and first-person singular pronouns to indicate submissiveness. To sum up, the higher the social class, the less likely one uses first-person singular pronouns and the less one uses emotion words.

Current virtual assistants are designed to be personified and with female personas in mind. It might not be problematic if users transfer gender stereotypes to machines. It can still be problematic if the virtual assistants are designed with gender stereotypes in mind as it could amplify existing gender stereotypes. Indeed, Alexa, the assistant analyzed by Hannon [2], uses words that correspond not only to words by women but also by people (male or female) who occupy a lower status in a relationship. Furthermore, there are differences between male and female users when embodying or designing different-gendered characters [14, 15]. Therefore, it is important to determine if users gender also affects their preferences when interacting with virtual assistants.

While previous work revealed effects for gendered voices and suggested negative consequences of virtual assistants using a submissive language, the interaction between these two factors is unclear. In this paper, we, therefore, conduct a study that investigated the effect of virtual assistants' voice and language. Through a controlled experiment we show that low-status language is appreciated while the voice's gender seems to be a less important factor.

2 METHOD

We conducted a study that investigates the effect of language (low-status / high-status) and voice (male-gendered / female-gendered) on the perception of virtual assistants. We used a repeated measure full-factorial 2×2 design with the independent variables language and voice resulting in four conditions. We used a Latin square design to avoid sequence effects. Participants were asked to rate four virtual assistants after performing a set of tasks. The virtual assistants were simulated using prerecorded audio controlled by an experimenter (see Figure **??** & 1).

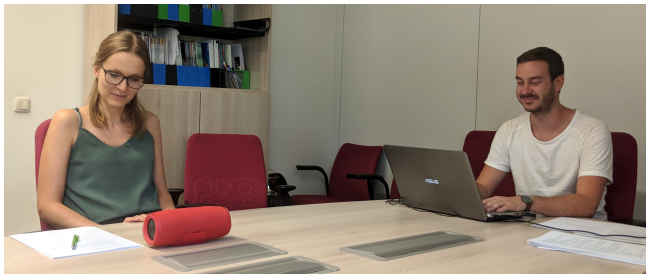

Figure 1: The study setup with the experimenter sitting on the right.

We designed four sets of tasks, each containing sub-tasks with different complexity levels (see Table 1 for an example). The tasks are based on common tasks users perform with virtual assistants. The tasks followed the same general pattern and started with a simple context-less question[1]. E.g., "Determine the current weather in your city." which was answered with "It is sunny and has 28 degrees." no matter of status. The initial sub-task was followed by more complex sub-tasks (see 1 for an example). An example of a more complex sub-task was letting the system recommend a song. After a participant asked the system to recommend a song it answered with "From which genre should the song be?" or "From which genre can I recommend you a song?". After a participant asked which genres are available the system answered with "There are the genres hip-hop, Schlager, rock, and pop." or "I found the genres hip-hop, Schlager, rock, and pop." After answering pop, the system responded with "'phenomenal' was found as a recommendation" or "I found 'phenomenal' as a recommendation for you".

To reduce biases caused by the voices themselves, we looked for consistently high-quality voices. Male and female voices should each come from the same software or platform to avoid confounding variables. As we conducted the study with German participants, we had to select two female and two male German voices. After testing a number of text-to-speech engines we decided for "Marlene" and "Hans" from Amazon Web Services as well as "Michael" and "Nadine" from fromtexttospeech.com.

To assess participants' satisfaction, we used the customer satisfaction questionnaire scale [1]. To learn about how participants assess the joint performance they achieve with the virtual assistant, we asked them to rate the perceived performance using six items contributing to a single scale [1]. To learn about effects on the user experience, we took the 10-item version of the AttrakDiff [3]. In line with previous work [16], we determined the big five personality traits for each condition to reveal if the four conditions are perceived to have different personalities. Thus, we asked participants

[1]Conversations were in German. Here we provide English translations.

Table 1: Example of the instructions for participants and the low- and high-status answers provided by the system. The example has been translated to English. Placeholders for names and items are in brackets. Participants had one minute before each task to think about the conversation and come up with names, items, location etc.

Instructions	Low-status answer	High-status answer
Create an entry for (day) (time) in the calendar.	Okay, I saved the entry for you.	Okay, the entry was saved.
Create a WhatsApp group.	Okay, who should I invite for you?	Okay, who should be invited?
(Say 5 names of the group members.)	Okay, I'm inviting these people. What name should I give the group?	Okay, these people have been invited. What name should the group have?
(Say name of the group.)	Okay, I created the group for you.	Okay, the group has been created.
Write (5 items) on the shopping list.	Okay, I put this on the shopping list.	Okay, this has been added to the shopping list.
How will the weather be on (day) in (location)?	It gets sunny and gets temperatures between 28 and 31 degrees.	It gets sunny and gets temperatures between 28 and 31 degrees.

to fill a 10-item version of the Big Five Inventory [12]. Finally, we added a single five-point Likert item that asked as for the dominance of the system.

The experiment took about 45 minutes per participant. After welcoming participants we provided an overview of the procedure, asked them to fill an informed consent form and asked for demographic information. Afterward, we introduced the questionnaires they had to answer after completing a task.

To ensure that all participants have at least a basic understanding of virtual assistants, we deliberately recruited participants with a background in technology. In total, 24 participants (12 female, 12 male) took part in the study. They were between 19 and 32 years old ($M = 23.81y, SD = 2.81$). Thirteen participants studied media informatics or computer science or have a degree in one of them. Two participants previously used virtual assistants several times per day. Three participants use a virtual assistant daily, two participants use it several times a week and two participants every week. Fifteen participants stated that they rarely use an SVA.

3 RESULTS

After completing all tasks, we explained to participants that the virtual assistant was simulated by the experimenter. Based on participants' feedback, we conclude that participants did not realize that the virtual assistant was a simulation. We used a two-way repeated-measures analysis of covariance (ANCOVA) with participants' gender as a covariate to determine the effects of language and voice. For

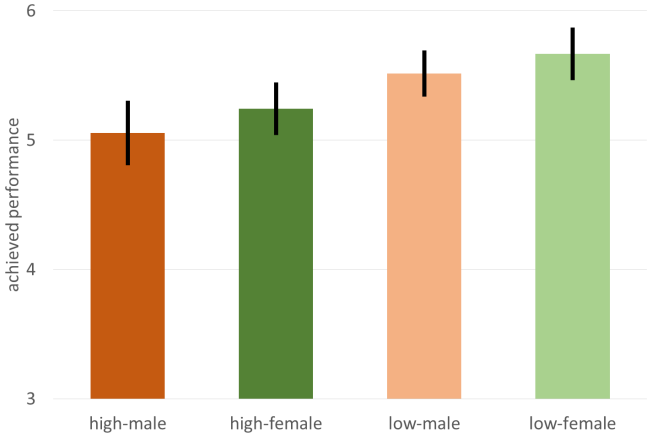

Figure 2: Achieved performance. Error bars show the standard error.

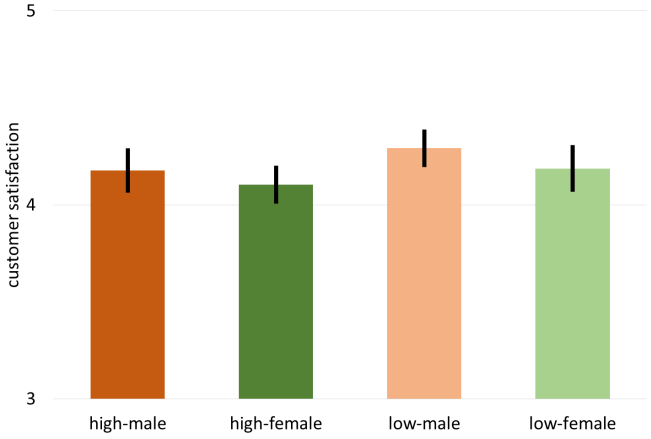

Figure 3: Customer satisfaction. Error bars show the standard error.

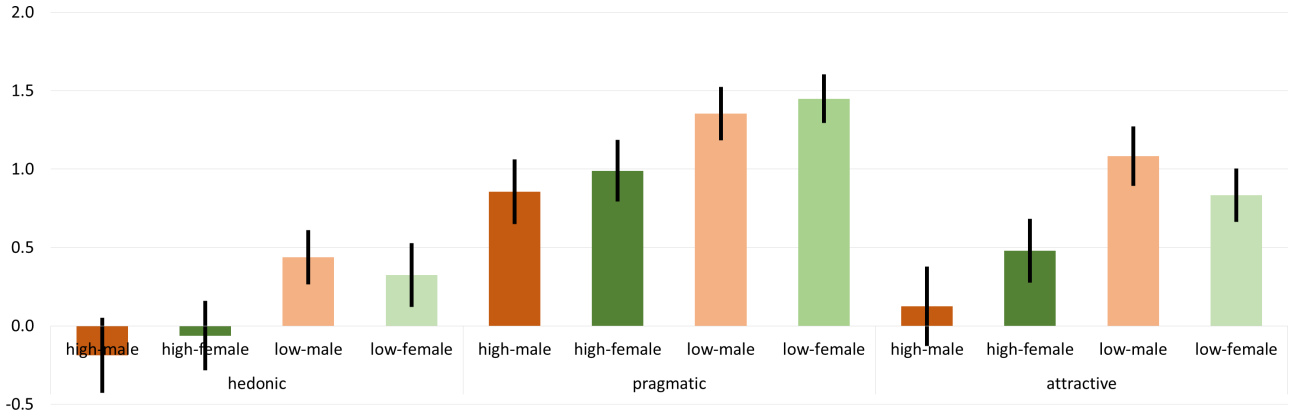

Figure 4: The subscales hedonic, pragmatic and attractive of the AttrakDiff. Error bars show the standard error.

the used full-factorial 2×2 design, the ANCOVA naturally prevents inflation of Type I errors.

The ANCOVA revealed a significant effect of language on achieved performance, $F(1, 22) = 10.12$, $p = .004$, $\eta_p^2 = .315$. Low-status language resulted in higher achieved performance ratings (see Figure 2). While we observed a higher achieved performance for female-gendered voices, we found no significant effect of voice on achieved performance, $F(1, 22) = 1.89$, $p = .183$, $\eta_p^2 = .079$). The ANCOVA revealed a significant interaction voice $\times$ gender effect, $F(1, 22) = 7.18$, $p = .014$, $\eta_p^2 = .246$. While we observed a slightly higher customer satisfaction for low-status language (see Figure 3), the ANCOVA revealed no significant effect of language on customer satisfaction, $F(1, 22) = 2.11$, $p = .160$, $\eta_p^2 = .088$). We observed a higher customer satisfaction for male-gendered voices, but the ANCOVA did not reveal a significant effect, $F(1, 22) = 1.12$, $p = .302$, $\eta_p^2 = .048$). We did not reveal a significant effect on any of the five personality traits (all p>.05).

We analyzed each of the AttrakDiff's sub-scales. The ANCOVA revealed a significant effect of language on the hedonic quality, $F(1, 22) = 10.41$, $p = .004$, $\eta_p^2 = .321$. Low-status language resulted in higher hedonic quality (see Figure 4). While we observed a higher hedonic quality for male-gendered voices, we found no significant effect of voice on hedonic quality, $F(1, 22) = 0.001$, $p = .982$, $\eta_p^2 < .001$). The ANCOVA revealed a significant effect of language on the pragmatic quality, $F(1, 22) = 12.10$, $p = .002$, $\eta_p^2 = .355$. Low-status language resulted in higher pragmatic quality (see Figure 4). While we observed a higher pragmatic quality for female-gendered voices, we found no significant effect of voice on the pragmatic quality, $F(1, 22) = 0.39$, $p = .537$, $\eta_p^2 = .018$. The ANCOVA revealed a significant effect of language on the attractive quality, $F(1, 22) = 10.44$, $p = .004$, $\eta_p^2 = .322$. Low-status language resulted in higher attractive quality (see Figure 4). We found no significant effect of voice on the attractive quality, $F(1, 22) = 0.07$, $p = .800$, $\eta_p^2 = .003$.

4 DISCUSSION & CONCLUSION

We conducted a controlled experiment that assessed the effect of virtual assistants' voice and language. Low-status language consistently received more positive ratings. Participants felt they achieved a higher performance. Assistants with low-status language had a higher hedonic, pragmatic, and attractive quality. We did not reveal significant differences between male- and female-gendered voices. The size of the effect for language we observed is clearly higher than the size of the effect for voice.

This indicates that users personify virtual assistants and that this personification influences future interaction with other persons. If this is indeed the case, interacting with low-status female-gendered assistants could increase the expectation that females have (or should have) a lower status. Our results also indicate that users appreciate low-status virtual assistants. While users might also prefer female-gendered voices, this seems to be less important. Consequently, we suggest that virtual assistants with low-status language should either be male-gendered or the gender should be randomly assigned with giving users the option to change it.

Our work only provides indicators for users' preferences. Longer-term studies are required to confirm the results. We believe, however, that it would be even more interesting to investigate if the interaction with personified computing systems changes how users of such systems interact with other people. We believe that this is an important direction for future work of systems that are designed for personification are becoming increasingly popular. Ultimately, we should investigate if users transfer the behavior that they used when interacting with computers to the interaction with other persons.

REFERENCES

[1] Thomas M Brill. 2018. *Siri, Alexa, and Other Digital Assistants: A Study of Customer Satisfaction With Artificial Intelligence Applications.* Ph.D. Dissertation. University of Dallas.

[2] Charles Hannon. 2016. Gender and Status in Voice User Interfaces. *interactions* 23, 3 (April 2016), 34–37. https://doi.org/10.1145/2897939

[3] Marc Hassenzahl. 2003. Funology: From Usability to Enjoyment. Kluwer Academic Publishers, Norwell, MA, USA, Chapter The Thing and I: Understanding the Relationship Between User and Product, 31–42.

[4] Irene Lopatovska and Harriet Williams. 2018. Personification of the Amazon Alexa: BFF or a Mindless Companion. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18).* ACM, New York, NY, USA, 265–268. https://doi.org/10.1145/3176349.3176868

[5] Thomas Messerli, Steve Duman, and Les Sikos. 2018. From Screen to Voice: Evidence of changed perceptions of voice-based virtual assistants. *IMPEC: Interactions Multimodales par Ecran* (2018), 79.

[6] 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. https://doi.org/10.1016/j.chb.2010.09.002 Current Research Topics in Cognitive Load Theory.

[7] Clifford Nass and Scott Brave. 2005. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship.* The MIT Press.

[8] Clifford Nass and Li Gong. 2000. Speech Interfaces from an Evolutionary Perspective. *Commun. ACM* 43, 9 (Sept. 2000), 36–43. https://doi.org/10.1145/348941.348976

[9] James W. Pennebaker. 2011. The secret life of pronouns. *New Scientist* 211, 2828 (2011), 42 – 45. https://doi.org/10.1016/S0262-4079(11)62167-2

[10] 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, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214

[11] 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 (CHI EA '17).* ACM, New York, NY, USA, 2853–2859. https://doi.org/10.1145/3027063.3053246

[12] Beatrice Rammstedt and Oliver P. John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality* 41, 1 (2007), 203 – 212. https://doi.org/10.1016/j.jrp.2006.02.001

[13] Valentin Schwind, Niklas Deierlein, Romina Poguntke, and Niels Henze. 2019. Understanding the Social Acceptability of Mobile Devices Using the Stereotype Content Model. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* ACM, New York, NY, USA, Article 361, 12 pages. https://doi.org/10.1145/3290605.3300591

[14] Valentin Schwind and Niels Henze. 2018. Gender- and Age-related Differences in Designing the Characteristics of Stereotypical Virtual Faces. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '18).* ACM, New York, NY, USA, 463–475. https://doi.org/10.1145/3242671.3242692

[15] Valentin Schwind, Pascal Knierim, Cagri Tasci, Patrick Franczak, Nico Haas, and Niels Henze. 2017. "These Are Not My Hands!": Effect of Gender on the Perception of Avatar Hands in Virtual Reality. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17).* ACM, New York, NY, USA, 1577–1582. https://doi.org/10.1145/3025453.3025602

[16] Michelle X. Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting Virtual Agents: The Effect of Personality. *ACM Trans. Interact. Intell. Syst.* 9, 2-3, Article 10 (March 2019), 36 pages. https://doi.org/10.1145/3232077