# The Impact of gendered voice in hospitality bots on user perception and experience

Nia Racheva
2841982
*Department of Artificial Intelligence, VU Amsterdam*
n.racheva@student.vu.nl

Anastasia Ciubucova
2856242
*Department of Artificial Intelligence, VU Amsterdam*
a.ciubucova@student.vu.nl

Lev Kondratchik
2856760
*Department of Artificial Intelligence, VU Amsterdam*
l.kondratchik@student.vu.nl

Anya de Saram-Larssen
2838616
*Department of Artificial Intelligence, VU Amsterdam*
a.c.d.saram-larssen@student.vu.nl

***Abstract* - The following study explores how different gendered voices (female and male), in a virtual hotel concierge bot, affect user satisfaction and perception in the context of a hotel. A within-subjects design was used with 8 participants (5 female and 3 male) in the real-world setting of a student hotel. Each participant interacted with both versions of the bot, which provided hotel-related recommendations and answered questions. Participants completed both pre- and post-surveys to rate the two bots on different attributes to gauge which one was preferred overall. Although the female-voiced bot was preferred, the statistical analysis, using the Wilcoxon signed-rank test, showed that there was no significant difference between the two conditions. This suggests that even though the participants showed a clear preference, the small sample size is not enough to generalise the findings to a larger population. However, our study still shows the potential of how qualities like the voice gender of a bot can influence user experience and improve human-computer interaction. Future research would benefit from a larger and more diverse sample size.**

***Keywords - Furhat bot, female voice, male voice, virtual hotel concierge, user perception, user satisfaction, human-computer interaction.***

## I. Introduction

This study examines how a virtual concierge agent can improve guests' experience in hotels while enhancing operational efficiency for the staff. The idea proposes to implement a Furhat-based agent that supports everyday tasks of the guests, but more specifically gives recommendations about local places that can be visited and answers questions the guests might have. The agent will be working in the hotel lobby, using natural speaking skills, making the interaction feel more natural. This is also a helpful option for customers who prefer personalised self-service for small problems. The issues we want to improve include inconsistent service, outlining cases such as busy times, small staff groups and guests' reluctance to seek help. These factors are often seen in conditions such as staff getting overwhelmed during peak hours, or just limited staff, especially during big events or the summer season. The main aspect we want to measure is how differently people will react to whether the bot has a male voice or a female voice. We want to find out whether there is a significant difference in how the guests will perceive 2 types of voices. So overall, we will have 2 different experimental conditions: a model with a female voice and a model with a male voice. The evaluation will include comparing all of the versions of the agent, measuring factors such as guest satisfaction, agent helpfulness and willingness for future use.

One of the studies explored how customers perceive chatbots in the hospitality industry, e.g., hotels. It concluded that while most respondents were unaware of chatbots or AI, younger ones were more open to using that technology. Although people appreciated convenience and the fact that they could contact it at any point in time, they were still more willing to contact real people for more complex problems. A key finding was that a lack of trust in AI or chatbots, as well as a lack of awareness, were the most important factors in why people did not want to adopt them [1].

Another study investigated how a chatbot's gender (male vs female) and the warmth of its language (high vs low) influenced users' perceptions of trust, helpfulness, and competence. Despite expectations based on stereotype theory, the results showed no significant effects of either gender or language warmth, nor their interaction, on any of the outcomes. Participants did not perceive female chatbots as more trustworthy or helpful, or male chatbots as more competent [2].

One more study explored how AI-powered, voice-based digital assistants affect the hotel service and user experience. Although adoption was limited, the findings showed that both guests and hoteliers recognised clear benefits: improved service efficiency, all-day accessibility and labour costs. Guests also appreciated the fact that they could control room settings or request services through natural conversations. On the other hand, there were some concerns about privacy, language limitations and system integration challenges. So the researchers propose that further potential improvement could be multilingual support or improved user education [3].

The last paper included in this literature review investigates how the perceived gender of a chatbot influences users' satisfaction and the activation of gender stereotypes. Through a controlled experiment using male, female, and non-gendered chatbots in both gender-neutral (banking) and gender-stereotypical (mechanics) domains, the researchers found that

users were more satisfied with non-gendered chatbots overall. However, when chatbots operated in stereotypically male domains, like mechanics, female chatbots received more gender-stereotyped evaluations, suggesting users apply biases more readily in such contexts [4].

The research question that guides this study is: Does the gender of the bots' voice and appearance affect user satisfaction and perception in a hotel context? We hypothesise that there will be a significant difference in how participants perceive the male-voiced bot versus the female-voiced bot, particularly in terms of friendliness, trustworthiness, adaptability and competence. Our null hypothesis is that there will be no significant difference in overall user satisfaction and preference between the two bots.

## II. Methods

### a. Study Design

This study utilised the within-subjects design, where each participant interacted with both conditions of the Furhat virtual hotel assistant: one with a male and one with a female voice, using neutral face to minimise facial bias. This is informative because it allows us to compare how one individual reacts to each version, instead of comparing different people's opinion in a between-subjects design. Since all participants experience each condition, variables such as a person's age, technology experience, personality etc., do not have as big of an impact on the results. The order of the conditions was intended to be randomised for each participant, meaning that some started with the male voice, and others with the female voice. This was to reduce order effects and ensure that any preference the participant showed towards the bot was due to the voice and not just simply the order in which it was presented. However, in practice most participants ended up choosing the female voice first. The independent variable is the gender of the agent's voice (female vs male) and the dependent variable was the participants preference and perception of the bot.

### b. Uncontrolled Variables

The study itself took place in a student hotel to closely mimic the environment of a real hotel and to recruit actual hotel guests as participants, and not just college students. This significantly increased the external validity of the results as it allowed for more realistic and natural reactions. However, conducting this study in a real-world setting allowed for a few uncontrolled variables, thereby decreasing the internal validity. Since the study did not take place in a controlled lab environment, background noise (conversations, lobby music, or air conditioning etc.) could have distracted participants during their interaction with the bots. Hotel Wi-Fi could have caused interaction lag if the signal was weak as well. Lighting conditions may have also affected how the visual appearance of the bots was perceived and could have unknowingly led to bias in how the voices were interpreted. As most of the study took place in a hotel lobby, the temperature may have been uncomfortable, affecting the overall mood and comfort levels of participants. Additionally, the mood of the participant pre-study might have also affected how they evaluated the bots. For example, if a participant was already irritable before the study, they might rate both bots less favorably, independent of the voice used. Prior experience and personal bias could have also played a role. If a participant had more encounters with male or female concierges/assistants in the past, they might be inclined to prefer one voice over the other. Finally, the participants' comfort level with technology could have made them more or less confident during the interaction, which may have influenced how positively they rated each bot.

Even though the environment could not be fully controlled, we recognise that several steps could have been taken to reduce the impact of these uncontrolled variables by keeping the setting consistent across all participants. For example, the bot (i.e. the laptop) could have been positioned in the same place so that all sessions were conducted in the same part of the hotel lobby to control for lighting and background noise. However, a measure we did take is that all participants were given the same standardised instructions before interacting with the robot to ensure they all had a consistent experience. There were still factors that we had no control over, such as the participants' mood and personal bias, which is a limitation of this study.

### c. Apparatus and Materials

Below is a list of all material and apparatus that was used during the experiment.

- Furhat robot: This was the main application through which participants interacted with both the male and female version of the virtual hotel concierge.
- Laptop: A laptop was used to run and display the Furhat robot interface.
- Consent form: A printed consent form was given to all participants prior to the start of the experiment. This form introduced all rules of the testing and confirmed that participants gave explicit permission for their data to be used in the study, and that they understood that they had the right to withdraw at any point throughout the experiment.
- Pre/Post Interaction surveys: Digital surveys were given before and after the experiment and each participant was required to answer them. This was crucial so that we could gather data about the demographics of the participants and to collect structured responses on which bot they preferred and how they rated the bot across different attributes.

d. Data Collection and Measurements

The target group for this study was hotel guests aged 18-65 with a basic grasp of technology. Participants were recruited through convenience sampling. This was done by approaching guests who were willing to take part in the experiment in the hotel lobby. Even though the target range was broad, the final sample consisted of 8 guests aged 18-25, 5 of which were female and 3 of which were male. This younger demographic can be explained by the fact that the experiment was run in a student hotel. 8 participants was considered reasonable enough of a sample size for this study, especially since a within-subjects design was used, meaning each participant underwent each condition and served as their own control. Data was collected using digital surveys, and Likert scales were used to measure participants' responses. A pre-survey was required by participants to fill out which collected demographic and background information including:

- Gender
- Age
- Frequency of hotel visits
- Comfort level with technology
- Prior experience with chatbots and virtual assistants
- Expectations for the interaction

After interaction with each bot, participants were required once again to fill in a digital post-survey. They rated both the female voices bot and male voiced bot on the following aspects:

- How well the bot understood their requests
- How adaptable the bot's responses felt
- Smoothness and naturalness of the interaction
- Likelihood of using a similar bot in future hotel visits
- Trustworthiness
- Usefulness of recommendations
- Which bot felt more natural and smooth
- Competence and reliability
- Accuracy in understanding input
- Friendliness and personability
- Overall preference

Each of these questions contributed to determining our core research question of whether the voice of the bot affects user satisfaction and perception in a hotel context. The participants answering these questions helped us determine which voice they preferred overall. These questions were answered using a Likert scale, allowing us to compare their feedback in a structured and organised manner.

## III. RESULTS AND DISCUSSION

a. Pre-interaction results

In order to understand better the sample we assembled, we decided to survey participants not only after the actual experiment (the interaction with the bot), but before it as well, so we knew how familiar they were with technologies, how often do they use the hospitality services, and what were their expectations from the bot they agreed to interact with.

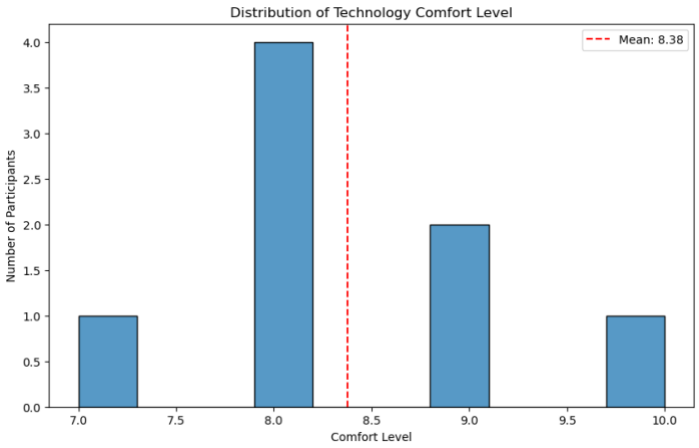


*Figure 1: Graph showing the distribution of users' comfort level*

Figure 1 shows that everyone except one participant reported frequent use of AI, and all rated themselves as mostly confident with technology (mean comfort score: 8.38/10). When asked about their expectations, most participants expressed interest in helpful, responsive, and accurate recommendations. This suggests that participants expected high-quality performance from the bot.

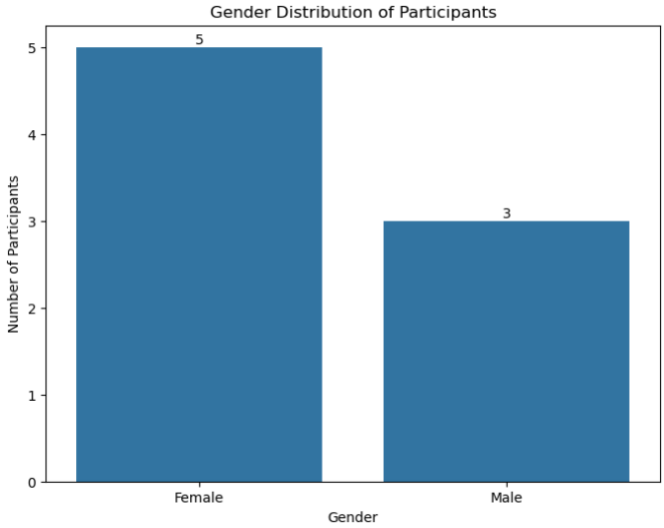


*Figure 2: Graph showing the distribution of genders in the sample.*

Figure 2 shows the gender distribution within the sample. Since we used convenience sampling, we were unable to ensure a fully balanced gender distribution. This is a known limitation and will be considered when interpreting the results, as it may have introduced bias.

In terms of travel hospitality experience, three participants reported staying in hotels more than five times per year, while the remaining five indicated staying between one and five times per year. This suggests that the majority of participants had moderate experience with hotel environments, making them reasonably familiar with typical concierge services and able to provide informed feedback on the virtual agent's performance.

b. Post-interaction results

Across both versions of the bot, participants rated their experience highly:

- Understanding requests: Mean = 4.25 (out of 5)
- Adaptability to input = 4.13
- Natural and smooth interaction = 4.13
- Willingness to use again = 4.25

This suggests that, independent of voice gender, the system was mostly perceived as natural and usable. While individual scores ranged from 2 to 5, the overall tendency was positive. Participants also agreed on using the system in future hotel stays, which is a decent indicator of user acceptance.

| CATEGORY | FEMALE VOICE | MALE VOICE |
|---|---|---|
| Trustworthy | 5 (62.5%) | 3 (37.5%) |
| Recommendations | 5 (62.5%) | 3 (37.5%) |
| Natural Smooth | 5 (62.5%) | 3 (37.5%) |
| Competent | 3 (37.5%) | 5 (62.5%) |
| Understands | 6 (75.0%) | 2 (25.0%) |
| Friendly | 5 (62.5%) | 3 (37.5%) |
| **Overall** | 6 (75.0%) | 2 (25.0%) |

*Figure 3: Table showing the participants' preferences between female and male bots in different aspects.*

When it comes to comparing the male and female versions of the bot, the first thing that needs to be clarified is that, since the preference-based questions only offered two response options, no data filtering or outlier removal was necessary. All participants completed both interaction conditions and provided valid answers for each comparison item.

Among the eight participants, 75% (6 out of 8) preferred the female voice overall. This trend continues in most individual attributes, including trustworthiness, recommendation quality, naturalness, and friendliness. The strongest preference was observed for "Understands input better," where 75% also selected the female agent.

On the other hand, if we look at the "Competent" category, it is the only category where male voice won, which is interesting considering that Baxter, McDonnell and McLoughlin paper (2018) also showed that when participants were asked whether male/female bots more competent in male-stereotyped domains like mechanics, most of them answered the male bot. This suggests that further research should be conducted to either gather more evidence supporting the notion that male bots are generally perceived as more competent or to determine whether this perception is context-dependent and influenced by domain-specific gender stereotypes.

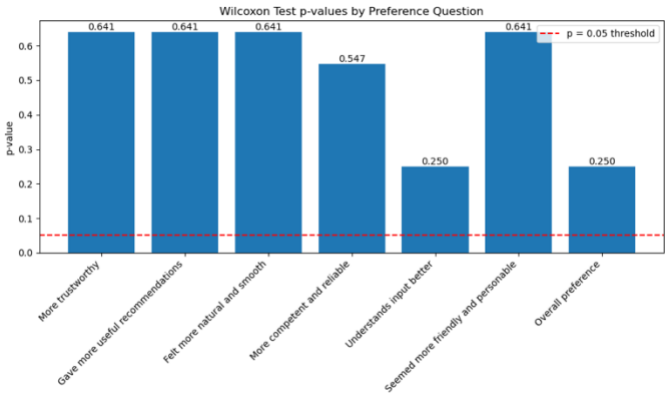


*Figure 4: Graph showing the Wilcoxon test p-values for each of the categories in the questionnaire, with the threshold of p=0.05*

Since we had a small sample and we also needed a non-parametric test because the results were not normally distributed to assess whether these differences were statistically significant, we used the Wilcoxon signed-rank test to compare user preferences, converting the female option into 1 and the male into 0 and therefore the median is 0.5. As shown in the bar chart of p-values, none of the results fell below the 0.05 significance threshold. The only two items that approached significance were "Understands input better" and "Overall preference", both with p-values of 0.25. All others exceeded p = 0.5, indicating no strong statistical evidence of preference bias between the two agent versions. Based on these results, we cannot reject the null hypothesis, so there is no significant difference in overall user satisfaction and preference between the two bots.

These results, however, on a small sample size, suggest a trend toward favouring the female agent. This, however, does not align with the outcome of the study by Bastiansen, Kroon and Araujo (2022), since their research showed that in general, there was no difference in how users perceived bots. However, in that

study, they used chatbots, so they did not have a fully animated head with an actual voice, the only difference between bots was the pronunciation (she/her or he/him) and the male/female avatar. This suggests that people can differently perceive different types of bots, not because of their initial predisposition towards a certain gender, but because of the way the bot looks and how it acts. In our study, the virtual concierge used a fully animated face and natural speech, which may have triggered more social and emotional cues compared to the simpler text-based bots used in the mentioned study.

### c. Limitations

Two of the limitations that were already mentioned are that the sample size was relatively small (n=8) and that the sample itself was not even in terms of gender (5 women and 3 men), therefore undermining the population validity. Additionally, the latter factor is arguably even more important, since our paper is heavily focused on gender, and considering the fact that the results showed a clear preference towards one of the bots, this could be affected by the sample effect.

Since we used the within-subject design, there is a certain possibility of the order effect. Because participants had already interacted with the bot once, during the second interaction with the bot (with a different gender), there was a chance of fatigue or that they could have already lost focus and simply decided then to answer the questions the same way as they answered them the first time.

Also, our bot used preprogrammed phrases, which can also be a potential weakness since there is a chance that participants associated specific phrases with one of the genders, so it could have been unusual for them to hear it from the opposite one. Besides, while participants provided binary and scaled ratings for various attributes of the virtual concierge, no open-ended responses or interview data were collected. As a result, we lack deeper insights into why participants preferred one version over the other or what specific aspects of the interaction shaped their impressions.

## IV. Conclusion

In conclusion, this study researched how the gender of a hotel concierge bot influences user satisfaction. Across eight participants, the female agent was generally preferred for attributes like friendliness, understanding, and overall impression, while the male agent was rated slightly more competent. Although these differences were not statistically significant, they suggest that users may respond differently to agents based on gender presentation, especially when the agent uses natural voice and animation.

Participants reported high satisfaction with the system overall, showing it could be useful in real hospitality settings. However, the small and gender-imbalanced sample, scripted dialogue, and lack of qualitative feedback limit the findings. Future studies should involve more participants and explore user preferences in more depth. Even so, the results give some ideas about how design choices like voice and appearance affect human-computer interaction.

## Statement of Contribution

The authors confirm their contribution to the paper as follows: bot programming and conducting of the experiment: Nia Racheva, Anastasia Ciubucova; data collection: N. Racheva, Anastasia Ciubucova; analysis and interpretation of results: Lev Kondratchik, Nia Racheva, Anastasia Ciubucova; draft preparation: Lev Kondratchik, Anya de Saram-Larssen. All authors reviewed the results and approved the final version of the manuscript.

## References


[1] Group, P. I. (2019). An Exploratory Study of Customer Perceptions of Usage of Chatbots in the Hospitality Industry. International Journal on Customer Relations. https://www.academia.edu/download/66041734/An_Exploratory_Study_of_Customer_Perceptions_of_Usage_of_Chatbots_in_the_Hospitality_Industry.pdf

[2] Bastiansen, M. H. A., Kroon, A. C., & Araujo, T. (2022). Female chatbots are helpful, male chatbots are competent? Publizistik. https://doi.org/10.1007/s11616-022-00762-8

[3] Buhalis, D., & Moldavska, I. (2021). In-room Voice-Based AI Digital Assistants Transforming On-Site Hotel Services and Guests' Experiences. Information and Communication Technologies in Tourism 2021, 30–44. https://doi.org/10.1007/978-3-030-65785-7_3

[4] Baxter, D. N., McDonnell, M., & McLoughlin, R. J. (2018). Impact of Chatbot Gender on User's Stereotypical Perception and Satisfaction. Electronic Workshops in Computing. https://doi.org/10.14236/ewic/hci2018.154