Your robot, my voice: enhancing android robot likability through personalization by cloning the user's voice

Johanna Magdalena Kuch, Marcel Heisler, Stina Klein, Silvan Mertes, Lennart Eing, Elisabeth André, Christian Becker-Asano

Angaben zur Veröffentlichung / Publication details:

Kuch, Johanna Magdalena, Marcel Heisler, Stina Klein, Silvan Mertes, Lennart Eing, Elisabeth André, and Christian Becker-Asano. 2025. "Your robot, my voice: enhancing android robot likability through personalization by cloning the user's voice." In *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 25-29 August 2025, Eindhoven, Netherlands*, edited by Emilia Barakova, Barbara Bruno, and Astrid Rosenthal-von der Pütten, 192–98. Piscataway, NJ: IEEE. https://doi.org/10.1109/ro-man63969.2025.11217611.

2025 34th IEEE International Conference on Robot and
Human Interactive Communication (RO-MAN)
August 25-29, 2025. Eindhoven, The Netherlands

Your Robot, My Voice: Enhancing Android Robot Likability through Personalization by Cloning the User's Voice

Johanna Magdalena Kuch[1], Marcel Heisler[2] Stina Klein[1], Silvan Mertes[1], Lennart Eing[1], Elisabeth André[1] and Christian Becker-Asano[2]

Abstract— **This study investigates whether personalized voice cloning can improve a robot's likability compared to a *design-congruent voice* and a distinctly *dissimilar voice*. Participants interacted with a gender-ambiguous android robot in three different voice conditions. We compared: (1) a *personalized voice clone* based on the participant's voice, (2) a *design-congruent voice* matching the robot's appearance, and (3) a *dissimilar voice*, which differs from both the participant's and the robot's features.**

The *cloned* and *design-congruent voices* significantly increased likability compared to the *dissimilar voice*, while anthropomorphism and familiarity showed no significant differences across conditions. Most participants did not immediately recognize their cloned voice until informed that one of the voices was a clone. However, most of the participants were successful when asked to pick out their cloned voice from those used. We assume that voice personalization through similarity to the user improves likability even before the user is aware of this similarity.

Our results show that personalized voice cloning is a simple alternative to other methods for the design of robotic voices. It significantly increases robot likability while requiring minimal user effort.

I. INTRODUCTION

In today's world, interaction with social robots is becoming increasingly important. A robot's voice is critical, as it strongly influences the mental image users form of the robot [1]. It affects users' willingness to approach the robot [2] and shapes essential perceptions such as human-likeness, naturalness, and emotional expression [3]. An unmatching voice can make the user feel uncomfortable or uncanny, leading to dismissive reactions [4]. Therefore, voice design must be carefully considered. But what is a suitable voice design?

Personalization of voices is known to have several advantages: Personalizing a voice improves its perceived quality [5]. In addition, matching voice characteristics to the user's characteristics can improve performance, perceived competence, perceived connection [6], and perceived social presence [7]. If no human voice recordings are used, this is referred to as speech synthesis. A major advantage of this is that no extensive recordings of real speakers are required. However, current methods for generating personalized synthetic voices are impractical for many human-robot interaction scenarios. The approach by Yarrington et al. [8] was developed for people who are at risk of losing their natural voice. The system creates a personalized synthetic voice by recording an individual speech inventory of the user, using the concatenation of phoneme pairs from these recordings for speech synthesis. This method requires extensive voice recordings, making it impractical for easily personalizing robot voices. For example, in a scenario where only brief interaction with a robot is required, the effort would be disproportionate.

Voice conversion or manipulation allows users to adjust the features of initial voices. This process benefits from achieving good satisfaction but requires fine-tuning from the user, making the system complex to use [9]. Easier, more user-friendly approaches like evolutionary sound design systems enable the refinement of voices iteratively [10], [11], they remain time-consuming. A more efficient solution is needed to minimize user effort while ensuring personalization.

An efficient approach to achieving a personalized voice might be voice cloning. Voice cloning builds on modern TTS synthesis systems but offers a key advance: the ability to replicate a speaker's vocal features from a short voice sample. The latest zero-shot approaches can now generate speech that matches the user's voice features without additional training or manual effort [12]. These developments make personalized voices accessible for real-world applications. The process is faster than manual adjustments and does not require additional time or effort from the user. Our study builds on these advances by investigating personalized voice design through voice cloning.

To evaluate this approach, we compare the likability of a user-similar *voice clone* with a *dissimilar voice*. The user-similar condition uses a cloned voice based on the participant's voice. The dissimilar condition features a voice selected from a pool of pre-generated voices to ensure a strong contrast with the user's voice. As a baseline, we include a *design-congruent voice*, which has been previously identified as a suitable match for the robot's visual design [13]. If the user-similar voice clone achieves equal or higher likability than the *design-congruent voice*, this indicates a promising alternative. Personalization based on similarity to the user could then replace more complex or time-consuming voice design methods in scenarios where a single user interacts with a robot.

Our study uses a gender-ambiguous android robot head [14]. This robot was chosen to avoid gender associations. The ambiguous design makes sure that all voice types, especially the *personalized clones* and distinctly *dissimilar voices*, can

[1]Chair for Human-Centered Artificial Intelligence, University of Augsburg, 86163 Augsburg, Germany
[2]Institute for Applied Artificial Intelligence, Stuttgart Media University, 70569 Stuttgart, Germany

be tested equally. In this way, the robot head is similarly suitable for male and female voices, allowing us to compare the different voice conditions.

Users attribute personality traits to synthetic voices, triggering a *Similarity-Attraction Effect* [15], where perceived similarity enhances likability. As similarity may also foster familiarity [16] and anthropomorphism [17] which are both positively linked to likability [18], [19], [20], we additionally examined perceived anthropomorphism and familiarity alongside likability.

In this work, we investigate how personalized voice cloning influences the perceived likability of an android robot head.

II. RELATED WORK

A. Likability, Anthropomorphism, and Familiarity in Android Robots

Humans prefer robots that communicate with speech [21]. Recent advances have made synthetic voices more similar to human speech but issues like limited pitch variation, can reduce perceived human likeness and likability [22]. Synthetic voices tend to be rated lower in anthropomorphism than natural voices [23]. To achieve likable voices, they should sound human-like [19]. Synthetic voices, in particular, benefit from human likeness [22]. Robots with more realistic voices tend to be perceived as more pleasant. They evoke less eeriness and are better accepted by users in various contexts. Additionally, they appear more anthropomorphic [24]. Anthropomorphic design enhances robot likability. It also increases perceived robot extraversion and agreeableness across different interaction scenarios [25]. Interestingly, perceived novelty is linked to anthropomorphism. The higher the novelty, the lower the anthropomorphism [20]. Therefore, it only makes sense that familiar robots are perceived as more convincing and trustworthy [26].

People create mental images of robots based on the robot's voices. They associate vocal features with the robot's appearance and function. Voice gender, naturalness, and accent shape these mental images [1]. Users create expectations about a robot's physical attributes based on its voice, for example human-like voices lead to expectations of human-like traits [27]. However, it is difficult to create a suitable voice design based solely on the appearance of a robot. The interaction context also determines how appropriate a voice is perceived [28]. Congruent designs improve consumer engagement [29]. Trust in robots is also shaped by the congruence between the first impression and the actual behavior of the robot [30].

B. Personalized Voice Design

Personalization has long been widespread in many fields of human-robot interaction, such as teaching [31], navigation [32], service [33] or even rehabilitation [34]. Voice design plays a crucial role in creating personalized user experiences. Therefore, we investigate a different approach to personalization in voice design. We create a similarity between the robot's and the user's voices. While personification by matching the robot's voice to the user's own, is a specific form of personalization, we refer to it simply as personalization throughout this paper.

The *Similarity-Attraction Effect* influences likability. People perceive robots as friendlier when recognizing similarities [35]. Same-gender robot voices increase acceptance and psychological closeness [17]. Applied to voices, users process speech more efficiently when it resembles their own [36]. TTS voices customized to the user are rated more credible and engaging [15]. Due to these numerous advantages of personalization and the potential we see in voice design, we want to investigate the approach of using **voice cloning** of the user's voice for a robot as a voice design. The robot could be more likable by using a voice that is very similar to the user's. Voice cloning offers a quick and straightforward approach to creating voices that match the user's. Voice cloning has rarely been used for robotic voices, as previous methods required extensive training data and computing power. First zero-shot models such as XTTS [12] overcome these challenges.

III. METHODOLOGY

This section describes the voices used in our study and explains the interaction between study participants and the robot. We compared three different voice conditions: (1) a personalized clone of the user's voice, (2) a design-congruent voice matching the robot's appearance, and (3) a voice that differs from both, the user's voice and the robot's appearance. We designed this setup to measure the effects on each interaction's perception of likability, anthropomorphism, and familiarity.

A. Participants

We recruited 50 participants from a university setting (M = 24.7 years, SD = 4.46, range: 18–38). Each participant received 10€ compensation. 27 of them were male, 23 were female.

B. Voice Cloning

To systematically examine the impact of voice personalization, we used XTTSv2 [12], a zero-shot voice cloning system capable of generating highly realistic speech from minimal input data. The model's architecture consists of three core components: (1) a variational autoencoder-based audio encoder that compresses mel-spectrograms into a discrete latent space, (2) a transformer-based autoregressive decoder that predicts the encoded speech tokens conditioned on text input, and (3) a HiFi-GAN vocoder that reconstructs waveform audio with high temporal and spectral quality. Before deployment in the user study, we systematically optimized key synthesis parameters. We recorded a dataset from six native German speakers (three male, three female), each reading ten sentences from the *Oldenburger Satztest* [37]. The *Oldenburger Satztest* was chosen because its balanced phoneme distribution and natural language structure should provide a good basis for creating German voice clones. We evaluated

synthesis quality using standard metrics: Character Error Rate (CER), Speaker Encoder Cosine Similarity (SECS), and Naturalness Mean Opinion Score (nMOS).

Our evaluation showed that using five reference sentences per speaker provided the best trade-off between speaker similarity and synthesis stability, it also supported the use of the parameters from [12] for speech synthesis and found the following parameters to work well for creating voice clones: $gpt\_cond\_len = 30$, $gpt\_cond\_chunk\_len = 4$, and $max\_ref\_length = 60$. With these settings the measured SECS ranged between 0.66 and 0.80 on average per speaker. The nMOS varied between 2.62 and 3.41. CER was between 0.36 and 1.03, though two speakers had higher values (up to 5.67) due to occasional artifacts in short utterances. For the *design-congruent* voice, we achieved a CER of 0.49, an nMOS of 3.38. We fine-tuned the XTTS encoder on 13 minutes of audio generated with the robot voice to increase SECS from 0.55 to 0.66. While longer reference recordings improved speaker fidelity, excessive input length introduced synthesis artifacts.

1) Personalized Cloned Voice: For each participant, a *personalized voice clone* was generated based on their recording of five sentences from the *Oldenburger Satztest* [37].

2) Dissimilar Voices: To create voices for comparison, we pre-generated six distinct voices using VoiceX [11], an evolutionary voice design tool that generates synthetic voices through random parameter initialization and iterative user selection. As a starting point, VoiceX generates a synthetic voice with randomized settings. The user can then refine the voice through a series of binary choices. We created six different voices by repeatedly selecting the variant that was subjectively the most gender-specific, resulting in three male and three female-sounding voices. Due to the gender-ambiguous design of the android robot head, it was possible to contrast these strongly gendered voices without a bias introduced by the robot's appearance. These *dissimilar voices* neither matched the robot's gender-ambiguous appearance nor resembled the participants' voices. To achieve this, we primarily used gender difference as the basis for the assignment: participants with deeper, male-sounding voices interacted with the robot speaking in a distinctly female-sounding voice in this condition and vice versa. Although we did not systematically assess the acoustic similarity between each participant's voice and the assigned voice, we ensured a clear perceptual difference through strong gender contrasts. Using this *dissimilar voices* allowed us to compare the effects of both design strategies, design congruence and *personalized voice clone*, against voices that did not meet either criterion.

3) Design-Congruent Voice: For the design-congruent voice condition, we used a voice originally developed by van Rijn et al. [13] as part of a larger research initiative on the creation of robot voices by humans in the loop. This combined crowd-sourcing, iterative user feedback, and perceptual modeling. The authors identified voice characteristics that are perceived as a good match for certain robot appearances. This resulted in voices for many robots that were design-congruent to the robots' appearance.

Although our robot head differs in embodiment (we used only the head and not a full-body android), it shares the same facial features, including the same wig, with the robot (Andrea) used in the study by van Rijn et al. Since only an image of the full-body android's head was used in the mentioned work, we assume that the design-congruent voice is transferable to our setup. We conducted an online pre-study with 49 MTurk participants to assess the gender ambiguity of the robot head used in our study versus the version used by van Rijn et al. [13]. Participants rated each version on perceived gender (female, male, neutral) and gender specificity (5-point Likert scales). Gender specificity was computed as the absolute difference between male and female ratings, weighted by how non-neutral the robot appeared. Both versions were seen as gender ambiguous, with both robots rated as similar neutral (Median = 3) and similar gender-specific (Median = 2).

To ensure comparable audio quality across all voice conditions and to adapt the originally English *design-congruent voice* for our German-language study, we cloned it using the previously described voice cloning system.

C. Robot Setup

For the study, we used a custom-designed android robotic head by the Japanese company A-Lab [14], as shown in Figure 1. The robot head features 14 pneumatic actuators. For animating the robotic head according to speech input, we implemented the approach proposed by Heisler et al. [14], [38]. This method utilizes a deep learning-based technique to generate lip-synchronized facial movements directly from an audio signal. The mesh-based approach maps predicted facial deformations onto the robot's actuators, allowing realistic mouth movements synchronized with speech. The software controlling this robot is described in [39] and was extended to support the required voice cloning capabilities. The robot's actuation system is powered by an air pressure mechanism, which means that no motor noises disturbed the participants' perceptions during the study.

D. Study Setup

Participants interacted individually with a gender-ambiguous android robot head [38] positioned on a freestanding table in a quiet room. Participants were seated directly opposite the robot at eye level to simulate a natural conversational scenario (Figure 1). At the beginning of the study, participants wore headphones with an integrated microphone input to record their voices. The study conductor remained unobtrusively present throughout, operating the robot via Wizard-of-Oz to ensure precise control over dialogue timing and synchronization.

E. Study Procedure

Initially, we provided the participants with an overview of the study. We told them they would interact with a robot in three different scenarios with different voices we had developed but did not inform them that one of these

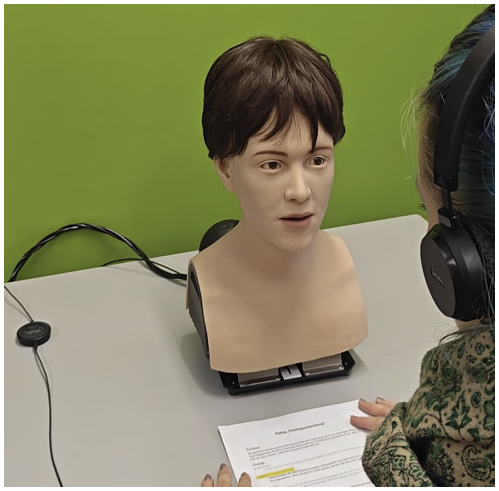

Fig. 1. Interaction between Robot and Participant

voices would be a clone of their own. Before proceeding, participants signed a consent form, which clarified that their participation was voluntary and that they could withdraw from the study without consequences.

First, participants were asked to read five sentences from the *Oldenburger Satztest* [37] aloud to create the *personalized voice clone* as described previously. Participants were not told that those sentences were used to create a voice clone. Instead, they were told that the recording was used to ”prepare the interaction.” This was done to prevent any influence of the knowledge about an existing voice clone on the perception of the robot and the voice. Next, participants filled out the questionnaire regarding their demographic data. Simultaneously, the study conductor created the robot's dialogue lines for the cloned voice. Because of artifacts possibly appearing in the audio files while generating the cloned dialogue lines, the study conductor checked for such artifacts. In case any major artifacts were present, the specific line(s) was regenerated.

Then, participants engaged in three interactions with the robot. All interactions were held constant, except for the voice, which varied according to the previously described voice conditions (see Section III-B). The order of the conditions was randomized. A scripted simulated dialogue [40] was used for the interaction. The dialogue revolved around preparing a meeting. The robot played the role of a colleague, and the participant talked about planning and missing documents. The dialogue was designed such that the robot and the participant had clearly defined roles that allowed comparable interaction between the three voice scenarios. After each interaction, participants completed a questionnaire assessing their perception of likability, anthropomorphism, and familiarity with the robot (see Section III-F).

At the end of the study, all three voice variants were repeated, and participants were asked whether they recognized that one of the voices was a clone of their own. Afterward, they tried to identify which voice they believed had been generated to match their voice and were asked to describe their perception of each voice using a free text field. Finally, we thoroughly debriefed the participants about the study's true purpose. We explained that one of the voices had been a clone of their own and that the study aimed to investigate the effects of personalized voices with voice cloning.

F. Evaluation Methods

Likability and anthropomorphism were measured using the respective subscales of the Godspeed questionnaire, each consisting of five items rated on a 5-point Likert scale [41]. Familiarity was assessed using two items adapted from prior research on novelty perceptions of synthetic voices [20]: ”Unknown/Well-known” and ”Unusual/Usual.” The familiarity score is calculated as the average of these two items.

Moreover, the *perceived gender* of the robot was assessed using three items: female, male, and neutral. Additionally, participants rated *voice suitability*, how well the voice suited the robot on a 5-point Likert scale. A Yes-No-Question was used to measure the participant's recognition of their voice clone. If they replied with *yes*, they were asked in an open question *”what made them recognize their voice.”* Then, they were asked to identify the interaction session containing their voice clone. To evaluate the perception of the voice clones, participants also described each voice they had heard in free-text responses. Finally, participants answered questions regarding their *prior experience* with text-to-speech (TTS) technology and android robots.

G. Analysis Plan

We planned to investigate the influence of the independent variable (IV) *voice* with its three levels: (1) *personalized voice clone*, (2) *design-congruent voice*, and (3) *dissimilar voice*, on the dependent variables (DVs) likability, anthropomorphism, and familiarity. Therefore, we planned to test for statistical significance using a one-way MANOVA with *voice* as the IV and likability, anthropomorphism, and familiarity as the DVs. If the MANOVA shows statistical significance, we perform separate post-hoc tests for each DV in separate one-way ANOVAs with a Bonferroni correction applied across all three tests. Tukey HSD tests will be conducted for pairwise comparisons between all three *voice* levels if any ANOVA is statistically significant.

IV. RESULTS

In this section, the quantitative and qualitative results of the survey will be reported.

A. Descriptive Results

1) Likability, Anthropomorphism, and Familiarity: Figure 2 gives an overview of the descriptive results. For likability, the *personalized voice clone* ($M = 3.61$, $SD = 0.79$) received the highest rating, closely followed by the *design-congruent voice* ($M = 3.57$, $SD = 0.70$), with the *dissimilar voice* rated noticeably lower ($M = 3.18$, $SD = 0.84$). Regarding anthropomorphism, ratings were similar across all voice conditions, with the *dissimilar voice* slightly higher ($M =$

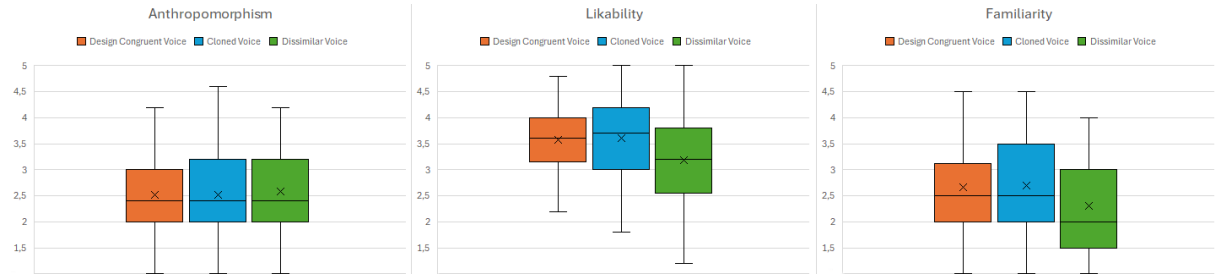

Fig. 2. Likability, anthropomorphism, and familiarity across the three conditions *Design-Congruent Voice*, *Cloned Voice*, and *Dissimilar Voice*.

2.59, *SD* = 0.72), followed closely by the *personalized voice clone* (*M* = 2.52, *SD* = 0.78) and the *design-congruent voice* (*M* = 2.51, *SD* = 0.81). For familiarity, the *personalized voice clone* received the highest rating (*M* = 2.70, *SD* = 0.96), closely followed by the *design-congruent voice* (*M* = 2.66, *SD* = 0.88), while the *dissimilar voice* was rated least familiar (*M* = 2.31, *SD* = 0.86).

2) Voice Suitability: Participants evaluated the suitability of each voice condition (design-congruent, cloned, and dissimilar) regarding how well they matched the robot's appearance. The *design-congruent voice* was perceived as the most suitable (*M* = 4.00, *SD* = 1.18). The *personalized voice clone* received a moderate suitability rating (*M* = 3.00, *SD* = 1.33). The *dissimilar voice* was rated lowest in suitability (*M* = 2.26, *SD* = 1.24).

3) Clone Recognition: Out of 50 participants, when asked if they recognized one voice as their voice clone, eight explicitly mentioned they recognized that one of the voices was a clone of their voice. When asked which voice condition contained their cloned voice, 7 identified the right voice. When knowing that a voice clone was involved, the majority of 38 participants correctly identified the cloned voice. Six participants incorrectly attributed their cloned voice to the *design-congruent voice*, one chose the *dissimilar voice*, and the remaining five were unsure.

B. Inferential Statistical Results

A repeated measures MANOVA was conducted to examine the effect of *voice* on likability, anthropomorphism, and familiarity. The analysis revealed a statistically significant impact, $F(6,44) = 2.72, p = .025$, indicating that the different voice conditions significantly impacted the DVs. As a follow-up, separate repeated-measures ANOVAs were performed for each DV with a Bonferroni correction. This set the significance level to $\alpha_{corrected} = .0167$.

A repeated measures ANOVA revealed a significant effect of *voice* on likability, $F(2,98) = 5.97, p = 0.004$. Subsequent post-hoc comparisons showed that the comparison between *design-congruent voice* and *dissimilar voice* was significant with $p = 0.024$, and the comparison between *personalized voice clone* and *dissimilar* was also significant with $p = 0.025$. No significant difference was found between the *design-congruent voice* and *personalized voice clone* ($p = 0.936$).

A repeated measures ANOVA was conducted to examine the effect of *voice* on anthropomorphism. The analysis did not yield significant results, $F(2,98) = 0.234, p = 0.792$, indicating that *voice* did not significantly influence the anthropomorphism ratings. Another repeated measures ANOVA was conducted to assess the effect of *voice* on familiarity. The results indicated a p-value of $p = 0.023$. Applying the Bonferroni correction for multiple comparisons ($\alpha_{\text{corrected}} = 0.0167$), this effect did not reach statistical significance, $F(2,98) = 3.92, p = 0.023$.

C. Qualitative Results

The *personalized voice clone* received predominantly positive feedback and was frequently described as *pleasant*, *warm*, and *familiar*. Participants perceived it as notably *friendly* and *natural*, describing it, for example, as "polite, open, interested, natural" and "pleasant to listen to and easy to understand." Another participant emphasized its realistic quality, noting that it "sounds realistic, like a voice in a phone call." At the same time, someone else described it as "very feminine and friendlier than the other voices." The cloned voice was also perceived as especially fitting for the robot: "Fits the robot best. Feels familiar and pleasant." Despite this largely positive reception, some participants pointed out minor drawbacks related to *unusual intonation* or a slightly mechanical impression, stating, for instance, "The intonation takes some getting used to." Only five out of 50 participants recognized their own voice from the beginning before being asked if they realized they were cloned. One, for example, mentioned: "Sounds like my own voice. Pitch and emphasis are similar to mine, making it very familiar, but also a bit eerie."

The *dissimilar voice* received a lot of negative feedback, described by participants as *unnatural*, *monotonous*, and overall unsuitable for the robot. However, some participants still recognized the qualities of the voices and suggested contexts in which it might be more suitable, such as "a storyteller," describing it as "inviting and warm," or as a "film dubbing actor." Another participant emphasized its cinematic quality, noting it was "somewhat unrealistic and cinematic," fitting the voice of a "cowboy character."

The *design-congruent voice* was predominantly described as *neutral*, *pleasant*, and *fluid*, and participants perceived it as appropriate for the robot, stating it "fits the robot

and feels familiar," and noting it "speaks fluently without interruptions." Some even described the voice as "could almost be human." Despite these generally positive impressions, participants mentioned minor drawbacks related to artificial or monotonous aspects, describing it as "relatively choppy," "stiff, metallic, unnatural," or suggesting that it "lacks expressiveness."

V. DISCUSSION

Our results show a significant effect of *voice* on the likability of the robot. Both a *personalized voice clone* and a *design-congruent voice* were rated significantly more likable than a *dissimilar voice*. There was no significant difference between a *personalized voice clone* and a *design-congruent voice*. Therefore, we conclude that voice cloning can effectively be used as a robot voice design method in one-on-one interactions between humans and robots. We found a trend regarding familiarity, but this did not remain significant after correcting for multiple comparisons. However, qualitative feedback showed that participants often perceived the personalized voice clone as particularly familiar and natural.

This research evaluates personalized voice cloning as an alternative to robot voice design, especially its influence on likability. Our findings show that personalized voice cloning achieves likability ratings comparable to those obtained through carefully designed, *design-congruent voices* [13]. Both *personalized voice clone* and *design-congruent voices* are significantly more likable than the *dissimilar voices*. Additionally, qualitative feedback underscored the familiarity and naturalness of the cloned voices. The approach we describe simplifies the creation of likable voice design through personalization by taking advantage of the *Similarity-Attraction Effect* [35], [15] and achieving similarity to the user directly through voice cloning instead of an elaborate design process [8], [9], [10], [11].

In contrast to previous studies [42], [43], we found no significant difference in anthropomorphism between the voice conditions, although they were differently well matched to the robot's visual features. Therefore, anthropomorphism might depend more on visual or behavioral congruence than on voice characteristics. These results contrast with previous findings [30], which emphasized multimodal congruence in robot design. The results of this study show that there might be cases where such congruence is not necessary. This could be due to the robot we used, which is very anthropomorphic by itself, or because we used human-like voices in every condition, and only bigger differences would lead to the effect described in previous studies. However, the topic needs further investigation.

A. Limitations

Our study has several limitations. While we aimed to minimize artifacts of the XTTSv2 voice cloning system, occasional artifacts may still have influenced user perceptions. We did not formally assess intonation or prosodic similarity between voices, which may have impacted comparability. Future studies might investigate if voice similarity enhances engagement, as demonstrated in related research on prosodic entrainment with children [44]. The short, scripted meeting-planning scenario limits generalizability to other interaction contexts or emotional scenarios. Also, voice cloning may become unfeasible in scenarios involving multiple interaction partners, which raises the question of whose voice should be cloned. Although the gender-ambiguous android head minimized biases related to voice-gender congruence, results may not generalize to robots with clearly gendered or distinct appearances. Furthermore, our setup did not account for other influential social characteristics such as age or ethnicity. Finally, our approach relied on pre-generated voices; real-time voice cloning without human verification still faces technical challenges, such as latency and artifacts.

VI. CONCLUSION

In this study, we explored personalized voice cloning as a novel method to enhance the likability of Android robots. Our approach utilizes the similarity between the robot's and the user's voices. Our results show that personalized cloned voices significantly increase the likability of interaction with the robot compared to voices incongruent with user or robot characteristics. However, we observed no significant effects on anthropomorphism or familiarity. Future research should test personalized voice cloning with different robot platforms and virtual agents. Additionally, combining cloned voices with appearance-congruent features could help overcome existing limitations. Also we plan to extend our approach by including other dimensions of voice identity (e.g., age, prosody, accent) and by directly measuring acoustic or perceived similarity. Personalized voice cloning opens new pathways towards effortless user-specific human-robot interactions.

ACKNOWLEDGMENTS

This work was supported by the Bavarian Research Foundation under the project "FORSocialRobots" AZ-1594-23.

REFERENCES

[1] C. McGinn and I. Torre, "Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots," in *HRI'19*, (Piscataway, NJ), pp. 211–221, IEEE, 2019.

[2] M. L. Walters, D. S. Syrdal, K. L. Koay, K. Dautenhahn, and R. te Boekhorst, "Human approach distances to a mechanical-looking robot with different robot voice styles," in *2008 17th IEEE International Symposium on Robot and Human Interactive Communication*, (Piscataway, NJ), pp. 707–712, IEEE, 2008.

[3] S. Ko, J. Barnes, J. Dong, C. H. Park, A. Howard, and M. Jeon, "The effects of robot voices and appearances on users' emotion recognition and subjective perception," *International Journal of Humanoid Robotics*, vol. 20, no. 01, 2023.

[4] W. J. Mitchell, K. A. Szerszen, A. S. Lu, P. W. Schermerhorn, M. Scheutz, and K. F. Macdorman, "A mismatch in the human realism of face and voice produces an uncanny valley," *i-Perception*, vol. 2, no. 1, pp. 10–12, 2011.

[5] Z. Shi, H. Chen, A.-M. Velentza, S. Liu, N. Dennler, A. O'Connell, and M. Mataric, "Evaluating and personalizing user-perceived quality of text-to-speech voices for delivering mindfulness meditation with different physical embodiments," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (G. Castellano, L. Riek, M. Cakmak, and I. Leite, eds.), (New York, NY, USA), pp. 516–524, ACM, 2023.

[6] D. Kao, R. Ratan, C. Mousas, and A. J. Magana, "The effects of a self-similar avatar voice in educational games," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CHI PLAY, pp. 1–28, 2021.

[7] N. Lubold, E. Walker, and H. Pon-Barry, "Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 255–262, IEEE, 3/7/2016 - 3/10/2016.

[8] D. Yarrington, C. Pennington, J. Gray, and H. T. Bunnell, "A system for creating personalized synthetic voices," in *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility* (A. Sears and E. Pontelli, eds.), (New York, NY, USA), pp. 196–197, ACM, 2005.

[9] H. J. Byeon, S. Ha, and U. Oh, "Avocus: A voice customization system for online personas," in *CHI'23* (A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, and A. Peters, eds.), (New York, New York), pp. 1–6, The Association for Computing Machinery, 2023.

[10] H. Ritschel, I. Aslan, S. Mertes, A. Seiderer, and E. Andre, "Personalized synthesis of intentional and emotional non-verbal sounds for social robots," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, (Piscataway, NJ), pp. 1–7, IEEE, 2019.

[11] S. Mertes, D. W. Don, O. Grothe, J. Kuch, R. Schlagowski, and E. André, "Voicex: A text-to-speech framework for custom voices."

[12] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, and J. Weber, "Xtts: a massively multilingual zero-shot text-to-speech model," in *Interspeech 2024*, pp. 4978–4982, 2024.

[13] P. van Rijn, S. Mertes, K. Janowski, K. Weitz, N. Jacoby, and E. André, "Giving robots a voice: Human-in-the-loop voice creation and open-ended labeling," 2024.

[14] M. Heisler, S. Kopp, and C. Becker-Asano, "Making an android robot head talk," in *Proc. of 32nd IEEE Int. Conf. on Robot and Human Interactive Communication (RO-MAN)*, pp. 1837–1842, 2023.

[15] C. Nass and K. M. Lee, "Does computer-generated speech manifest personality? an experimental test of similarity-attraction," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (T. Turner and G. Szwillus, eds.), (New York, NY, USA), pp. 329–336, ACM, 2000.

[16] R. L. Moreland and R. B. Zajonc, "Exposure effects in person perception: Familiarity, similarity, and attraction," *Journal of Experimental Social Psychology*, vol. 18, no. 5, pp. 395–415, 1982.

[17] F. Eyssel, D. Kuchenbrandt, S. Bobinger, L. de Ruiter, and F. Hegel, "'if you sound like me, you must be more human'," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (H. Yanco, ed.), ACM Conferences, (New York, NY), p. 125, ACM, 2012.

[18] K. Haresamudram, I. Torre, M. Behling, C. Wagner, and S. Larsson, "Talking body: the effect of body and voice anthropomorphism on perception of social agents," *Frontiers in robotics and AI*, vol. 11, p. 1456613, 2024.

[19] K. Kühne, M. H. Fischer, and Y. Zhou, "The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study," *Frontiers in neurorobotics*, vol. 14, p. 593732, 2020.

[20] J. M. Kuch, J. Nasir, S. Mertes, R. Schlagowski, C. Becker-Asano, and E. André, "Evaluating gender ambiguity, novelty and anthropomorphism in humming and talking voices for robots," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*, pp. 2219–2225, IEEE, 2024.

[21] N. Chatterji, C. Allen, and S. Chernova, "Effectiveness of robot communication level on likeability, understandability and comfortability," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1–7, IEEE, 2019.

[22] A. Baird, E. Parada-Cabaleiro, S. Hantke, F. Burkhardt, N. Cummins, and B. Schuller, "The perception and analysis of the likeability and human likeness of synthesized speech," in *Interspeech 2018*, (ISCA), pp. 2863–2867, ISCA, 2018.

[23] J. M. Kuch, F. Melchior, and C. Becker-Asano, "Effects of gender neutralization on the anthropomorphism of natural and synthetic voices," in *Proc. of 32nd IEEE Int. Conf. on Robot and Human Interactive Communication (RO-MAN)*, pp. 2080–2085, 2023.

[24] S. Schreibelmayr and M. Mara, "Robot voices in daily life: Vocal human-likeness and application context as determinants of user acceptance," *Frontiers in psychology*, vol. 13, p. 787499, 2022.

[25] A. S. Arora, M. Fleming, A. Arora, V. Taras, and J. Xu, "Finding "h" in hri," *International Journal of Intelligent Information Technologies*, vol. 17, no. 1, pp. 1–20, 2021.

[26] S. Saunderson and G. Nejat, "Robots asking for favors: The effects of directness and familiarity on persuasive hri," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1793–1800, 2021.

[27] M. de Cet, M. Cvajner, I. Torre, and M. Obaid, "Do your expectations match? a mixed-methods study on the association between a robot's voice and appearance," in *ACM Conversational User Interfaces 2024* (M. Dubiel, L. A. Leiva, J. Trippas, J. Fischer, and I. Torre, eds.), (New York, NY, USA), pp. 1–11, ACM, 2024.

[28] I. Torre and S. Le Maguer, "Should robots have accents?," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 208–214, IEEE, 2020.

[29] C. Ma, A. Fan, and S. A. Lee, "Unveiling the role of congruity in service robot design and deployment," *International Journal of Contemporary Hospitality Management*, vol. 36, no. 12, pp. 4150–4170, 2024.

[30] I. Torre, J. Goslin, L. White, and D. Zanatto, "Trust in artificial voices," in *Proceedings of the Technology, Mind, and Society*, (New York, NY, USA), pp. 1–6, ACM, 2018.

[31] P. Baxter, E. Ashurst, R. Read, J. Kennedy, and T. Belpaeme, "Robot education peers in a situated primary school study: Personalisation promotes child learning," *PloS one*, vol. 12, no. 5, p. e0178126, 2017.

[32] J. de Heuvel, N. Corral, B. Kreis, J. Conradi, A. Driemel, and M. Bennewitz, "Learning depth vision-based personalized robot navigation from dynamic demonstrations in virtual reality," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6757–6764, IEEE, 2023.

[33] N. Gasteiger, M. Hellou, and H. S. Ahn, "Optimizing human-robot interaction through personalization: An evidence-informed guide to designing social service robots," in *2021 18th International Conference on Ubiquitous Robots (UR)*, pp. 53–56, IEEE, 7/12/2021 - 7/14/2021.

[34] R. Feingold-Polak and S. Levy-Tzedek, "Personalized human robot interaction in the unique context of rehabilitation," in *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (J. Masthoff, E. Herder, N. Tintarev, and M. Tkalčič, eds.), (New York, NY, USA), pp. 126–127, ACM, 2021.

[35] E. P. Bernier and B. Scassellati, "The similarity-attraction effect in human-robot interaction," in *2010 IEEE 9th International Conference on Development and Learning*, pp. 286–290, IEEE, 2010.

[36] B. Chen, N. Kitaoka, and K. Takeda, "Impact of acoustic similarity on efficiency of verbal information transmission via subtle prosodic cues," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2016, no. 1, 2016.

[37] V. Kuehnel, B. Kollmeier, and K. Wagener, "Entwicklung und evaluation eines satztests für die deutsche sprache i: Design des oldenburger satztests," *Zeitschrift für Audiologie*, vol. 38, pp. 4–15, 1999.

[38] M. Heisler and C. Becker-Asano, "Learning to control an android robot head for facial animation," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (D. Grollman, E. Broadbent, W. Ju, H. Soh, and T. Williams, eds.), (New York, NY, USA), pp. 530–534, ACM, 2024.

[39] M. Heisler and C. Becker-Asano, "An android robot head as embodied conversational agent," in *56th International Symposium on Robotics* (ISR Europe, ed.), (Stuttgart, Germany), pp. 93–99, 2023.

[40] A. Glasbergen-Plas, S. Gryllia, L. P. Robles, and J. Doetjes, "Scripted simulated dialogue: a new elicitation paradigm," in *Speech Prosody 2022*, (ISCA), pp. 683–687, ISCA, 2022.

[41] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[42] G. Trovato, J. G. Ramos, H. Azevedo, A. Moroni, S. Magossi, H. Ishii, R. Simmons, and A. Takanishi, "Designing a receptionist robot: Effect of voice and appearance on anthropomorphism," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2015)*, (Piscataway, NJ), pp. 235–240, IEEE, 2015.

[43] B. Sarigul, I. Saltik, B. Hokelek, and B. A. Urgen, "Does the appearance of an agent affect how we perceive his/her voice?," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (T. Belpaeme, J. Young, H. Gunes, and L. Riek, eds.), (New York, NY, USA), pp. 430–432, ACM, 2020.

[44] I. Molenaar, "Towards hybrid human–ai learning technologies," *European Journal of Education*, vol. 57, no. 4, pp. 632–645, 2022.