Bachelor Thesis

Emotions in Voice Assistant: Balancing Humanization and Robotization in Visualization and Vocal Design to Meet User Needs

Jiahui Zhang 2859882

Communication Science

Faculty of Behavioral, Management and Social Science (BMS)

Supervisor: Dr. Joyce Karreman

University of Twente

Date: 2024.07.01

Abstract

Background

The rapid advancement of generative AI and voice assistants has expanded their potential applications, moving beyond straightforward task implementation to more complex conversational interactions. Design considerations, such as humanization through voice and visualization, are critical in shaping user experiences and emotional perceptions.

Purpose

This study aims to explore how emotional values perceived from voice assistants can be balanced between humanization and robotization. It investigates user acceptance, preferences, and the psychological and emotional reasons underlying the need for humanized voice assistants.

Method

The study employed a qualitative approach, utilizing prototype testing, concept evaluation, and exploratory interviews. Participants interacted with voice assistant prototypes featuring varying degrees of humanization, including facial expressions and voice characteristics. Data were collected through open-ended questions to capture diverse perspectives on humanized versus robotic elements.

Results

Findings indicate that users generally prefer humanized features in voice assistants, attributing emotional resonance to human-like voices and appreciating avatars with dynamic animations. Humanized design elements, such as facial expressions and tone, enhance user satisfaction and emotional engagement. However, the balance between humanization and robotization must be carefully managed to avoid negative reactions.

Conclusion

The study highlights the importance of integrating emotional support features into voice assistants. While users are open to advanced features, both functional and design-wise, future research should further investigate the impacts of highly realistic avatar animations and the psychological benefits of humanization. These insights provide valuable guidance for designers and engineers in developing emotionally supportive voice assistants that enhance user experience.

Keywords: Voice assistant, Humanization, Prototype test, User Experience, Emotional Perception.

Content

1. Introduction **5**

1.1 The functionality of Voice Assistant *5*

1.2 The user experience of VA *6*

2. Theoretical framework **9**

2.1 Visualisation of the Agent *9*

2.2 Vocal feedback and prosodic feature of Voice assistant *12*

2.3 Emotional Perception from Voice Assistant *13*

3. Methodology **15**

3.1 Research Design *15*

3.2 Instruments *16*

3.3 Interview Protocol *18*

3.4 Validity and Reliability of Protocols *23*

3.5 Participants *26*

3.6 Data Analysis *27*

4. Results **36**

4.1 Likability of four elements *36*

4.2 Attribute Perception Analysis *39*

4.3 Comparison of Emotional Perception of the Four Prototypes *50*

4.4 Humanization vs. Robotization: Preferences and Needs *52*

5. Discussion and Conclusion **58**

5.1 Main Findings *58*

5.2 Discussion *60*

5.3 Practical Implementation *62*

5.4 Limitation & Future Work *64*

5.5 Conclusion *65*

Appendix A: **70**

Appendix B: **79**

1. Introduction

Voice assistants (VAs) are increasingly becoming ubiquitous technologies, seamlessly integrating into users' lives. They serve various scenarios across multiple devices, whether as standalone smart speakers or embedded in other technology products such as smartphones, computers, cars, and advanced home appliances. According to Hoy (2018), voice assistant comes from the idea of basically interacting with our computers by talking with them. The internet holds immense power, and it would be incredibly valuable if users could interact with it through voice commands at any moment, regardless of their location or other limitations. This capability allows individuals to communicate with the internet even when their hands or feet are occupied, enabling seamless information sharing as long as they can speak. This high level of accessibility greatly expands the utility and reach of the Internet, making it widely accessible to a broader range of users.

1.1 The functionality of Voice Assistant

Voice assistants require internet connectivity to receive verbal input from users, analyze it, and provide feedback. Typically, people utilize voice assistants for a variety of simple tasks, such as playing music, setting alarms, or obtaining weather updates. These tasks save considerable time and effort, especially for individuals facing behavioural limitations. Additionally, users rely on voice assistants to address a wide range of queries, spanning from random to specific or professional inquiries, with the assistant tapping into the internet to deliver precise responses. Furthermore, the integration of voice assistants into car design has seen a surge in demand, reflecting the rapid evolution of the automotive industry. Beyond their practical functions, VA can also provide emotional or social support, such as serving as companions, engaging users in casual conversation, and providing leisure and entertainment.

Voice assistants are powered by a sophisticated array of technologies, including speech recognition, natural language processing (NLP), machine learning, internet connectivity, cloud computing, natural language generation (NLG), and robust privacy protection features. As AI continues to advance, integrating VAs with AI models has become commonplace, enabling VAs to exhibit more human-like characteristics such as more natural conversation, unique personalities, varied speech styles, and adaptive responses based on user data. Additionally, contextual understanding allows VAs to provide broader knowledge and discern user intentions more efficiently, propelling them toward the next level of technological advancement.

1.2 The user experience of VA

Designing a Voice Assistant generally involves two main categories: functional aspects and design considerations. Successful design requires a balanced focus on both areas. Different design styles within these elements can significantly influence user satisfaction, as users perceive and prefer designs that align with their expectations and preferences, even when the functional capabilities remain constant. It's essential to recognize that focusing solely on functionality, as mentioned above, may not suffice for VA products to achieve excellent user experience. It's equally crucial to enhance the interaction experience, considering that users are multifaceted beings with practical, emotional, and psychological needs. Fulfilling these needs is imperative for achieving superior user experience performance in VAs.

From a functional perspective, designers prioritize enhancing response accuracy and efficiency, the assistant's ability to understand commands and context, natural language processing, and ensuring privacy protection. On the other hand, design considerations primarily revolve around two elements: visualization and verbal feedback. Visualization can include an avatar or simple light indications, while voice feedback encompasses pitch, intonation, tone, and speed variations. Subsequently, one might ponder how user experience influences VA product interaction beyond technological aspects. From a design perspective, what considerations should user experience designers take into account to create a compelling and highly rated VA?

In terms of design, a Voice Assistant (VA) can either incorporate an avatar or agent or rely solely on voice interaction, similar to products like Amazon's Alexa and Apple's Siri. Study of Shamekhi et al. (2018) suggests that incorporating an avatar can positively influence users' perceptions, making the assistant seem more intelligent and trustworthy. This is because avatars with human-like characteristics can build rapport and enhance the user satisfaction. Additionally, effective design of a Voice Assistant's (VA) image can enhance user engagement and encourage continued usage of the product. For instance, scholars (Waytz et al., 2014) have discovered that employing text, images, videos, and voice to anthropomorphize virtual agents can significantly enhance the user experience. They also claimed when the VA agent provides vivid emotional expressions in its design, users are more likely to perceive a sense of humanity, further deepening their interaction with the assistant.

Therefore, both the visual and vocal features of voice assistants can significantly impact users' perceptions, ultimately influencing overall user experience satisfaction. Considering these two design elements, the study has formulated its main question into two research questions:

RQ1: How does the presence of an agent in the visual representation of a Voice Assistant (VA) affect users' perception emotionally and psychologically, ultimately impacting the User Experience of VA?

Visualization of a voice assistant is not necessary but more as a nice to have for most of the voice assistant products on the current market. It raises the question of how users might prefer such visualizations based on their needs. Are there any emotional or psychological benefits provided by the visualization design of a voice assistant which can increase the satisfaction and enhance the user experience of a voice assistant product.

RQ2: How does vocal feedback of voice assistants influence users' emotional perception and subsequently impact the User Experience of VA?

Similar to visualization but even more critical as the main element of a voice assistant, how does vocal feedback impact users emotionally and psychologically, and influence their willingness to use and overall user satisfaction of the voice assistant?

2. Theoretical framework

Variability in appearance among Voice Assistants, and more broadly among virtual assistants, can influence how humans perceive them. Acceptance of this diversity can contribute to successful robot design. Beer et al. (2017) advocate that Human-Robot Interaction research should prioritise finding the most effective design and development approaches for these robots. Additionally, Prakash and Rogers (2015) highlight that a robot's appearance can significantly impact user experience with the product. This underscores the importance, as noted by Beer et al. (2017), of aesthetics in shaping socially assistive robot faces and forms, as users' initial impressions can shape their expectations and acceptance.

2.1 Visualisation of the Agent

Voice assistants are evolving beyond mere internet-connected devices that respond to queries with voice. With the integration of visual interfaces or "faces," users can interact with these assistants in ways that extend beyond vocal feedback alone. By incorporating a visual component, users gain access to a more diverse range of interactions, enhancing the overall user experience and making interactions with the assistant more intuitive and engaging sometimes. It’s interesting to find out that voice assistants are more commonly visualised in automobiles than in home appliances. While smart speakers from companies like Amazon, Google, or Apple's Siri lack physical faces, some incorporate digital displays or visual cues, such as a glowing light or animated graphics, to indicate their responses. This raises the question of whether voice assistants should have a physical face with which users can interact, allowing for reactions not only through voice but also through digital facial expressions or body language cues.

Wienrich et al. (2022) highlight that users often anthropomorphize voice assistants, attributing human-like qualities to them, especially when they exhibit humanised features or behaviours. Therefore, the perception of voice assistants is closely tied to anthropomorphism. This suggests that integrating facial expressions or other human-like visual elements could enhance users' interactions with voice assistants, providing a more intuitive and engaging experience.

Human Look or Robot Look Appearance

Virtual Assistants (VAs) come in various forms, with some featuring visual representations like avatars, models, or images to accompany their voices during interactions with users. Some voice assistants either resemble a human, a cartoon, or even have a body to enrich the interaction experience. Prakash and Rogers (2015) found that, generally, people prefer a virtual assistant with a clear appearance, either highly human-like or distinctly robotic, over a mixed human-robot appearance. They also found that the human-likeness of a robot had a significant impact on people's likability, trust, and perceived usefulness towards the robot.

Why is human-likeness so crucial in this context? Shamekhi et al. (2018) proposed that an agent's enhanced social presence from its continuous presence, intuitive and pleasant interactions across multiple modes, and higher task capability linked to a more realistic visual character could explain this phenomenon. On the other hand, interestingly, users don't just see these assistants as mere software, hardware, or algorithms. According to Carolus & Wienrich (2022), users often assign tangible characteristics to these assistants, like personalities, genders, and ages. Despite knowing these assistants aren't real humans, users sometimes treat them as alternatives for human interaction. The study emphasises that people generally perceive these assistants as human or humanoid entities, attributing them with human-like traits and appearances. To effectively convey personality and human-like traits, the evolution of virtual assistants could prioritise embodiment design (Bonfert et al., 2021). Enhancing visual attractiveness has been underscored as a means to achieve a more effective embodiment effect (Khan & De Angeli, 2009).

Facial Expressions and Body Language in Virtual Assistant Visualization

To delve deeper into the concept of human-like appearance for virtual assistants, it's crucial to examine facial expressions and body language. Enhancing feedback to be more emotionally engaging, resembling human responses, might lead users to perceive interactions as more natural and emotionally resonant, fostering a stronger bond with voice assistants. One approach to achieve this effect could involve designing AI agents to resemble humans visually and behave in a human-like manner, such as through facial expressions. According to Ekman and Friesen's study (1975), facial expressions are a vital component in successful human-human social interactions. Moreover, they determined that facial expressions play a key role in the development of emotionally responsive robots, which are also known as affective robots. Emotional facial expressions can be described as specific configurations of facial features that represent distinct emotional states. Some basic emotions have been found to be universally recognizable across different cultures and norms (Ekman and Friesen, 1975). Shi et al. (2018) discovered that positive emotions with high arousal facilitate emotional connections between users and voice assistants (VAs). The study noted that participants showed more facial expression changes when VAs expressed emotions such as joy, eagerness, and excitement through their own facial expressions or text box movements. Facial expressions and other non-verbal cues can be effectively leveraged in affective robot design to enhance social interaction and communication between humans and robots, as well as to convey emotional states (Bates, 1994).

When considering human-like visualisation, focusing on facial expressions or emotional cues akin to emojis emerges as a central aspect of imitating human behaviour to achieve a lifelike visual representation. Among the various forms of human communication like verbal, written, and facial expressions, the latter is particularly potent. Facial expressions have remarkable power, swiftly conveying emotions and intentions, often perceived by others on a subconscious level (Revina & Emmanuel, 2021). Research conducted by Shi et al. (2018) indicates that emotions characterised by positive valence and high arousal facilitate the establishment of emotional connections between users and virtual assistants. Participants showed greater variability in facial expressions when visual aids included facial expressions conveying joy, eagerness, and excitement.

However, adopting a human-like avatar or appearance isn't always advantageous. If the design is partly human-like but not fully realistic, it can trigger the uncanny valley phenomenon. Mori et al. (2012) specifically proposed that a person's response to a robot resembling a human would abruptly shift from empathy to revulsion as it approached, but did not attain, a fully lifelike appearance.

2.2 Vocal feedback and prosodic feature of Voice assistant

Vocal feedback from voice assistants is offered in different types on the market, it can be designed on different aspects of elements for instance intonation.

Human voice vs. Computer voice

According to Knote et al. (2019b), voice or virtual assistants can be categorised into several types based on various factors like communication mode, interaction direction, query input, response output, action, assistance domain, accepted commands, and more. Their cluster study revealed that Embodied virtual assistants as the largest class among voice or virtual assistants, offering both speech and visual output. Their research suggests that compared to assistants without voice or visual output, this type excels in facilitating seamless human-like interactions and enhances the interaction between users and the assistant. That is showing both vocal and visual output can be important. While between human voice and robotic voice, it can be agreed that robotic voice can be more efficient as Sarigul et al. (2020) discovered in their study since people has shorter reaction times to robotic voice than to the human voice. While effectiveness is important, UX isn't solely about it. Overall satisfaction also plays a crucial role. Emotional effects can significantly influence satisfaction but might be overlooked if only effectiveness is evaluated.

Sound is one of the five human senses and plays a pivotal role in perceiving cognitive information naturally. Beyond the literal meaning, the tone of voice significantly influences how the receiver perceives the message. Human voices, with their acoustic properties, effectively convey emotions and are readily recognized by other humans (Bachorowski, 1999). Similarly, when virtual assistants utilize human-like voices, they can convey emotional nuances such as humor, empathy, or contextual understanding, which significantly influence user satisfaction (Hsu & Lee, 2023).

Additionally, Kim et al. (2020) discovered that non-verbal vocal features like a soft tone, varied speech speed with occasional slower and inconsistent pacing, variable pitch, and adaptable intonation can enhance intimacy, similarity, and connectedness with a voice agent. These features also contribute to a more enjoyable and user-friendly interaction experience.

Moreover, Davis et al. (2019) found that human voices tend to engage users more effectively than computer-generated voices. One significant issue with computer-generated voices is the absence of prosodic features such as stress, pitch, intonation, and emotion, which participants quickly notice. Qualitative feedback described computer voices as annoying, obnoxious, and distracting. The natural rhythm inherent in human speech is also easily detected by users, making unnatural computer-generated rhythms stand out.

2.3 Emotional Perception from Voice Assistant

In terms of emotional perception, research indicates that a human-like appearance enhances the user experience when interacting with a Virtual Assistant (VA), as it facilitates emotional feedback and provides psychological support. Moreover, employing vivid and highly human-like visualizations of agents enhances interaction enjoyment, as users find it more engaging and intriguing to engage with such avatars, thus positively influencing user experience. Shamekhi et al. (2018) found that agents equipped with social-interactional intelligence and non-verbal cues enhance interaction intuitiveness and enjoyment. Additionally, the presence of a human-like agent can foster a closer user-agent relationship, characterized by trust and respect (Bonfert et al., 2021).

Furthermore, Virtual Assistants (VAs) perform better when they exhibit personable qualities, incorporating cognitive patterns, emotions, behaviors, and psychological mechanisms (Founder, 2012). These studies above suggest that users may perceive emotional interaction feedback or prompts from a voice assistant positively when it exhibits human-like characteristics, whether through visualization or verbal communication. However, it remains unclear whether these positive emotions are universally perceived as essential for voice assistant products in general, which could be a relevant subtopic for further study. In conclusion, from the current literature review, there are two areas worthy of further exploration: 1) understanding how humanized elements, whether through visualization or vocal feedback, emotionally impact user perceptions and satisfaction, and 2) identifying the specific needs for which users may prefer a humanized Voice Assistant with emotional intelligence. These topics highlight important avenues for future research in enhancing user experience with voice assistants.

3. Methodology

3.1 Research Design

This study aimed to examine the design elements, both visual and verbal, that could influence users' overall perception of the User Experience (UX) of a Voice Assistant. Qualitative methodology has been employed to explore the factors or reasons that elicit positive or negative feedback from users during interactions with the voice assistant, in a general, non-specific scenario.

Conducting in-depth interviews where participants will be offered chances to interact with different prototypes of voice assistants, could help unveil the root causes of these preferences. The qualitative study utilised in-depth interviews within prototype testing and concept evaluation were conducted to investigate participants' acceptance and preferences regarding four different prototype combinations. The aim was to explore the potential for humanising and robotizing voice assistants, enabling participants to compare these features and reflect deeply on their attitudes and needs regarding humanization and robotization in voice assistants.

The in-depth interviews also combined with the Wizard of Oz method, specifically for reboosting purposes. The Wizard of Oz methodology, initially developed and applied in 1973 by Don Norman and Allen Munro, was designed as a cost-effective way to test prototypes when they cannot function independently (Rosala & Ramaswamy, 2024). This has been applied in the study to ensure fluent interaction between participants and the prototypes. It was only used when the interviewer needed to reset the prototypes to detect commands from participants when they became unresponsive.

3.2 Instruments

Design of the four prototypes

In this study, initially, four prototypes of voice assistants ranging from robotic to human-like visualisations, along with variations in two-level intonations from robotic to humanoid, were presented to participants.

How to design the visualisation

The visualisation of the voice assistant can be divided into two distinct designs. One design features an avatar, represented by a cartoon human face, which conveys humanised characteristics. The other design is a bubble, devoid of any humanised features.

Additional efforts have been invested in the avatar design using Cinema 4D, creating it from scratch with three distinct emotional expressions including smiling, eye blinking, and head nodding and shaking. This animated visualisation has been implemented to give participants greater opportunities to observe and assess the importance of these emotional expressions. The bubble uses a video that shows it fluctuating and changing sizes during interaction.

Figure 1

Design of Avatar

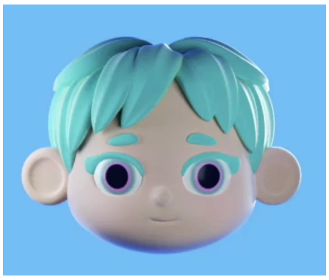
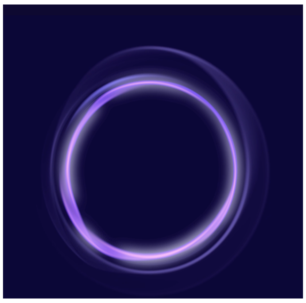


Figure 2

Design of Bubble



How to design the voice feedback

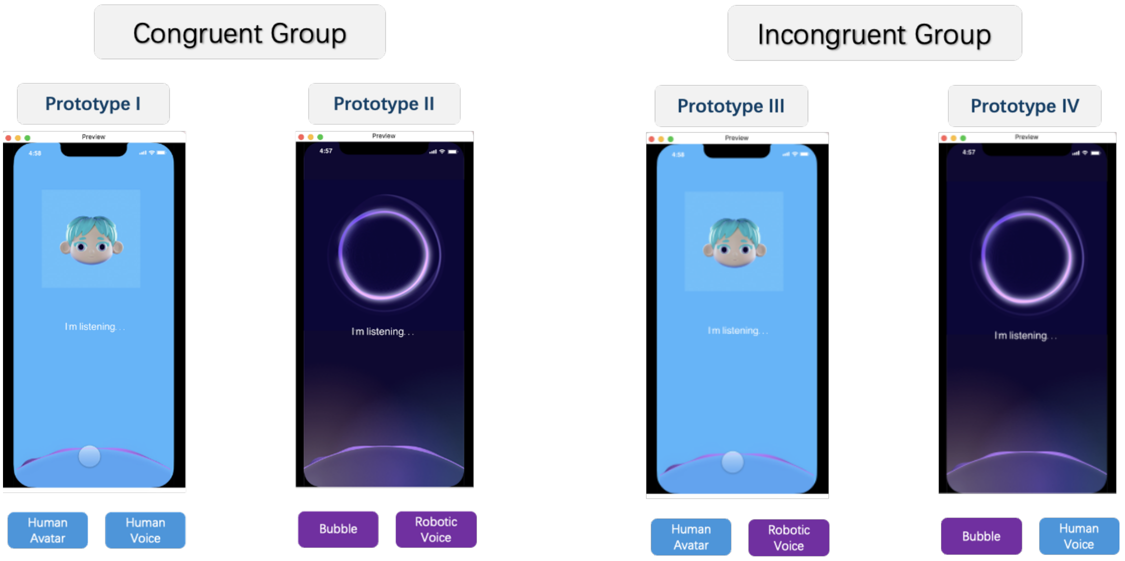
Voice feedback was tested in two different types: one recorded using a human voice and the other generated automatically by the prototype, resulting in a robotic sound. The human voice feedback, recorded by the interviewer, was attributed with humanised characteristics. In contrast, the robotic voice feedback was generated by the software, which read the script using a digital voice.

Four Prototypes - combined two-level of visualisation and two-level of verbal feedback

After generating the visualisations and voice feedback, the four elements were combined to create four prototypes. Four Prototypes including: Prototype I with Human Avatar and Human Voice, Prototype II with Bubble and Robotic Voice, Prototype III with Human Avatar and Robotic Voice, and Prototype IV with Bubble and Human Voice.

Figure 3

Four Prototypes for test



These four prototypes have been divided into two groups: congruent and incongruent. Sixteen participants were evenly distributed between these two groups.

Congruent Group:

Group one: The congruent group included Prototype I, which featured a Human Avatar and Human Voice, and Prototype II, which featured a Bubble and Robotic Voice. These two prototypes had consistent features, either humanised or robotic, to allow participants to experience and compare the differences distinctly.

Incongruent Group:

Group two: The incongruent group included Prototype III, which featured a Human Avatar and Robotic voice, and Prototype IV, which featured a Bubble and a Human Voice. These two prototypes had inconsistent features, with both humanised and robotic, to allow participants to evaluate and experience.

Participants were sequentially invited to experience either Prototypes I and II or III and IV during the interview sessions. This arrangement ensured an equal number of rounds for each prototype to be the first experienced in the interview. These prototypes were generated and interacted with via pre-set commands on the Protopie platform to ensure that participants were able to visualise the interactions, converse with the voice assistants, and receive feedback.

Table 1

Tools and Platform Used

Tool/Platform Name	**Usage**
Cinema 4D	Design, Model and Generate Avatar with Animations
Veed	Generate Human Voice
Protopie	Design and Generate Voice Assistant into 4 Prototypes
Microsoft Teams/Zoom	Interview, record and transcript generation

3.3 Interview Protocol

The interview was structured into five parts: a warm-up section, testing and evaluation of the first prototype, testing and evaluation of the second prototype, comparison of the two prototypes within each group, discussion on humanization and robotization, and a wrap-up session, totalling 60 minutes. The interview guide is attached as Appendix A.

Table 2

Interview Procedure

Sessions	**Detailed discussion**	**Timeline**
Warm up	Discussion about their pre-knowledge, current usage, current attitudes and expectations of voice assistant	8 min
1st prototype concept evaluation	Participants interacted with the first prototype through scenarios and evaluated its elements, followed by discussions on emotional perceptions towards this prototype.	15 min
2nd prototype concept evaluation	Participants interacted with the second prototype through scenarios and evaluated its elements, followed by discussions on emotional perceptions towards this prototype.	15 min
Sessions	Detailed discussion	Timeline
Comparison	Compare the 1st and 2nd prototype, share preference and the suitable scenarios for each prototype	10 min
Humanization and Robotization	Based on the concept evaluation, share preferences, attitudes and expectations towards humanization or robotization of ideal voice assistant	10 min
Wrap up	Summarise important findings with participants	2 min

Warm-Up

The interview begins with a warm-up section designed to introduce the study and obtain informed consent from the participants. This section also gathers background information on participants' prior experience with voice assistants. Research topics in this part focused on the participants' familiarity and frequency of use with various voice assistants, the scenarios in which they typically use these assistants, their general feelings towards them, and their preferences regarding visual interfaces and human-like voices. This background information set the stage for understanding the context of the participants' experiences and preferences, which is essential for analysing how the visual and vocal aspects of voice assistants influence their perceptions.

Prototype Testing (1st and 2nd Prototypes)

Participants then moved on to interact with two different versions of the voice assistant prototype, referred to as "Monday." This section involves two rounds of testing, where participants engaged with each prototype through predefined scenarios such as greeting the assistant, checking the weather, asking for restaurant recommendations, and setting navigation. The purpose was to observe participants' interactions with both prototypes, gather their initial impressions, and capture their emotional responses to the design elements. This combined prototype testing helps answer RQ1 by examining how the visual representation of the voice assistant affects users' emotional and psychological perceptions and RQ2 by assessing how the vocal feedback with humanised features influences users' emotional perception and subsequently impacts their user experience.

Comparison of Two Prototypes

After interacting with both prototypes, participants are asked to compare the two versions. This section focuses on discussing preferences, emotional connections, and the reasons behind favouring one design over the other. Participants reflect on their experiences and provide insights into which elements they found more appealing and why. This comparative analysis is crucial for understanding how the visual representation and vocal feedback of voice assistants influence emotional perception and user experience, directly addressing RQ1 and RQ2.

Humanization and Robotization

In this part, participants evaluate the importance of human-like features in voice assistants, such as facial expressions and vocal tones. The discussion explores when and why participants prefer human-like features and how these features impact their emotional support and user experience. This section delves into the psychological aspects of user interaction with voice assistants, providing insights into how humanised vocal feedback influences emotional perception and user experience, thus addressing RQ2.

Wrap-Up

The interview concludes with a wrap-up section where participants are thanked for their participation. They are given an opportunity to share any additional thoughts or feedback that might not have been covered during the interview. This final part ensures that any overlooked aspects are captured, providing a comprehensive understanding of the participants' perceptions and experiences.

In summary, the interview is structured to systematically explore how the presence of an image or agent in the visual representation of a voice assistant (RQ1) and the humanised vocal feedback (RQ2) influence users' emotional and psychological perceptions and their overall user experience. Each part of the interview builds on the previous one, linking the gathered data to the research questions and providing a comprehensive view of user interactions with voice assistants, particularly regarding their emotional perceptions.

Table 3

Scenarios and Tasks for interaction

4 Steps of Interactions	Tasks
Greeting	Greeting to the Voice Assistant
Scenario 1 - Functional Tasks	Check weather - Ask for restaurant recommendation - Set navigation
Scenario 2 - Emotional Support	Ask for help when feeling upset - Play the music - Tell VA feel lonely - Set alarm
Thank you	Thank you / Bye bye

The selection of usage scenarios and tasks is crucial for testing the prototypes, as it can affect the difficulty of the test. This, in turn, may influence how users perceive the capability and usability of the voice assistant, ultimately impacting their overall satisfaction with the interaction. However, it's important to note that the overall user experience satisfaction of interaction with the prototype is not the main focus of this study. Therefore, the selection of usage scenarios will be categorised into two main fields at a basic level. The first field is functional, encompassing tasks such as weather checking. The second field is emotional and conversational, involving activities like engaging in small talk or conducting quick checks. The Discussion Guide provided in Appendix A will outline the final interview discussion.

Pilot Interview

A pilot interview was conducted to finalise and refine the interview questions and flow. During pilot testing, real-time note-taking was emphasised to enable interviewers to promptly record keywords and ratings, ensuring accurate reflection of participant feedback while they were still engaged with the topic. Sequential presentation of prototypes was also highlighted, with the first prototype aimed at minimising initial learning curves, thus facilitating easier interaction with subsequent prototypes. This approach aimed to mitigate potential biases in perception and satisfaction due to varying learning costs. Overall, the study flow was well-organised, with the participant of the pilot interview reporting no cognitive or logical difficulties in understanding and responding to the questions.

3.4 Validity and Reliability of Protocols

Validity

Validity in this study has been established through confirming credibility, transferability, dependability, and confirmability.

Credibility was ensured through triangulation which involved capturing both verbal and non-verbal cues during data collection, including emotional expressions, attitudes, and subconscious emotions observed. These data were validated through prompt feedback sessions with participants.

The main research questions and concepts were phrased in multiple ways to ensure participants comprehended and assessed them uniformly. Detailed accounts of the research process and results, akin to thick description, were provided to ensure transferability. This included contextual explanations of findings from interviews and presenting participants' exact responses verbatim. These efforts aimed to facilitate comprehensive understanding and accessibility of the study for the readers. Moreover, data saturation was confirmed through iterative data analysis, where emerging themes remained consistent without yielding new insights on the same research topics.

Dependability was demonstrated by maintaining clear and comprehensive records. Transcripts, verbatim statements on each research topic, the process of code generation, and definitions of codes were documented in separate files. This approach substantiated the reliability and transparency of the research process.

Confirmability was established by transparently sharing the process of code frame generation, employing systematic coding techniques, and providing clear definitions to minimise bias in data interpretation. These methods aimed to ensure the objectivity and neutrality of the study's findings.

Overall, these strategies collectively supported the validity of the study, ensuring that the research was credible, transferable, dependable, and confirmable.

Reliability

Consistency in data collection was ensured by employing standardised methods and procedures across all participants and research topics. This included using consistent interview protocols, such as interview guides and note-taking sheets, and employing observation techniques, including recording and double-checking main questions and ratings, to verify understanding across all interviews.

Inter-coder reliability in this study was assessed using Cohen's kappa for the five codebooks, comparing results with another coder based on analysis of four randomly selected interviews.

Table 4

Cohen's Kappa of five codebooks

Measurements	Cohen's Kappa	N of Valid Cases
Avatar	0.96	25
Bubble	0.94	32
Human Voice	0.846	26
Robotic Voice	0.94	23
Humanization and Robotization	0.94	47

Transparency in the data processing procedure was maintained by sharing transcripts, recordings, and verbatim excerpts that illustrate different code schemes. The process of generating code schemes and the codebook, which includes definitions of codes, were also made accessible. Quotes were used to vividly and authentically illustrate explanations of the codes to participants.

Reflection on the researcher’s role, biases, and potential influences was conducted through pilot interviews to refine protocols and procedures. Iterative data analysis was also utilised, and data results were shared with peer interviewers to ensure comprehensive coverage of the phenomenon under study.

3.5 Participants

The study focused on participants with prior experience using voice assistants across various applications, who expressed a willingness to continue using voice interactions and remain receptive to technology products in the future. Sampling encompassed individuals familiar with diverse types of voice assistants. Sixteen participants, aged 22 to 49 years (mean = 28 years, SD = 6.6), were recruited, including 5 males and 11 females, all of whom reported previous experience with voice assistants.

Table 5

Participants List

Participant No.	Age	Gender	Usage of Voice Assistant
P1	23	Female	Alexa and Google Assistant
P2	23	Female	Siri
P3	32	Male	Siri
P4	24	Male	Alexa and Google Assistant
P5	31	Female	Siri, In Car Assistant, Humi
P6	33	Male	Siri, In Car Assistant, Humi
P7	49	Female	Siri, In Car Assistant, Humi
P8	24	Female	Siri
P9	22	Female	Siri
P10	30	Male	Siri
P11	26	Female	Siri
P12	24	Female	Siri
P13	29	Male	Google Assistant, Siri
P14	30	Female	Google Assistant, Siri
P15	30	Female	Siri, In Car Assistant, Humi, Other Smart Speaker
P16	23	Female	Siri, In Car Assistant, Other Smart Speakers

3.6 Data Analysis

Likability of Four Elements

Reysen's (2005) likability scale is a widely used measurement tool designed to assess how individuals perceive the likability or attractiveness of others. It includes a series of statements that respondents rate based on their level of liking or admiration towards a person or group. This scale is structured to capture subjective perceptions of social attractiveness, evaluating various attributes or behaviours that contribute to likability. Typically, respondents use a Likert scale ranging from 1 to 7 to indicate their degree of agreement or endorsement of each statement. Higher ratings indicate greater likability, while lower ratings suggest lesser likability.

In the context of this study, the likability scale has been modified and applied to explore participants' perceptions of liking towards both humanised and robotic features. By adapting this scale, the research aims to delve deeper into how individuals perceive and respond to these features, shedding light on nuanced attitudes and preferences. By leveraging Reysen's likability scale (2005) in this manner, the study seeks to uncover insights that can inform the design, acceptance, and integration of human-like and robotic features in various contexts. This approach not only facilitates a quantitative assessment of likability but also provides a structured framework for understanding subjective evaluations and preferences related to technological and humanistic attributes.

Emotional perceptions

The 7-point Likert Scale was chosen for its effectiveness in capturing nuanced responses from participants regarding their emotional connection with the prototypes. This scale allows for a range of responses, from strongly disagree to strongly agree, providing granularity in participants' emotional perception towards the four prototypes. The 7-point Likert Scale emerged as a common variation, offering a broader range of response options than 5-

point scale, to allow respondents to express their level of agreement or disagreement in more detail, thereby enhancing the sensitivity and precision of data collection. Using this scale, the researcher was able to discern the differences in emotional perception and connection between participants and the four prototypes.

Attributes Analysis of the four elements

An attribute analysis of the four elements involved systematically categorising qualitative data using attribute codebooks. These tools were essential for organising likes, dislikes, and underlying psychological connections into predefined codes that represented specific attributes or themes relevant to each element under study. The use of codebooks ensured consistency and accuracy in identifying patterns of preferences and expectations associated with each element.

Table 6

Code Book for Avatar

Category	Code	Definition	Example	Participants
Humanization	Higher expectation on the humanised animation	The participant has higher expectations for the animation to be more human-like.	"I expect it to behave more like a human."	P3, P5, P6, P8, P9, P11, P14, P15, P16
	More human and natural conversation	The participant perceives the conversation as more human and natural.	"The conversation feels very natural and human."	P1, P2, P4, P7, P9, P16
	Too human	The participant feels the interaction is excessively human-like.	"It seems too human."	P13
Entertainment Visual Effect	Entertainment	The participant finds the visual effect entertaining.	"It's very entertaining to watch."	P10, P12, P14, P16
	Cute	The participant describes the visual effect as cute.	"It's really cute."	P1, P4, P15
	Please, relax to see	The participant finds the visual effect pleasing and relaxing.	"It's pleasing and relaxing to see."	P1, P3

Table 7

Code Book for Bubble

Category	Code	Definition	Example	Participants
Neutral	Neutral	The participant expresses a neutral perception, neither positive nor negative.	“I don't have any strong feelings about it.”	P8, P10, P14, P16
	No feeling	The participant explicitly states having no particular feelings or opinions.	“It doesn't evoke any feelings in me.”	P9, P10, P14
Positive Feeling	Better focus	The participant mentions it’s helpful for focus or concentration.	“I can focus much better when I use it.”	P1, P15
	Natural	The participant describes the experience as feeling natural or intuitive.	“It feels very natural to use.”	P5
	Satisfying	The participant finds the experience satisfying.	“It's quite satisfying to interact with.”	P1
	Comfortable	The participant describes the experience as comfortable.	“I feel very comfortable using it.”	P5
	Calm	The participant feels calm or relaxed.	“It makes me feel calm.”	P3, P13
	Peaceful	The participant experiences a sense of peace.	“Using it feels peaceful.”	P4, P5
Nice Design	Aesthetically pleasing	The participant finds the design visually appealing.	“It's really aesthetically pleasing.”	P1, P2, P4, P6
	Cool	The participant describes the design as cool or trendy.	“It looks really cool.”	P11
	Cute	The participant finds the design cute or charming.	“It's so cute!”	P12
	Dynamic	The participant perceives the design as dynamic or lively.	“The design is very dynamic.”	P5, P6, P7
	Well developed	The participant appreciates the detailed and well-thought-out design.	“It's very well developed.”	P6

Table 8

Code Book for Human Voice

Category	Code	Definition	Example	Participants
Emotional Connection	Emotional connection	The participant feels an emotional connection with the interaction.	"I feel a strong emotional connection."	P1, P4, P5, P7, P8, P9, P10, P13, P14
Positive Engagement and Enjoyment	Positive feeling/happy feeling	The participant experiences positive or happy feelings during the interaction.	"It makes me feel happy."	P2, P3, P9, P10, P12
	Excitement and interesting	The participant finds the interaction exciting and interesting.	"It's really exciting and interesting."	P1, P16, P11, P10
	Adorable/Cute	The participant describes the interaction as adorable or cute.	"It's so cute!"	P12
Secure	Reliable	The participant perceives the interaction as reliable.	"It seems very reliable."	P9
	Warm, kind and friendly	The participant finds the interaction warm, kind, and friendly.	"It's very warm and friendly."	P1, P8, P9, P15
	Relaxing and pleasing	The participant finds the interaction relaxing and pleasing.	"It's relaxing and pleasing to interact with."	P3, P4, P8, P10
Comprehensio n	Easy to understand	The participant finds the interaction easy to understand.	"It's very easy to understand."	P2
	Understood/C omprehended	The participant feels that they understood or comprehended the interaction.	"I understood it perfectly."	P9
Higher Expectation	High expectation	The participant has high expectations for the interaction.	"I have high expectations for this."	P3, P7

Table 9

Code Book for Robotic Voice

Category	Code	Definition	Example	Participants
Neutral	Neutral	The participant expresses a neutral perception, neither positive nor negative.	"I don't have any strong feelings about it."	P7, P11, P14, P16
Robotic	No emotions/No connection	The participant feels the interaction lacks emotions or a personal connection.	"It doesn't evoke any emotions or connections."	P1, P14
	Not appropriate for emotional talk	The participant finds it unsuitable for emotional conversations.	"Not appropriate for emotional talk."	P3
	Monotone	The participant perceives the communication as monotone.	"It's very monotone."	P4, P9, P10
	Cold	The participant describes the interaction as cold.	"It feels cold."	P5, P15
	Robotic	The participant perceives the interaction as robotic.	"It sounds robotic."	P4, P8, P12, P15, P16
Efficient Communication	Precise/Simple/Straightforward	The participant appreciates the precision and simplicity.	"It's very straightforward and precise."	P1, P12
	Clear to hear	The participant finds the communication clear to hear.	"It's clear to hear."	P2, P3
Professional and Rational	Professional	The participant perceives the communication as professional.	"It sounds professional."	P6
	No emotions (Positive)	The participant finds the lack of emotions to be a positive trait.	"It's good that there are no emotions."	P13

Attributes Analysis of User Needs of Humanization and Robotization

By employing the code book, researchers ensure consistency and rigour in identifying patterns and themes concerning how individuals perceive and interact with humanised and robotic elements. Moreover, it facilitates a nuanced exploration of the psychological motivations and reasons behind these perceptions, shedding light on the deeper emotional and cognitive connections people may have with such technologies. Through systematic coding and analysis, the code books enable researchers to uncover hidden insights into the psychological benefits and underlying needs that drive perceptions and preferences towards humanization and robotization.

Table 10

Code Book of Humanization - Positive

Category	Code	Definition	Counts
Emotional Needs	Higher Capability	The participant perceives a high level of understanding and empathy.	2
	Supportive for negative emotions	The participant feels supported in dealing with negative emotions.	5
	Accompany	The participant feels accompanied and not alone.	4
	Take care of me	The participant feels taken care of.	2
	Discussion of private topics	The participant feels comfortable discussing private topics.	2
	Comfort me	The participant feels comforted.	1
	Dealing with stress	The participant feels helped in dealing with stress.	1
	Facilitates easier connection and expression	The participant finds it easier to connect and express themselves.	1
	Human reaction	The participant perceives human-like reactions.	1
	Not too rational	The participant finds the interaction not excessively rational.	1

	Encourage	The participant feels encouraged.	1
Functional Benefits	Comfortable and Natural Conversation	The participant finds the conversation comfortable and natural.	3
	Satisfying conversation	The participant finds the conversation satisfying.	1
	Fun	The participant finds the interaction fun.	4
	Intelligence	The participant perceives the interaction as intelligent.	1
	Less frustration	The participant experiences less frustration.	1
	Happy and delightful	The participant feels happy and delighted.	2

Table 11

Code Book of Humanization - Negative

Codes	**Definition**	**Example**	**Counts**
Trustworthy	The participant finds it very trustworthy.	"I find it very trustworthy."	4
Scary	The participant finds it somewhat scary.	"It's a bit scary."	2
Privacy	The participant is concerned about privacy.	"Privacy is a major concern for me."	1
Too Realistic	The participant feels it is too realistic.	"It seems too realistic for comfort."	1
Safe	The participant feels safe using this.	"I feel safe using this."	1

Table 12

Code Book of Robotization

Code	**Definition**	**Counts**

Reliable	Consistently performs as expected	1
Professional	Displays predicted standards of conduct and performance	2
Functional Well	Operates effectively and efficiently	1

4. Results

This section presents the findings, including ratings for two key factors — visualisation and verbal feedback — for each of the four prototypes. Additionally, it details the emotions perceived for each prototype and identifies the appropriate use case for each.

4.1 Likability of four elements

Table 13

Likability Scores of the Four Elements

Elements	Mean Likeability Score
Human Voice	5.65
Robotic Voice	4.09
Human Avatar	5.10
Bubble	4.56

The mean likeability scores were evaluated across four different combined prototypes.

Element One: Human avatar as visualisation of voice assistant

The first element is the human avatar, which is used as the visualisation for prototypes I and III. The average overall likability rating for the avatar is 5.1 on a scale from 1 to 7. Most first impressions of the avatar describe it as cute and adorable. The alternate perception regards the avatar as a human-like entity, prompting increased engagement in social behaviours.

Quotes

"Because the avatar is so cute and like, just nice to look at." - P1, Female, 23 y.o.

"Pleasure, relax and more natural. It gives me more feeling that I'm talking to a person. It has higher awareness I perceive." - P3, Male, 32 y.o.

Element two: Bubble as visualisation of voice assistant

The second element is the bubble, which is used as the visualisation for prototypes II and IV. The average overall likability rating for the bubble is 4.56 on a scale of 1 to 7. Most first impressions of the bubble described it as calm and peaceful. There were no extreme negative or positive reactions, with most comments being neutral.

Quotes

"I think the design is what mostly brings me comfort because I'm very used to this sort of design for abstract assistant and that makes me feel kind of comfortable and used to it." - P4, Male, 24 y.o.

Element three: Human voice as verbal feedback of voice assistant

The third element pertains to the use of human voice for verbal feedback in Prototypes I and IV. On average, participants rated the likability of this verbal feedback at 5.65. Initial impressions of the human voice were largely positive, described as comfortable, encouraging, and pleasing. Moreover, the human voice evoked expectations regarding the capabilities of the voice assistant. Participants felt a greater sense of naturalness, leading to reduced awkwardness and increased willingness to engage in conversation with the human voice assistant.

Quotes

"It was nice and satisfying, aesthetically pleasing. It just sounded very gentle and comfortable and understanding." - P1, Female, 23 y.o.

"With the robotic voice I feel that and then feel a little bit awkward, but with the human voice, I don't feel it's weird or awkward" - P2, Female, 23 y.o.

Element four: Robotic voice as verbal feedback of voice assistant

The fourth element concerns the use of a robotic voice for verbal feedback in Prototypes II and III. On average, the likability rating for the avatar is 4.09 on a scale from 1 to 7. Initial impressions of the avatar often characterise it as familiar yet cold. The robotic voice is commonly associated with existing voice assistants on the market, lacking the naturalness of human speech. While most participants did not express dislike for this element, many mentioned growing accustomed to it over time.

Quotes

“I feel cold because it sounds, yeah, a bit more like a robot, like Siri, and it doesn't really have those emotional expressions.” - P2, Female, 23 y.o.

Conclusion

Overall, the human voice received the highest likability rating, drawing significant positive feedback in both Prototype I and Prototype IV. Its ability to evoke a pleasant sensation among participants and facilitate more natural conversations contributes to this acclaim. Additionally, participants rated the human avatar above average, primarily valuing its entertainment and amusement factor. The bubble element received a slightly above-average rating following the second element. Lastly, the robotic voice was evaluated by participants as calm but somewhat cold, lacking novelty. While likability reflects the general attitudes of participants toward each element, the attributes of each element provide a more detailed understanding of what users specifically like or dislike about each element.

4.2 Attribute Perception Analysis

After analysing the likability of the four elements, it is crucial to understand how participants perceived the characteristics and attributes of each element.

Attributes of Visual Elements - Human Avatar and Bubble

Attributes on avatar. During the interview, participants examined the four elements individually and described them using various attributes. For the visualization element, specifically the avatar, predominant attributes included Entertainment Visual Effect and Humanization, which were identified as the main characteristics of the Human Avatar.

Table 14

Attributes of Avatar

Attributes of Avatar		No. of Mentions	Participants No.
Humanization	Higher expectation on the humanised animation	9	P3/P5/P6/P8/P9/P11/P14/P15/P16
	More human and natural conversation	6	P1/P2/P4/P7/P9/P16
	Too human	1	P13
Entertainment Visual Effect	Entertainment	4	P10/P12/P14/P16
	Cute	3	P1/P4/P15
	Please, relax to see	2	P1/P3

Humanization emerges as the predominant observation regarding the avatar. This cartoon-like assistant, characterised by minimal facial expressions such as head nodding and eye blinking, conveys signals of human-like behaviour. Participants acknowledge this aspect, leading to heightened expectations, particularly in initial interactions with the prototypes. Like they would link the avatar to a real human person. The perception of the avatar as cute and human-like fosters a sense of face-to-face interaction, which enhances a more human and natural conversation. Moreover, the advantage of humanization lies in mitigating the uncanny valley effect, as the avatar resembles a human but retains a cartoonish quality.

Regarding humanised animations, participants anticipated the addition of more vivid and sophisticated features to enhance entertainment value further. Over half of the participants felt that the current animations were insufficient and expected more realistic human-like animations. They believed that more advanced features would make the interaction more enjoyable and create a more authentic conversational experience. Some also mentioned that these improvements would contribute to a more consistent and coherent experience, especially when paired with a human voice.

While humanization is not always an advantage, some participants felt that a too-human appearance could trigger negative feelings, such as social pressure. Two participants mentioned that if the voice assistant appeared too human, they would feel bossy asking it to perform tasks they could manage themselves. Additionally, a highly realistic avatar could trigger the uncanny valley effect, although the current slightly cartoonish design avoids this issue and is not perceived as unsettling.

Quotes

“So if it's more personal, I would like that the character has more of a facial expression so that when I am talking to it and I'm looking at it, it feels like I have someone in front of me instead of something.” - P9, Female, 22 y.o.

“If it looks more real then I feel more interested in that. I think it's more like consistency like I see a human face and I hear a human voice so these are consistent.” - P14, Female, 30 y.o.

"Maybe because it's not fully emotionally advanced. Like, but if it can smile and blink and all of it, I would think it's a higher rate (Now she gave 6 out of 7), It has blue hair and very turquoise blue eyes, that it's less distinct from humans.(Less scary)" - P8, Female, 24 y.o.

Entertainment Visual Effect. One of the notable attributes was entertainment visual effect, as acknowledged by a select few participants who value the avatar. During interactions, participants primarily focused on tasks, relying more on verbal feedback when the visualisation serves as a mere companion to the conversation. However, the avatar's presence serves a dual purpose beyond functional utility. It offers entertainment and joy, which some participants appreciate. Some participants also mentioned the avatar is cute or pleased and relaxed to look at it while talking with it. This aspect of supplying entertainment and enjoyment stands out as one of the key reasons why participants value the avatar, despite its limited practical functionality for communication.

Quotes

"The avatar is so cute and like, just nice to look at. It makes you feel more comfortable to talk to the assistant. It doesn't make you feel judged or, you know." - P1, Female, 23 y.o.

"I would need an avatar in general. It must fit into many kinds of scenarios because now the expressions are limited. And in some scenarios, people could easily get bored of it and feel unreal. Entertainment is the main reason I might need that, to not get bored." - P10, Male, 30 y.o.

Attributes of the Bubble. Regarding the bubble, its main attributes were identified as Neutral, Positive Feeling, and Nice Design.

Table 15

Attributes of Bubble

Attributes of Bubble		**No. of Mentions**	**Participants No.**
Neutral	Neutral	4	P8/P10/P14/P16
	No feeling	3	P9/P10/P14
Positive Feeling	Better focus	2	P1/P15
	Natural	1	P5
	Satisfying	1	P1
	Comfortable	1	P5
	Calm	2	P3/P13
	Peaceful	2	P4/P5
Nice Design	Aesthetic pleasing	4	P1/P2/P4/P6
	Cool	1	P11
	Cute	1	P12
	Dynamic	3	P5/P6/P7
	Well developed	1	P6

Neutral. A few participants perceived the bubble neutrally, experiencing neither positive nor negative feelings during interactions with it. In scenarios, this neutrality can be beneficial, since it minimizes emotions and feelings.

Quotes

"In this scenario I was upset and at least the bubble is not making something worse. It's good." - P3, Male, 32 y.o.

"Doesn't give me any emotion, which is why I kind of like it. It's something that is abstract, not human." - P13, Male, 29 y.o.

Nice Design. In general, the bubble was perceived as aesthetically pleasing, cool, and dynamic.

Positive Feeling. The bubble had a positive effect on participants during conversations, as they mentioned that its dynamic effects, dilation and contraction, helped them focus more and feel more natural. Participants also found it satisfying, calm, comfortable, and peaceful to look at. Participants noted that it provides users with imaginative space, allowing them to concentrate more on the conversation itself. Unlike avatars which can offer entertainment, the bubble is not inherently negative, it tends to be perceived as neutral, offering a versatile and more flexible canvas for interaction.

Quotes

"It's a large one and its dynamic effect is very natural. And makes people feel comfortable and peaceful." - P5, Female, 31 y.o.

"It's something like people will watch and feel satisfied so that they can focus better. It's easier to focus on the conversation and just have a better experience." - P1, Female, 23 y.o.

Attributes of Vocal Elements - Human Voice and Robotic Voice

Attributes of human voice. The human voice was perceived through several aspects, with five main attributes identified: Emotional Connection, Positive Engagement and Enjoyment, Security, Comprehension, and Higher Expectation.

Table 16

Attributes of Human Voice

Attributes of Human Voice		No. of Mentions	Participants No.
Emotional Connection	Emotional Connection	9	P1/P4/P5/P7/P8/P9/P10 /P13/P14
Positive Engagement and Enjoyment	Positive Feeling/Happy Feeling	5	P2/P3/P9/P10/P12
	Excitement and Interesting	4	P1/P16/P11/P10
	Adorable/Cute	1	P12
Security	Reliable	1	P9
	Warm, Kind and Friendly	4	P1/P8/P9/P15
	Relaxing and Pleasing	4	P3/P4/P8/P10
Comprehension	Easy to Understand	1	P2
	Understood/Comprehended	1	P9
Higher Expectation	High Expectation	2	P3/P7

Emotional Connection, the most frequently mentioned attribute, was perceived by 9 participants in various ways. Generally, they stated that the human voice stood out immediately, creating a specific emotional connection. Participants found it surprisingly natural, akin to talking to a real person on a phone call. The human voice touched them with its natural vocal characteristics, such as human-like intonation, speech flow, and pitch. This evoked their emotions immediately, making them feel closer and more connected, even imagining a real person behind the voice assistant. One participant mentioned that it felt overly intimate, making her feel uncomfortably close to the voice assistant.

Quotes

"It sounds very exciting to help me. And that makes me kind of positive. Like it kind of, the avatar or the, the assistant kind of sounds like it's smiling. It makes me want to smile. Just kind of like a sense of satisfaction because of how smooth the replies were and how smooth the conversation was." - P1, Female, 23 y.o.

"It gives me a connection, like with a human. I think that the voice definitely is what triggers the most emotional connection, more like the social connection." - P4, Male, 24 y.o.

Positive Engagement and Enjoyment. Beyond connection, the human voice also provided excitement and interest to users. Hearing the human voice made them perceive the voice assistant as a real human, which made communication feel more intriguing and stimulating. Additionally, users found the human voice adorable and cute, enhancing their happiness during interactions and positively engaging them to communicate more with the voice assistant. Several participants mentioned that the human voice initially captures their attention and helps them focus more on the conversation.

Quotes

"Actually I will be more willing to talk about it. It will trigger my interest to talk to that robot." - P11, Female, 26 y.o.

"It just sounds more interesting than the robotic voice." - P16, Female, 24 y.o.

Security. The human voice has been valued for its warmth, kindness, friendliness, and relaxing qualities by users. These positive sensations contribute to their feeling of safety during interactions, reinforcing the sense of closeness mentioned earlier.

Quotes

"The assistant kind of sounds like it's smiling. It makes me want to smile." - P1, Female, 23 y.o.

"It's just more comfortable and relaxing to talk with, due to the human-like voice" - P4, Male, 24 y.o.

"It was pleasant actually. Like it was kind, and it didn't feel unnatural." - P8, Female, 24 y.o.

Comprehension. The human voice demonstrates greater understanding, even when delivering responses that are content-wise identical to those of the robotic voice. Users find it easier to understand and perceive the human voice as conveying more empathy and comprehension. This human-like quality enhances comfort during interactions and emphasises the humanity of the voice assistant, despite its robotic nature. The improved comprehension abilities also help participants feel more at ease and facilitate smoother communication during conversations, thereby increasing their willingness to use the voice assistant.

Quotes

"I feel like I was way more understood, in a way that I'm talking to someone who is real, then just a voice on the phone. I think more of a reality. So I feel like having a better conversation." - P9, Female, 22 y.o.

"It's more human-like, kind of like my real friend instead of a robot. It will make me not hesitate to seek help from her." - P10, Male, 30 y.o.

Higher Expectation. The human voice triggered higher expectations among some participants, as they perceived it as more human-like and therefore expected the voice assistant to understand their emotions and provide emotional feedback. This assumption led them to imagine that the voice assistant could comprehend their feelings and possibly detect their emotions.

Quotes

"More pleasing. Tone and speed are more different. More positive. It will give me less patience and higher expectations." - P3, Male, 32 y.o.

"I would expect it to be more emotional in feedback." - P7, Female, 49 y.o.

"It feels like the voice really wanted to help." - P9, Female, 22 y.o.

Attributes on robotic voice. Overall, the robotic voice has been evaluated as Neutral, Robotic in general, but also Efficient in Communication by some individuals.

Table 17

Attributes of Robotic Voice

Attributes of Robotic Voice		No. of Mentions	Participants No.
Neutral	Neutral	4	P7/P11/P14/P16
Robotic	No emotions/No connection	2	P1/P14
	Not appropriate for emotional talk	1	P3
	Monotone	3	P4/P9/P10
	Cold	2	P5/P15
	Robotic	5	P4/P8/P12/P15/P16
Efficient Communication	Precise/Simple/Straight forward	2	P1/P12
	Clear to hear	2	P2/P3
	Professional	1	P6
	No emotions (Positive)	1	P13

Neutral, Similarly to the robotic bubble, participants found the robotic voice quite neutral, lacking noticeable emotional expression during communication which is not necessarily to be negative. The neutral sound reminded a few participants of Siri, Alexa or Google assistant which they already got used to.

Quotes

"It just feels like Alexa and google assistant" - P4, Male, 24 y.o.

"It's just neutral. I don't like it. I don't dislike it. Yeah." - P7, Female, 49 y.o.

Robotic, the robotic voice is often perceived as one of its major disadvantages, characterised not only by its robotic nature but also by being monotone, cold, and lacking in emotions. This was particularly evident in scenario 2, where participants rated their interaction with the robotic voice assistant poorly. They expected emotional engagement and feedback from the voice assistant but received responses in a robotic voice. Participants found it too impersonal and uncomfortable to discuss personal emotions or private topics, as the voice assistant lacked empathy.

Quotes

"I don't feel connected to it. I don't feel comfortable, like, having a nice conversation with it, and it doesn't feel that trustworthy." - P1, Female, 23 y.o.

"It has no emotions at all." - P10, Male, 30 y.o.

"This one is just like putting the words next to each other in sentences, but also like you hear that it's a tool talking to you. " - P8, Female, 24 y.o.

Efficient communication. The robotic voice has also been perceived as beneficial, especially for simple and straightforward tasks. Participants found it very easy to communicate with the robotic voice when they needed it to perform straightforward tasks. They appreciated that they could issue simple commands without engaging in human-like conversation or emotional considerations. This perception of the robotic voice as professional and rational made it easier for them to assign tasks to the voice assistant.

Quotes

"I like it, cause it's not that human. So I actually prefer robotics if you ask me to compare especially for tasks like this because I'm not talking to them as I'm talking to another human. When I talk to other humans I'll be not making requests all the time. I feel I'm a bit bossy if I ask a human to do things for me." - P13, Male, 29 y.o.

“It sounds very professional.” - P6, Male, 33 y.o.

Conclusion

The human avatar and human voice, both containing humanised features, were perceived clearly through the avatar's facial expressions and the human-like intonation of the voice. The human avatar triggered higher expectations for advanced animations, mainly for entertainment purposes. The human voice is better connected with users through its emotional tone and feedback, which they found reliable, warm, and relaxing. This enhanced the subconscious belief that a voice assistant with humanised features has a higher capability for emotional comprehension, even though it does not.

On the other hand, both the bubble and robotic voice were perceived as either neutral or beneficial for non-emotional tasks. The non-emotional voice was seen as professional, and the pleasing movement of the bubble helped participants concentrate better, allowing them to manage tasks more efficiently and straightforwardly. These findings lead to the next topic: the emotional perceptions participants had of the four prototypes.

4.3 Comparison of Emotional Perception of the Four Prototypes

Table 18

Mean value of emotional perceptions among four prototypes

Prototype No.	Elements Involved	Mean of Emotional Perceptions Rate
Prototype I	Human Avatar & Human Voice	4
Prototype II	Bubble & Robotic Voice	2.62
Prototype III	Human Avatar & Robotic Voice	2.62
Prototype IV	Bubble & Human Voice	4.75

The prototype with a human voice received the highest ratings for emotional perception among the four prototypes. Notably, Prototype IV, which features the human voice paired with the bubble, was rated higher than Prototype I, which combines the human voice with the avatar. Participants explained that the absence of changing animations in the bubble allowed them to focus more on the vocal responses. The bubble enhanced the connection provided by the human voice, making interactions feel more like a phone call with friends or family. This focus on the voice alone fostered a stronger sense of connection, even without seeing an actual face.

The human voice in Prototype I and IV triggers stronger emotional perceptions, though these are not always positive sensations. Stronger emotional perceptions can be beneficial, particularly in scenarios like Scenario Two, where users seek support when feeling lonely or unwell. In these cases, participants found the human-like voice more helpful, perceiving its supportive and kind feedback as caring. They valued the emotional connection provided by the human voice, feeling that it genuinely tried to address their emotional issues as set in the scenario. However, participants also noted some negative aspects. Treating the voice assistant as a human made them feel bossy when giving commands and introduced social burdens like greetings and saying thank you, making the interaction unnecessarily tedious. This decreased their willingness to use the voice assistant, as it introduced unwanted social activities. Conversely, the robotic voice in Prototype II and III was seen as making interactions easier and simpler, as participants did not feel bad about issuing commands to a robot. They appreciated that the robotic voice did not evoke emotions and felt it conveyed a sense of professionalism and intelligence.

In general, the emotional feedback of a voice assistant can be both helpful and burdensome, depending primarily on the use case and whether it requires emotional engagement. Participants found prototypes II and III, which used a robotic voice, to be too cold, discouraging them from discussing their feelings.

Congruent and incongruent groups did not show much difference in this study, as participants primarily focused on the voice rather than the visual elements. This led them to concentrate on the experience of communicating with robotic and humanised voices. The avatar and bubble were considered less important, resulting in evaluations skewed towards the voice, with participants not perceiving the prototype as a cohesive whole. Only two participants mentioned preferring to see a human face during long conversations, as it feels more like a face-to-face interaction.

Quotes

"It does make the whole interaction feel a lot more natural because the greeting at the beginning and the thank you at the end makes a lot more sense." - P13, Male, 29 y.o.

4.4 Humanization vs. Robotization: Preferences and Needs

Table 19

Humanization preference likert scale

How Humanised do you want your VA to be? Scale 1 - 7	Counts
1 - As Like Robot as Possible	N = 1
2 - Very Robotic	N = 1
3 - A bit Robotic	N = 0
4 - In between	N = 2
5 - A bit humanised	N = 3
6 - Very humanised	N = 3
7 - As Like Human As Possible	N = 6
Mean =	5.38
SD =	1.86

Between the preference for humanization or robotization, this study revealed distinct trends: over half of the participants (N=9) favoured voice assistants being very humanised or as human-like as possible, while a smaller group (N=2) preferred them to remain robotic. Some participants (N=5) expressed a desire for a middle ground in voice assistants, seeking a balance between humanization and robotization to avoid feeling unsettled or unsure. This disparity in preferences stems from participants' understanding, interview experiences, comparisons between robotic and humanised assistants, and their specific usage preferences in real-life scenarios.

Individuals, who preferred humanization, appreciated the benefits of human-like qualities in voice assistants, such as emotional support and natural conversation, yet some of them also value maintaining a clear distinction between human and machine.

Participants inclined towards humanised voice assistants cited emotional support as a primary reason for their preference. They discussed various usage scenarios where a more human-like interaction was deemed beneficial.

Table 20

Expectations and Needs of Humanization

Positive Sides of Humanization			
Emotional Supports of Humanization	**Counts**	**Functional Benefits of Humanization**	**Counts**
Understanding & empathy	5	Higher Capability	2
Supportive for negative emotions	5	Comfortable and Natural Conversation	3
Accompany	4	Satisfying conversation.	1
Take care of me	2	Fun	4
Discussion of private topics	2	Intelligence	1
Comfort me	1	Less frustration	1
Dealing with stress	1	Happy and delightful	2
Facilitates easier connection and expression	1		
Human reaction	1		
Not too rational	1		
Encourage	1		

Emotional Supports

Emotional support stands out as the primary benefit uniquely provided by humanised voice assistants. This support encompasses various aspects, as depicted in table above. Participants who experienced humanised voices expressed expectations and desires for voice assistants capable of empathising with their emotions, particularly negative ones. They value emotional conversations that go beyond task-setting, allowing them to discuss personal emotional issues and receive supportive interactions when feeling down. Humanised assistants are perceived as having the potential to uplift moods and provide comfort akin to interacting with a real human, displaying emotional intelligence and genuine reactions. Participants found it easier to discuss private issues when they knew they were interacting with a robot that could also communicate in a manner that felt like talking to a human. Some participants also expressed a desire to imbue voice assistants with personality traits, fostering a deeper connection similar to relationships with friends or pets. This perceived humanity makes it easier for users to open up emotionally to the assistant.

Quotes

"I think it would be nice if they can comfort me when I'm upset or sad. When I need a human, but not a real human. It can never be a real human." - P2, Female, 23 y.o,

"Yes, humans have emotions, so we need to have some emotional communication" – P5, Female, 30 y.o.

"I wish it can feel like talk with a real person. It can understand more feeling." – P6, Male, 33 y.o.

Functional Benefits. Humanization also demonstrates another functional benefit highlighted in this study: it offers a comfortable and natural conversational style between the voice assistant and users. Even when delivering identical sentences, participants found human voices easier to understand and more comfortable to perceive compared to digital and robotic voices. This ease of understanding, coupled with human-like intonation and pitch, brought them a sense of enjoyment and interest while reducing frustration. Participants appreciated the humanization as facilitating easier and more comfortable communication.

Quotes

"The conversation with this voice is just satisfying, it's so natural." – P4, Male, 24 y.o.

“It feels it’s more capable than a robot,” – P3, Male, 32 y.o.

“I don’t know why. It just sounds more interesting to talk with, like a human, a friend of mine. Is it your voice?” – P11, Female, 26 y.o.

Table 21

Concerns of Humanization

Negative Sides of Humanization			
Needs of privacy	**Counts**	**Uncanny Valley**	**Counts**
Trustworthy	4	Scary	2
Privacy	1	Too Realistic	1
Safe	1		

Humanization also introduces risks to users, particularly when the voice assistant exhibits high levels of intelligence. Participants expressed concerns about trusting a humanised assistant that learns quickly and can think like a human, even though they understand it is ultimately programmed and not truly human. Additionally, there is the risk of encountering the uncanny valley effect when the humanised robot closely resembles a human. This phenomenon blurs the boundary between interacting with a real human and a robot pretending to be human, leading to discomfort and reluctance to share personal thoughts with the assistant. One participant found it difficult to accept the idea that a robot could be as capable as a human, describing it as a challenging concept to embrace.

Quotes

“If it sounds too like a human, I will think twice before I share my thoughts with it. Like I will not share everything with a person as well.” – P6, Male, 33 y.o.

"It sounds scary if it looks like a bubble but sounds like a human. It can never be like a real human but it sounds like it can." – P14, Female, 30 y.o.

Robotization of Voice assistant

Table 22

Needs of Robotization

Positive of Robotization	**Counts**
Efficient conversation	2
Less social burden	2

Table 23

Concerns of Robotization

Negative of Robotization	**Counts**
Inappropriate for emotional talk	2

Robotization is more prevalent as most voice assistants currently on the market are not fully able to converse like real humans. Participants with less emotional needs find it efficient to communicate with robotic voice assistants since they do not expect deep or emotional responses, given their limitations. They preferred not to discuss emotions or feelings with a robot. Moreover, participants perceived less social burden when interacting with robotic voices because they clearly understood that these voices are not human. The straightforward communication style of robotic voice assistants is preferred over more humanised alternatives, especially by those who cannot accept the concept of a robot being human-like. These individuals find comfort in the clear distinction between human and machine, preferring interactions that are pragmatic and devoid of emotional complexity. This preference reflects a practical approach to using technology, where efficiency and task-oriented communication are prioritised over the potential complexities of emotional engagement with a human-like voice assistant.

Quotes

"I don't want to tell it I don't feel good. It just sounds like it doesn't care at all." - P4, Male, 24 y.o.

"Robot voice just gives me it's very professional. Like a computer knowing everything." – P12, Female, 24 y.o.

"I don't feel guilty to tell it to do tasks. Since it has no emotions." – P13, Male, 29 y.o.

Conclusion

A balanced approach should aim to enhance user comfort and trust by maintaining clear and predictable interactions, while also incorporating beneficial human-like qualities that improve user experience and engagement. This approach ensures that voice assistants can provide emotional support and natural conversation without causing confusion or discomfort about their role as artificial entities. This balance allows users to enjoy the advantages of human-like interaction while maintaining a clear understanding that they are interacting with a machine, thereby optimising the usability and effectiveness of voice assistant technology in various contexts.

5. Discussion and Conclusion

5.1 Main Findings

This study employed a combination of prototype simulations and concept test with Wizard of Oz to examine the importance of humanization and robotization involved in visualization and vocal design of voice assistant. It delved into participants' physical and functional needs, fostering a deeper understanding through prototype stimuli.

Answering to the research questions, *RQ1: How does the presence of an image or agent in the visual representation of a Voice Assistant (VA) affect users' perception emotionally and psychologically?*

The mixed cross-testing of two primary elements revealed that most participants focused primarily on voice feedback, paying attention to vocal characteristics such as intonation, tone, flow, pace, and pitch, more than on visualization. This observation aligns with the concept of a "voice assistant," emphasizing the importance of vocal interactions.

However, visualization can be seen as a nice-to-have feature when it adds additional value, especially for providing fun and entertainment. Users also appreciate clear and coherent movement consistency between the animation of an avatar and the vocal feedback. This coherence helps users connect the face with the voice, enhancing the perception of the assistant as human rather than robotic. Perceiving the assistant as human can bring psychological benefits, as indicated in the results, and users generally perceive it as more advanced and well-developed, implicitly suggesting it is imbued with human-like intelligence.

Answering to the RQ2: *How does vocal feedback of voice assistants with humanised features influence users' emotional perception and subsequently impact their User Experience (UX)?*

Voice feedback with humanized features, in this study has shown a greater impact on human emotional perceptions compared to humanized visualization. Users expressed stronger emotional reactions and provided more feedback based on the humanized voice elements rather than the visual aspects. Generally, the human voice is an endearing element that obviously fosters natural conversation and emotional connection between humans and voice assistants (VAs) which have been appreciated by most of the participants in this study. The human voice generates the natural conversation feeling, with the advancement of generative AI, people are viewing voice assistants in more diverse ways of using it than before. However, perceptions still vary among users based on their prior knowledge and their anticipated role of VAs in their daily life.

For instance, some participants have prior experience with traditional voice assistant products like Siri or Amazon Alexa, still would expect VA talking as robotic. On the other hand, another group of users interacts with more advanced voice assistants combined with ChatGPT or other generative AIs. These varying levels of experience shape participants' expectations: traditional users typically use VAs for basic tasks such as setting alarms or checking the weather, whereas advanced users expect more complex interactions, including casual conversations or emotional support through chat.

Due to these different needs, the requirements and expectations for vocal emotional feedback from VAs also vary. Users relying on VAs for simple tasks may prefer efficient and time-saving interactions without conversational functions, appreciating robotic sounds for their practicality. Users who desire a more human-like interaction with voice assistants not only expect the assistant to sound human but also anticipate that it can intelligently understand their emotions and provide suggestions on various topics at the same time the voice sounds positive and emotionally supportive. As highlighted in the results, when voice assistants sound robotic, it can diminish users' willingness to share personal topics with a robot. Participants who have a high need for emotional conversations are generally more satisfied with the human voice compared to the other three elements. They perceived emotions sensitively through intonation, pitch, and tone, aiming to make conversations with the voice assistant as realistic as those with a human.

Overall, the voice characteristics of a voice assistant play a crucial role in shaping its satisfaction and user experience across various scenarios and use cases. Using a completely human-like voice isn't always necessary or advisable. Instead, the key lies in employing voice elements in a manner that aligns with diverse user expectations and usage contexts.

5.2 Discussion

This study explores the theoretical implications that both the voice and visualization of a voice assistant can impact users' emotional perceptions during interactions. It emphasizes that beyond the capabilities of the voice assistant itself, design considerations such as preferences for humanized features implicitly influence user experience. The research questions were formulated based on the linkage between humanization and emotional perceptions. Whether through facial expressions or voice characteristics, humanization design has been shown to affect how users perceive emotional interactions with voice assistants.

As observed in the study by Carolus and Wienrich (2022), humanized facial or bodily features establish a cognitive link for users to imagine the voice assistant as resembling a real human, despite knowing it is a robot. Their research also reveals that user attribute personality traits, characteristics, and even age to these humanized features. Similarly, Castillo et al. (2018) demonstrate that users' perceptions can be significantly influenced by the appearance of embodied conversational agents. These findings are consistent with the results of this study, underscoring that a humanized face or humanoid appearance can evoke emotional responses akin to those experienced in human interactions.

Similarly, researchers learned from the study by Hsu and Lee (2023) that when voice assistants (VAs) display characteristics resembling human language, such as tone and phrasing, as well as positive behaviors like politeness and helpfulness, users experience greater enjoyment in using VAs, develop higher levels of trust in them, and are more likely to continue using them. Building on this understanding, researcher of this study explored more emotions, sensations, and psychological needs that are better fulfilled by a humanlike voice.

Moreover, the critical question remains: how do voice assistants with humanized features impact overall user experience? Insights gleaned from the study by Zhou et al. (2019) similar as study mentioned before, also indicate that humanized features, such as personality traits, can enhance user enjoyment during interactions with robots—a finding also echoed in the results of this study. Beyond enjoyment, the psychological benefits of humanization identified in this study underscore the hidden advantages of perceiving voice assistants as more human-like. For example, warmth and emotional support have been recognized as crucial attributes of an intelligent conversational voice assistant, as highlighted in Gelbrich et al.'s work (2021).

Broadly speaking, the theoretical implications of this study align with its comprehensive findings and are consistent with existing theoretical frameworks. The methodology, including prototype testing and concept evaluation, provided researchers with precious opportunities to explore comprehensive and abstract concepts beyond the capabilities of the prototypes alone. The open-ended questions in the qualitative study investigated key research questions from diverse perspectives, such as preferences for humanized versus robotic elements, and expectations regarding the humanization or robotization of future voice assistant products, drawing comparisons with current prototype versions. However, the qualitative nature of the study introduces certain limitations, which has been discussed in the corresponding section on limitations.

5.3 Practical Implementation

The foundational idea behind this study originated from the perspective of a UX researcher. Its primary novelty lies in providing designers, engineers, and other stakeholders involved in voice assistant products with an exploratory guide. This guide aims to foster insights into whether users perceive emotional values from voice assistants and how to strike a balance between humanization and robotization. Humanization, being more nuanced than robotization, can evoke both positive and negative reactions.

Drawing insights from this study emphasizes the importance of considering the emotional responses users may expect during interactions with voice assistants. The degree of humanization and emotional expression should align closely with the specific goals of the voice assistant product. For instance, conversational usage scenarios highlight the potential benefits of integrating emotional support features, as discussed in this paper. Furthermore, this study serves as a resource for product designers, offering insights into how to tactfully incorporate emotional aspects without unnecessarily stirring emotions. However, the study's prototypes were limited in their ability to animate fully, despite indications that users who appreciated avatars desired more dynamic animations akin to those seen in Disney movies.

Additionally, while cartoon-like avatars were generally well-received in this limited sample, striking a balance between being endearing and avoiding human-like qualities that could unsettle users was noted. Regarding voice characteristics, findings highlighted a preference for human voices due to their perceived emotional resonance, contrasting with robotic voices that were valued for their professionalism and efficiency. Furthermore, human voice naturally facilitates conversations that are considered acoustically pleasing and subconsciously comforting. Unlike robotic voices, human voices have a profound ability to convey emotions, which is particularly valuable in contexts requiring emotional engagement.

Overall, participants in this study expressed openness to more advancements in both functional capabilities and design elements of voice assistants. This indicates a willingness to embrace future developments that enhance user experience across various dimensions. Further designs of voice assistants should consider the impact of humanized features, especially the human voice. Designers should assess whether emotional capabilities are necessary based on the product's intended use. For example, designer should decide if the voice assistant is designed solely for task execution, or it more inclines to conversational usage. The emotional needs identified in the study results might offer conceptual guidance for designing voice assistants that are visually and vocally emotionally supportive.

5.4 Limitation & Future Work

The limitations of this study are twofold. Firstly, it relies solely on qualitative verbatim and coded data, which may not fully demonstrate the interrelationships between variables. While qualitative research aims to explore possibilities and understand the needs behind designing features with more human-like qualities, it's crucial to establish if there is a significant correlation between humanized features and emotional satisfaction. In terms of measurement methods, qualitative research cannot quantify the specific differences in emotional perception between human-like voices and robotic voices. Conclusions can only be drawn based on participants' feedback and ratings of relative preference. However, what can be determined is that, as shown in the conclusions, participants' preference for emotionally perceptible scenarios and the underlying psychology can be understood.

Secondly, from a technological standpoint, the study's avatar animations, while present, may not sufficiently capture how closely an avatar must resemble a human to significantly enhance positive emotional perceptions. To address this, future research interested in the design of humanized features for voice assistants should consider employing more advanced animations to delve deeper into the potential impacts of humanization on the quality of voice assistant interactions.

For further studies in this area, it is recommended to explore the effects of highly realistic avatar animations on enhancing emotional engagement and satisfaction within voice assistant interactions. With the rapid advancement of voice assistants and generative AI, the future of these technologies extends beyond simple task implementation. Moreover, a quantitative study could be conducted to explore, for example, the correlation between user satisfaction with voice assistants and the degree of human-like voice used, or the correlation between user satisfaction with voice assistant and the degree of emotional perceptions. Therefore, there is a growing need to study and understand the broader contexts in which voice assistants are used conversationally, to better meet user needs. Privacy issues, data protection, and concerns surrounding the usage of generative AI should also be carefully considered and evaluated throughout the study.

5.5 Conclusion

In conclusion, in this qualitative study, as a graduate in Communication Science, the researcher integrated topics from both intelligent technology and social sciences to explore design directions applicable to voice assistants. Especially in today's rapidly advancing generative AI landscape, voice assistants present a multitude of possibilities. Through prototype testing, concept evaluation, and exploratory interviews, this study investigated user acceptance and preference for anthropomorphic voice assistants. It also delved into the psychological and emotional reasons why users perceive a need for anthropomorphized voice assistants. Ultimately, insights gained from this research inform future design directions, emphasizing the demand for emotionally supportive voice assistants and the underlying motivations for their necessity. Emotion is one of the most important influences in human life. The research results indicated that as voice assistants evolve, they can simultaneously provide emotional value through both voice and visual feedback, thereby expanding their usability in various scenarios and genuinely assisting users. In this study, anthropomorphic design has been shown to better connect with users, offering more diverse possibilities and enhancing the contextual benefits of voice assistants, thereby implicitly improving the overall user experience.

Reference list

Andrist, S., Tan, X. Z., Gleicher, M., & Mutlu, B. (2014). Conversational gaze aversion for humanlike robots. *Proceedings Of the ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/2559636.2559666

Bachorowski, J. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science*, *8*(2), 53–57. https://doi.org/10.1111/1467-8721.00013

Bainbridge, W. A., Hart, J. W., Kim, E., & Scassellati, B. (2008). The effect of presence on human-robot interaction. *IEEE Press*. https://doi.org/10.1109/roman.2008.4600749

Beer, J. M., Liles, K. R., Wu, X., & Pakala, S. B. (2017). Affective Human–Robot interaction. In *Elsevier eBooks* (pp. 359–381). https://doi.org/10.1016/b978-0-12-801851-4.00015-x

Bonfert, M., Zargham, N., Saade, F., Porzel, R., & Malaka, R. (2021). An evaluation of visual embodiment for voice assistants on smart displays. *Association for Computing Machinery*. https://doi.org/10.1145/3469595.3469611

Cambre, J., & Kulkarni, C. (2019). One voice fits all? *Proceedings of the ACM on Human-computer Interaction*, *3*(CSCW), 1–19. https://doi.org/10.1145/3359325

Carolus, A., & Wienrich, C. (2022). "Imagine this smart speaker to have a body": An analysis of the external appearances and the characteristics that people associate with voice assistants. *Frontiers in Computer Science*, *4*. https://doi.org/10.3389/fcomp.2022.981435

Castillo, S., Hahn, P., Legde, K., & Cunningham, D. W. (2018). Personality Analysis of Embodied Conversational Agents. *In Proceedings Ofthe 18th International Conference on Intelligent Virtual Agents*. https://doi.org/10.1145/3267851.3267853

Davis, R. O., Vincent, J., & Park, T. J. (2019). Reconsidering the voice principle with non-native language speakers. *Computers and Education/Computers & Education*, *140*, 103605. https://doi.org/10.1016/j.compedu.2019.103605

Fernandes, T., & Oliveira, E. (2021). Understanding consumers' acceptance of automated technologies in service encounters: Drivers of digital voice assistants adoption. *Journal of Business Research*, *122,* 180–191.

Funder, D. C. (2012). Accurate personality judgement. *Current Directions in Psychological Science*, *21*(3), 177–182.

Gelbrich, K., Hagel, J., & Orsingher, C. (2021). Emotional support from a digital assistant in technology-mediated services: Effects on customer satisfaction and behavioural persistence. *International Journal of Research in Marketing*, *38*(1), 176–193. https://doi.org/10.1016/j.ijresmar.2020.06.004

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, *37*(1), 81–88. https://doi.org/10.1080/02763869.2018.1404391

Khan, R., & De Angeli, A. (2009). The attractiveness stereotype in the evaluation of embodied conversational agents. In *Lecture notes in computer science* (pp. 85–97). https://doi.org/10.1007/978-3-642-03655-2_10

Kim, J., Kim, W., Nam, J., & Song, H. (2020). "I Can Feel Your Empathic Voice": Effects of nonverbal vocal cues in voice user interface. *CHI 2020 Late-Breaking Work*. https://doi.org/10.1145/3334480.3383075

Knote, R., Janson, A., Söllner, M., & Leimeister, J. M. (2019). Classifying smart personal assistants: An empirical cluster analysis. *Proceedings of the Annual Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2019.245

Mori, M. (2012). The Uncanny Valley: The original essay by Masahiro Mori. *IEEE Robotics & Automation Magazine*, *19*(2), 98–100. https://doi.org/10.1109/mra.2012.2192811

Revina, I. M., & Emmanuel, W. R. S. (2021). A survey on human face expression recognition techniques. *Mağalaẗ Ǧam’aẗ Al-malīk Saud : Ùlm Al-ḥasib Wa Al-ma’lumat*, *33*(6), 619–628. https://doi.org/10.1016/j.jksuci.2018.09.002

Reysen, S. (2005). Construction of a new scale: the Reysen likability scale. *Social Behavior and Personality*, *33*(2), 201–208. https://doi.org/10.2224/sbp.2005.33.2.201

Rosala, M., & Ramaswamy, S. (2024, April 22). *The Wizard of Oz Method in UX*. Nielsen Norman Group. https://www.nngroup.com/articles/wizard-of-oz/#:~:text=Summary%3A%20The%20Wizard%20of%20Oz,technologies%20at%20a%20low%20cost.

Sarigul, B., Saltik, I., Hokelek, B., & Ürgen, B. A. (2020). Does the appearance of an agent affect how we perceive his/her voice? *HRI ’20: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/3371382.3378302

Shamekhi, A., Liao, Q. V., Wang, D., Bellamy, R. K. E., & Erickson, T. (2018). Face Value? Exploring the effects of embodiment for a group facilitation agent. *CHI 2018 Paper*. https://doi.org/10.1145/3173574.3173965

Shi, Y., Yan, X., Ma, X., Lou, Y., & Cao, N. (2018). Designing emotional expressions of conversational states for voice assistants. *CHI 2018 Late-Breaking Abstract*. https://doi.org/10.1145/3170427.3188560

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113–117.

Wienrich, C., Ebner, F., & Carolus, A. (2022). Giving Alexa a face - implementing a new research prototype and examining the influences of different human-like visualisations on the perception of voice assistants. In *Lecture notes in computer science* (pp. 605–625). https://doi.org/10.1007/978-3-031-05412-9_41

Zhou, M. X., Mark, G., Li, J., & Yang, H. (2019). Trusting virtual agents. *ACM Transactions on Interactive Intelligent Systems*, *9*(2–3), 1–36. https://doi.org/10.1145/3232077

Appendix A:

Interview guide

Welcome to the interview. This study is for the bachelor’s thesis at UT. The whole project is to find answers about how avatar and verbal feedback may impact on the voice assistant experience. During the interview, I'll present two versions of a prototype—a Voice Interaction system—in different design settings. For each version, you'll have the opportunity to engage with them across various scenarios. I'll provide you with the scenarios beforehand, allowing you to prepare for interaction. Following each interaction, I'll inquire about your overall perceptions of the design and your thoughts on the Voice Assistant. The whole session will last 1 hour.

In the second round, we'll repeat the process with the other design version of the Voice Assistant. Finally, I'll ask a few questions about your general preferences and your feelings regarding these two versions, especially in terms of their appearance and voices.

Before we kick off the interview, I would kindly ask you bare in minds few information:

- This is still a basic prototype that can only interact with pre-set sentences. Sometimes, the prototype may not catch your words, so please be patient and repeat your command if necessary.
- Please kindly note that you'll be introduced to several scenarios, some of which may match reality, but they're only for test purposes.
- If you don't feel comfortable continuing, please let me know. You can pause or cancel this interview at any time. The functionality of this prototype is very limited; the purpose of this study is not to test the ability of the voice assistant but rather the design elements.
- Please keep this in mind when sharing your thoughts. However, your thoughts are valuable and important to this research.

- This interview will be recorded and archived in both audio and video formats, but access will be restricted solely to the researcher, only the transcript of the interview will be submitted to the University of Twente. Personal data will not be disclosed to third parties. All recorded data will be deleted after the completion of the thesis report, estimated to be by the end of July 2024. This study has been approved by ethic committee of university of Twente.
- Before we begin, do you voluntarily consent to participating in this interview, granting researcher Jiahui Zhang the right to record and utilise the data obtained for this study? Are you comfortable with both video and audio recording?
- Do you have any further questions?

Warm up:

- Have you ever used voice assistant products like Alexa, Google Assistant, Amazon, Apple's Siri, or Xiao AI?
- How frequently do you use them?
- In what scenarios do you typically utilise voice assistants: for example, in the car, at home, or on the phone?
- How do you feel about voice assistants in general?
- Do you usually interact with voice assistants that have a visual interface?
- Do you prefer voice assistants with a human-like voice?

Group 1 - Human Image vs AI Image

Scenario: greeting section

For your initial interaction with this Voice Assistant product, which we've named **Monday**, your first step is to greet him. Below are the sentences/words you can use for greeting; you can choose one and try to say it with him.

Greeting words	1. Nice to meet you on Monday! 2. Greetings 3. How are you 4. Hello Monday!

- Are you feeling ready to start the journey now? Do you have any questions before you start?

  Yes: let’s go!

  No: what is your question?

Scenario 1: Hang out in good weather.

Step 1:

You're looking out the window, and it's such a gorgeous day. You're thinking about heading out for a bite at a nice restaurant. But before you go, you want to make sure the weather's just right. You could ask your voice assistant something like:

Check the weather with VA.	1. How is the weather? 2. How is the weather today? 3. What is the weather? 4. Can you tell me the weather today? 5. Weather today? 6. Can you check the weather for me please?

Step 2: The weather seems really good outside, so you decide to head out. But you're not sure where to go. You're thinking about asking the VA for some recommendations. You could ask your voice assistant something like:

Ask VA for recommendations	1. Any suggestions? 2. Any restaurant recommendations? 3. Any place to recommend? 4. Do you have some recommendations?

Step 3: Imagine your voice assistant finds some information about restaurants on Google for you, but it won't open the Google page. After reviewing a few options, you select the 1st one and decide to go. Now, you want to set the navigation for this restaurant. But you are busy on your hands, so you want the VA to help set the navigation so you can check the arrival time. You could ask your voice assistant something like:

Set Navigation	1. Can you check and set the navigation to the first restaurant. 2. Can you set the navigation to the first restaurant. 3. Can you set the navigation.

Scenario 1 is done here.

First impression:

- Now the first scenario is done. What is your first impression of this voice assistant? Do you need a break? Do you have any questions?

Scenario 2: Mental support

Now, let's leave behind everything from the previous scenario, as it's a new day. Lately, you've been consumed by your thesis/work, but now that you've finished, you're feeling a bit lonely. Feeling the absence of someone to talk to at home, you turn to your voice assistant, hoping it can offer some support.

Step 1: You first need to wake it up by saying:

Wake it up	1. Hey Monday! 2. Hi Monday. 3. Hello Monday 4. Monday

Step 2: You can tell him you are not feeling well.

Ask the VA for help.	1. I don't feel well. 2. I don't feel good. 3. I still don't feel good. 4. I still don't feel well. 5. I'm not feeling well. 6. I'm not feeling good.

Step 3: You feel like maybe music can help. So you decide to ask VA to play some music for you.

Ask the VA to play the music	1. Can you play the music of? 2. Can you play the music via? 3. Can you play the music? 4. Can you play some music for me? 5. Can you play the music for me?

	6. Can you play some music?

Insert the music here

Step 4: The music ends but you are not feeling better, you feel lonely now. You may want to tell the Voice assistant to give you some help. (and you feel like maybe calling someone can help.)

Ask the VA for help	1. I feel a bit lonely. 2. I feel lonely. 3. I am a bit lonely. 4. I feel alone. 5. I feel I am a bit lonely 6. I feel I am lonely 7. I still feel a bit lonely 8. I still feel lonely

If yes - finish scenario 2

If no - jump to scenario 3

Now you feel like you may need a little nap. Maybe around 1 hour, so you ask the VA to set an alarm in one hour for you. You may say to it as:

Ask the VA for help	1. Can you set an alarm in 1 hour? 2. set an alarm in 1 hour. 3. set an alarm after 1 hour 4. Set the alarm in 1 hour. 5. set the alarm after 1 hour. 6. Set alarm in 1 hour

	7. set alarm after 1 hour 8. set the alarm for me? 9. Set alarm in an hour 10. Set alarm in 60 min.

One round is finished.

Overall impression:

Participants are kindly reminded to focus solely on design aspects rather than functionality when providing their feedback.

- Previously, you mentioned the impression as xxx (insert the answer from previous). How do you feel now, specifically in terms of its design?
- Why do you feel this way? Which element of the design evokes such a feeling, whether it's related to the voice or face?
- How do you feel emotionally when interacting with it?
- Is there any aspect that triggers a specific emotion, such as awkwardness, warmth, kindness, or sadness? If so, why do you think that is?

For voice

- What do you notice about the voice feedback during interaction, such as its intonation, pitch, and tone?
- How do you perceive the voice feedback, particularly regarding the intonation? Do you find it more positive or negative?
- When do you feel positively about it, and what contributes to that sentiment?
- Conversely, when do you feel negatively about it, and what factors contribute to that?
- Give a rate from 1 - 7 to the voice, 1 represents I don't like it at all, and 7 represents I like it very much. How much will you rate it?

For face:

- What do you observe from the facial reactions during interaction, such as its emoji and body language?
- How do you perceive facial design? Do you find it more positive or negative?
- In what instances do you feel positively about it, and why?
- Conversely, when do you feel negatively about it, and what contributes to that sentiment?
- Give a rate from 1 - 7 to the face, 1 represents I don't like it at all, 7 represents I like it very much. How much will you rate it?
- On a scale of 1 to 7, where 1 means 'I don't feel any emotions at all' and 7 means 'I feel a lot,' how emotionally impactful do you find this prototype overall?

Do you need a break?

Start the second version - the same.

Final Comparison:

1. Compare these two versions of design and share which one you prefer and why.
2. Are there any situations where you might prefer the first version?
3. Similarly, are there any scenarios where you might prefer the second version?
4. How do you perceive the differences between these two versions?
5. Which version do you feel a stronger emotional connection to?
6. Do you like or dislike either version, and what factors influence your preference?
7. Human as a ChatGPT AI base, or robotics as normally you use, as Siri or Alexa. 1-7
8. Generally speaking, do you find human-like facial expressions and tones important for a voice assistant product? Why or why not? Please provide separate responses for facial expressions and voice.
9. When do you prefer a voice assistant with human-like facial expressions or tones, and when do you not?

10. Do you believe emotionally expressive features are important for a voice assistant product? Why or why not?
11. How do you typically perceive emotional support from voice assistant products?

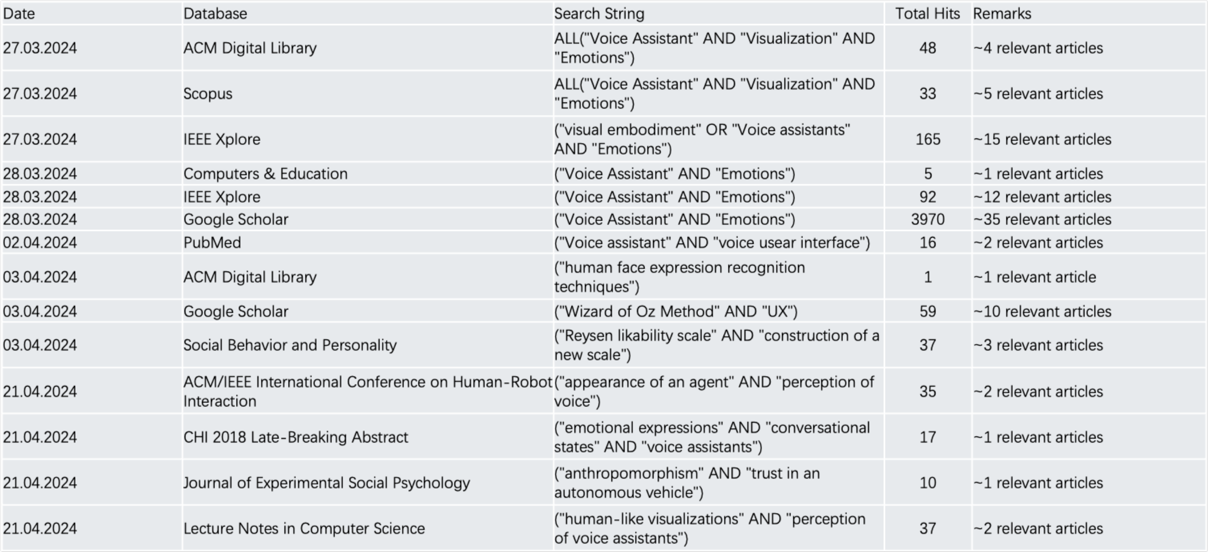
Wrap up

Thank you for your participation and for generously sharing your time with us. Is there anything else you'd like to share before we conclude this interview?

Wishing you all the best and have a wonderful day!

Appendix B:

Reference Log



Date	Database	Search String	Total Hits	Remarks
27.03.2024	ACM Digital Library	ALL("Voice Assistant" AND "Visualization" AND "Emotions")	48	~4 relevant articles
27.03.2024	Scopus	ALL("Voice Assistant" AND "Visualization" AND "Emotions")	33	~5 relevant articles
27.03.2024	IEEE Xplore	("visual embodiment" OR "Voice assistants" AND "Emotions")	165	~15 relevant articles
28.03.2024	Computers & Education	("Voice Assistant" AND "Emotions")	5	~1 relevant articles
28.03.2024	IEEE Xplore	("Voice Assistant" AND "Emotions")	92	~12 relevant articles
28.03.2024	Google Scholar	("Voice Assistant" AND "Emotions")	3970	~35 relevant articles
02.04.2024	PubMed	("Voice assistant" AND "voice usear interface")	16	~2 relevant articles
03.04.2024	ACM Digital Library	("human face expression recognition techniques")	1	~1 relevant article
03.04.2024	Google Scholar	("Wizard of Oz Method" AND "UX")	59	~10 relevant articles
03.04.2024	Social Behavior and Personality	("Reysen likability scale" AND "construction of a new scale")	37	~3 relevant articles
21.04.2024	ACM/IEEE International Conference on Human-Robot Interaction	("appearance of an agent" AND "perception of voice")	35	~2 relevant articles
21.04.2024	CHI 2018 Late-Breaking Abstract	("emotional expressions" AND "conversational states" AND "voice assistants")	17	~1 relevant articles
21.04.2024	Journal of Experimental Social Psychology	("anthropomorphism" AND "trust in an autonomous vehicle")	10	~1 relevant articles
21.04.2024	Lecture Notes in Computer Science	("human-like visualizations" AND "perception of voice assistants")	37	~2 relevant articles