Exploring the impact of Chatbot Voice Gender on User Satisfaction - Group 21

SACHA VUCINEC, MAUD OVERBEEK, ALEXO CASTRO YÁÑEZ, YFKE SMIT, DOUWE BRINK

1 INTRODUCTION

Using automated dialog systems is becoming increasingly popular for communicating efficiently with users in many different domains such as customer service, planning services, assistance and online chatting [5]. Questions that users have can be easily answered by so-called chatbots, without the need for extra human resources. Optimizing the user experience in these machine to human interactions is key for many businesses and organizations [8]. Within all the different factors that play a role in the user-perceived satisfaction from the interaction with the chatbot system, we would like to remark the importance on the way in which the system present its responses to the user. We may think now in many features that could be adjusted when presenting system output as the speed of chatbot replies, the formality of those replies, or the use of text-to-speech with all its involved configurations (e.g. tone, speed and gender of the voice). This paper is focused on seeing how the gender of the text-to-speech voice embedded in a recommendation system may have an effect in user experience when using the system .

In many popular AI conversational agents such as Apple's Siri, Microsoft's Cortana, Amazon's Alexa and many more, a female voice is used by default [4] [3]. Recent studies have suggested there is a difference in stereotypical perception of chatbots by users and in user satisfaction when using those chatbots dependent on the perceived gender of the chatbot by the user [7] [2], with a significant difference in favour of non-gendered chatbots by users. It is important to study whether this gender bias difference extends to different forms of chatbots and dialogue systems within diverse fields, so future AI-powered voice products can be configured to maximize user satisfaction. We will work with a voice assisted recommendation system and test whether there will be a difference in satisfaction between users who interact with a chatbot with a female voice and recognise it as females, users who interact with a male voice and recognise it as male and users who are not able to recognise the gender of the voice they interact with. Accordingly, the following hypothesis is advanced:

> H1: There will be a significant difference in user satisfaction based on the assigned gender of the chatbot by the user (male, female, not sure)

Additionally, we hypothesize that there is an effect between gender of participant and the chatbot voice gender. The following hypothesis is proposed to be tested:

> H2: There will be a significant difference in user satisfaction based on the predefined chatbot gender configuration (male or female) and on participants gender.

In previous research user satisfaction using chatbots was measured by a questionnaire [1] In this paper, we follow this approach, using a questionnaire that uses the Likert scale [6]. Our approach is further laid out in Section 2. In Section 3 the results of the experiment are presented. Section 4 contains the discussion on our approach and limitations of this study.

Author's address: Sacha Vucinec, Maud Overbeek, Alexo Castro Yáñez, Yfke Smit, Douwe Brink.

2 METHODS

2.1 Participants

A total of 32 students participated in this experiment. In terms of gender identity, 16 participants identified themselves as female, 16 as male, and no one as “other" (so this category is not included anymore in the data). Participants were divided into four groups based on their gender and the gender of the text-to-speech voice used during the interaction (see Table 1). Each group consisted of eight participants. Five participants were not sure when assigning the gender of the text-to-speech voice. This is explained further in section 4. There was also one participant who assigned the incorrect gender to the text-to-speech voice making this data point invalid.

Table 1. Distribution of participant groups.

Group	Participant Gender	Text-to-Speech Voice Gender
A	Male	Male
B	Male	Female
C	Female	Male
D	Female	Female

2.2 Experimental Design

The experiment was conducted in person and was based on a 2 (chatbot voice gender: male vs. female) × 2 (participant gender: male vs. female) between-subjects design.

The study aimed to explore between-group differences in user satisfaction, particularly focusing on the influence of the chatbot’s voice gender. User satisfaction was assessed using a self created questionnaire addressing overall user satisfaction, specific aspects of voice settings (e.g., understandability), and participant response validation through both positively and negatively framed questions, which can be seen in Table 2.

2.3 Materials

The experiment used a chatbot for recommending restaurants in the Cambridge area utilizing the pyttsx3[1] library to convert text responses to speech. Two Windows System voices[2] were used: ‘David’ for the male voice and ‘Zira’ for the female voice. Participants were provided with headphones for the interaction. Following the interaction, participants completed the questionnaire shown in Table 2.

2.4 Procedure

Before setting up the experiment the ethics checklist was checked. During the experiment participants provided informed consent before being briefed on the chatbot’s functionality. The text-to-speech voice gender was then set according to the participant’s group assignment, with a male chatbot voice for participants of groups A and C, and a female chatbot voice for participants of groups B and D. Participants interacted with the chatbot, completing two tasks as outlined in Table 2.

[1]https://pypi.org/project/pyttsx3/.
[2]An installation guide can be found here.

Table 2. Tasks and questionnaire for participants following chatbot interaction.

Tasks:
Task 1: Find the phone number of a restaurant that serves Chinese food in the north part of town.
Task 2: Find the name of a restaurant serving any food in any price range in the center that is touristic.
Questionnaire:
Q1: I would recommend this system to a friend.
Q2: The system's voice responses were unclear and hard to follow.
Q3: I would not use this system again.
Q4: The system's voice was clear and understandable.
Q5: I did not like the recommendation system.
Q6: I like the way the information is presented.
Q7: In case of using the system again, I would like it to have a different voice.
Q8: I am overall satisfied with the system.
Q9: What gender would you assign to the voice of the chatbot?
Q10: What is your gender?

2.5 Measurements

Participants rated questions 1 to 8 in Table 2 on five-point Likert scale from 1 (completely disagree) to 5 (completely agree). We deal with negatively framed questions just by assigning to the corresponding item six minus the actual value of the question.

2.5.1 Manipulation checks. In the last section of the experiment, we use question 9 to check if the participants could correctly classify the text to speech voice gender and question 10 asked about the participant's gender. One unique participant classified incorrectly the gender of the voice, so we only used 31 out of 32 participant responses in the data analysis. Five of the remaining participants were not sure about the system voice gender, all of them being male and all but one having interacted with the male voice. The 26 remaining participants were able to identify the gender of the voice correctly. Due to the disparate distribution of users that were not sure about chat voice gender along the independent variables, we have decided to consider as correctly assigned all the not sure responses in order to test the hypothesis H2, and only after comparing the results between the different groups by gender assignation.

2.5.2 Dependent variable. The user satisfaction when interacting with the chatbot was measured along two different dimensions: the overall satisfaction of the user and the user's preference to change the voice, what we will call voice satisfaction (overall satisfaction: 1 = not recommended, not repeatable, not liked; 5 = recommended, repeatable, liked; voice satisfaction: 1 = will not repeat voice, 5 = will repeat voice). The overall satisfaction is measured in four items related with questions 1, 3, 6 and 8 in Table 2, while voice satisfaction is measured with one item corresponding with question 7. We then define our satisfaction dependent variable as the sum (or average) of the two considered dimensions.

2.5.3 Control variable. User-perceived system comprehensibility was used as a control variable and was measured using two different items: the system explainability (1 = information poorly presented, 5 = information well presented) and the voice clarity and understandability (1 = not clear, not understandable; 5 = clear, understandable).

3 RESULTS

3.1 Descriptive Statistics

Table 3. Statistical analysis of selected variables

	Descriptive Statistics			Correlation matrix		
Variables	Mean	Standard Deviation	95% CI	1	2	3
1. Overall Satisfaction	3.234	0.998	(1.196, 5.272)	1	0.259	0.450
2. Voice Satisfaction	3.064	1.263	(0.485, 5.644)	0.259	1	0.353
3. System Comprehensibility	3.925	0.734	(2.426, 5.424)	0.450	0.353	1

The descriptive statistics related with our dependent and control variables for the whole group of participants can be found in Table 3. The matrix of correlations is also provided, in which we can see that overall and voice satisfaction have a correlation around 26%, despite being two dimensions of the user satisfaction variable. This reaffirms the usage of both of them.

3.2 Normality tests

In order to select the appropriate statistical tests, we have performed a Shapiro-Wilk normality test to each variable to make sure that we can assume that it follows a normal distribution. We cannot reject the null hypothesis that the overall satisfaction comes from a normal distribution (W = 0.977, p = .72), whereas we have significant evidence that both the voice satisfaction (W = 0.912, p = .014) and system comprehensibility (W = 0.909, p = .012) do not follow a normal distribution. However, if we combine both overall and voice satisfaction variables into a general satisfaction variable by adding (or averaging) them, we can keep the assumption of normality. This is because we have not found significant evidence to reject this assumption (W = 0.957, p = .27), so we can perform parametric tests when comparing the satisfaction variable.

3.3 Gender-based analyses

We conducted gender-based analyses based on the following groups: A, B, C, and D. Descriptive statistics for each group are presented in Table 4 and Table 5.

Table 4. Overall Satisfaction by group

Group	Overall Satisfaction Mean	Overall Satisfaction Median
A	3.4	3.5
B	3.3	4.3
C	3.0	4.0
D	2.6	3.3

Table 5. System Comprehensibility by group

Group	System Comprehensibility Mean	System Comprehensibility Median
A	3.6	3.7
B	4.2	4.0
C	3.7	4.0
D	3.7	3.3

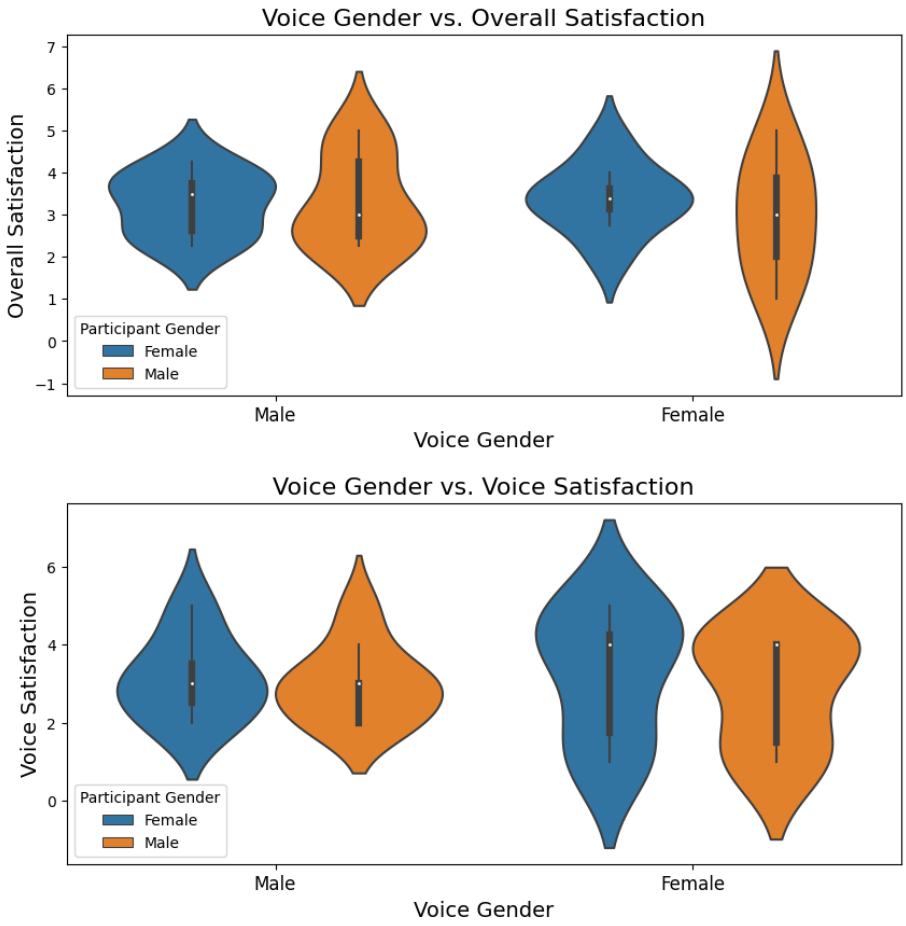

Fig. 1. Probability distributions per group

The probability distributions in Figure 1 show that for voice satisfaction, the male participants had similar distributed data, as do the female participants. However, the mean voice satisfaction for the female voice gender is higher for both male and female participants and those groups have more outliers.

3.4 Statistical tests

To test H1 a one-way ANOVA was conducted between the groups of users who correctly identified male voice, users who correctly identified female voice and users who did not identify voice gender. Preliminary analysis showed no violations of the assumptions of the one-way ANOVA test, as satisfaction variable is supposed to follow a normal distribution as discussed on section 3.2 and there is no significant evidence to reject homoscedasticity (p=.15). The results showed that there was no significant difference on user satisfaction between the three groups ($F[2,29] = 2.45$, $MSE = 2.24$, $p = .104$).

We used a two way ANOVA to test H2, i.e. how user satisfaction differs when using the system on the four participant gender × chatbot gender combinations. Results revealed that there was no significant difference in user satisfaction between the four combinations ($F[2,26] = 1.802$, $MSE = 1.566$, $p = 0.185$). Simple main effect analysis showed that both the chatbot gender ($p = 0.096$) and participant gender ($p = 0.083$) have a marginally significant effect on user satisfaction variable.

4 DISCUSSION

This study aimed to measure the impact of chatbot voice gender on user satisfaction in a restaurant recommendation system. We conducted an experiment involving 33 participants, assigning them to four groups based on both participant and chatbot voice gender. Five out of the 33 participants were unsure when answering question 9, meaning that they were unable to assign a gender to the chatbot voice. For this experiment we decided to include their results for the proper group, as four out of the five participants belonged to group A. Removing these participants from the dataset would result in losing 50% of the data for group A.

Our results can contribute to the existing knowledge regarding user interactions with chatbots. The results showed an effect of a higher user satisfaction with a female voice, but they were insignificant and thus not conclusive. This implies that in this specific context, users may not have gender preferences when it comes to the gender of chatbot voices.

There are certain limitations of this research that could account for the insignificance of the results, which can be further investigated in future research. For example, the study involved a small sample size of 33 participants, which impacts the generalizability and significance of the results. A bigger group of participants would allow for a better perspective on effect relations between the dependent and independent variables. In particular, it would greatly improve the relationships modelled by the two-way ANOVA, since this test requires more data, the more independent groups are considered. Additionally, we also managed certain limitations related with experiment measures and considered dependent variables. We have measured the voice satisfaction dimension with one unique question in our questionnaire. This resulted in the variable having a distribution more similar to a categorical variable than to a continuous variable and thus being far away from being normally distributed. We suggest collecting more data on this point (as done with the overall satisfaction) and averaging it out, in order to find the variable value. This procedure will capture the user preferences better and is more likely to result in a normally distributed dependent variable, which is desired for using high-power parametric tests. We have limited our work to work only with the user satisfaction as a dependent variable, for which we have defined two different dimensions of this variable (overall satisfaction and voice satisfaction) and then we have combined them just by average. Even though this variable merge seems to work as allows us to preserve normality assumption on data, it is an oversimplification of the original data and more advance techniques could result in better data interpretation. For example, we suggest the use of a MANOVA (Multivariate ANOVA) test

which allows us to study the dependence of both overall satisfaction and voice satisfaction without having to combine them, and so preserving their original data. Another possible test improvement would be to take into account that the variables overall satisfaction and voice satisfaction are correlated as discussed in section 3.1, so performing an Analysis of Covariance (ANCOVA) or its multidimensional alternative (MANCOVA) could help us to take this correlation into account. Our study utilized a single measure for voice satisfaction, which might not capture the full spectrum of user preferences.

In conclusion, our research does not indicate a significant effect of chatbot voice gender on user satisfaction in a restaurant recommendation context. However, the insignificance of the results can be a result of the small scope of our research. In further research with larger samples and more extensive user satisfaction measures, more significant results could be found.

REFERENCES

[1] Balaji, D. Assessing user satisfaction with information chatbots: a preliminary investigation, September 2019.

[2] Baxter, D., Mcdonnell, M., and McLoughlin, R. Impact of chatbot gender on user's stereotypical perception and satisfaction. pp. 1–5.

[3] Costa, P., and Ribas, L. Ai becomes her: Discussing gender and artificial intelligence. *Technoetic Arts: A Journal of Speculative Research 17*, 1-2 (2019), 171–193.

[4] Feine, J., Gnewuch, U., Morana, S., and Maedche, A. Gender bias in chatbot design. In *Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019, Revised Selected Papers 3* (2020), Springer, pp. 79–93.

[5] Følstad, A., and Brandtzæg, P. B. Chatbots and the new world of hci. *Interactions 24*, 4 (jun 2017), 38–42.

[6] Joshi, A., Kale, S., Chandel, S., and Pal, D. Likert scale: Explored and explained. *British Journal of Applied Science Technology 7* (01 2015), 396–403.

[7] McDonnell, M., and Baxter, D. Chatbots and Gender Stereotyping. *Interacting with Computers 31*, 2 (04 2019), 116–121.

[8] Verhagen, T., van Nes, J., Feldberg, F., and van Dolen, W. Virtual Customer Service Agents: Using Social Presence and Personalization to Shape Online Service Encounters*. *Journal of Computer-Mediated Communication 19*, 3 (04 2014), 529–545.