Humanising in-home Conversational Agents: Improving Anthropomorphism and Emotional Prosody

KIPKEMBOI CEPHARD, Swansea University, United Kingdom

ABSTRACT

Conversational agents have widely been accepted for our day-to-day use. While the growth has been widely accepted, this study finds the need to improve the interaction between this agent and human beings through emotional intelligence. Previous research has found that most users are enthusiastic about the new gadget but seem to find it boring later. Emotional prosody and anthropomorphism have played a big role in this result and need to be improved. This study considers "Shneiderman's Golden Rules of Design" to produce a novel conversational agent that conforms to HCI design principles. We formulate research questions and propose to measure our hypothesis using interviews and a Likert Questionnaire on 20 participants between 15 and 35 years old. We use MMDAgent, Balabolka, and FL Studio to embed emotions into a synthetic voice. This study indicates how emotions elevated by acoustics can be depicted in the form of graphs, waveforms, and audio spectrums. The findings of this between-group in-lab experiment will bridge the gap in the literature on the subject of human emotions toward conversational agents.

CCS Concepts: • **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Additional Key Words and Phrases: Humanising Synthetic Voice, Human Acceptance, Conversational Agents

ACM Reference Format:

Kipkemboi Cephard. 2024. Humanising in-home Conversational Agents: Improving Anthropomorphism and Emotional Prosody. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION: DESCRIBING THE PROBLEM

Voice-based agents have become increasingly popular in human everyday life. These agents are essential in various settings, such as in an office, home, or a self-checkout unit at a convenience store. Despite the numerous benefits, these agents come with a caveat that users have yet to accept completely. One of the primary challenges users face is the synthetic voice used in Text-to-speech (TTS) technology, which has very low prosody [7]. Although options such as Concept-to-speech (CTS) and the use of pre-recorded feedback from trained voice actors to produce a more natural voice tend to produce better results, they require significant time and resources to produce [1, 7].

Even though TTS voice quality has improved in recent years, users still experience a noticeable loss of interest after some time, particularly with children who expect their Conversational Agents (CAs) to grow with them [9]. Compared to social robots that use audio, haptic, and visual cues to communicate, CAs have more difficulty achieving high anthropomorphism since they use voice as the only interaction modality. This feature of utilizing one input modality hinders the conversational agents from forming deep connections with the users.

In the era of artificial intelligence and machine learning, CAs have grown closer to human likeness than before, although they still struggle to achieve the same form of social intelligence portrayed in human beings. Additionally, their lack of personality and low sense of "life" is highly evident in their responses to questions relating to sensitive topics. Despite their growing popularity among different age groups and contexts in human life, CAs tend to give users formal responses that are unsuitable for casual conversations. Humans rarely use formal conversations, for example, between friends, family, or relatives.

Despite advancements in Natural Language Processing (NLP), which helps CAs mimic human communication, they still struggle to understand social cues such as emotional prosody in sentence pauses, speech tone, and sound timbre. NLP also requires users to take turns with the device during the interaction, unlike how humans multitask seamlessly in back-and-forth conversations [8]. Furthermore, the human voice is too complex and needs to be conveyed and measured beyond phrases and sentence construction. This complexity arises primarily from socio-diverse nuances like age, gender, and race that impact the user's accent. Furthermore, humans can form healthy emotional bonds with non-human figures, such as dogs, which have successfully been interwoven into the human social-cultural environment [13]. Similar relations can be achieved for in-home agents when they can convey an understanding of contextual emotions and feelings through their responses.

"This research paper explores and addresses the gap in how CAs perceive and portray context when communicating. It also highlights the importance of avoiding the 'uncanny valley'". [20]. While envisioning to achieve a model that sounds natural, it is crucial to be cautious of eerie and uncomfortable feelings [5] that may arise with a more socially intelligent agent design.

2 LITERATURE REVIEW AND RELATED WORK

In this study, we conduct an integrative literature review of prior academic work, focusing on the progress made in improving synthetic voice and the drawbacks of using this approach. This study also highlights the progress achieved and what needs to be improved in conversational agents' emotional prosody and anthropomorphism. With the conclusions derived from the findings of prior work, we find gaps and generate research questions that will be the center of research in this study.

Synthetic voice is generated by conversational agents traditionally through the Text-to-Speech technique. This can be adjusted for penalization in terms of speed rate and gender. These days, there are two predominant methods of synthesizing speech: parametric synthesis and unit selection, which were tested side by side with the "big-five" human behaviors by Aylett et al. [3]. However, these approaches failed to deliver similar traits they were tested against [3]. Prior studies have agreed that this option is more fitting for existing agents due to its flexibility and affordability. However, they also note that synthetic voice lacks the quality of utterance and prosody depicted by human voice [7]. Furthermore, a recent study by Iizuka and Mori [12] illustrates how to naturalize and improve the acceptance of synthetic voice through spontaneous feedback, which received better reception than a synthetic voice that reads the input. Studies have also proven that the context of the synthetic voice being used can improve its impact on how users perceive it. Dubiel et al. [6] study demonstrates that adding a debating nature to the voice makes it more lively and pleasant. However, the study needs to find tangible results on the expected interaction feedback received from the users.

The attempt to make conversational agents more appealing and acceptable with a more human-like likeness, in this case through a more realistic voice with high quality indistinguishable from human recorded voice, is the technique of Anthropomorphism and Social Intelligence. It also enables artificially generated responses to sound into context and bidirectional making machine social agents [12]. In the empirical study of Li and Suh [15], it is evident that researchers have yet to fully agree on a universally acceptable amount of human likeness that can be adopted into these agents without social drawbacks [15]. Previous studies have concluded that integrated intelligence and personalizing CAs increase adaptability and acceptance Moussawi et al. [22]. The feeling of relatable human traits is not simply implied but perceived by users as suggested in the study of [21, 22]. Hence, it is inferred that the success of these relationships between humans and CAs does not entirely rely on the quality and "gender" embodied in the agent.

On the contrary, some other studies tend to differ from these findings based on the adverse impact on user motivation to continue using these agents. The hypothesis of the study conducted by Gursoy et al. [11] dwells on the negative impacts caused by improving human traits in a system to the point when users tend to feel like the device's core functionalities have been neglected Lin et al. [16].

Finally, emotional Prosody is a phenomenon that has yet to be vastly tackled by existing work, with just a few scratching the surface and providing limited background research. However, existing materials provide intriguing discussions that are important to this study. In a study to create speakers with human-like emotions by breaking down nonverbal cues embedded in words rather than the entire scope of the sentence, the study of Lu et al. [19] has proven that speakers can attain a more emotional state while preserving the prerequisite design functionalities. Another study highlights the perceived emotions portrayed in human-to-agent conversation through a single utterance in the contextual semantics and acoustics [18]. However, due to the limited training dataset, the results can not be entirely valid, regardless of the findings. Generally, the emotionally aware framework could still portray the expected emotional cues even with these limitations.

2.1 Research Questions

The main aim of this study is to understand what emotions are present and to what extent the users find it acceptable for CAs to portray human-like emotions in their speech. The literature review's findings prove that for a successful study, the agent has to be able to perceive emotions portrayed in the users' sounds and respond in the same contextual manner. Furthermore, the literature depicts insufficient knowledge of emotional prosody regarding emotional agents. Hence the need for this study where, we will focus on the scope of these questions:

> ***RQ 1 :*** What emotions make synthetic voices sound natural and motivate users to interact with conversational agents without inducing discomfort?
>
> ***RQ 2 :*** Do naturally sounding synthetic improve trust and build a stronger bond from the user towards the conversational agent and to what extent is the bond formed?

3 PROTOTYPE DESIGN AND OVERVIEW

To ensure the prototype meets the prerequisites of the research questions, this study will not utilize any existing speech recognition machine learning models. Although MMDAgent offers *Julius* [14], an inbuilt speech recognition plugin, the study will use the "Wizard of OZ" prototyping approach, which proved to be successful in the study of Iizuka and Mori [12]. This is a human-computer interaction (HCI) prototyping technique where a human controls the agent remotely to answer prompts and perform tasks like the traditional agents. Still, the user interacting with the agent is unaware of the wizard. This study considers "Shneiderman's Golden Rules of Design" and overall heuristic design principles.

3.1 Universal Usability

This study aims to suggest a novel conversational agent capable of interacting with users of different skills, from experts to novices. First, the agent interacts with different subjects, offering flexibility and plasticity from commentating on children's cartoon characters to answering complex arithmetic questions of an office manager. While it seems a far-fetched idea, this design principle will enable the agent to easily adapt to different environments and get accepted by different users. This study demonstrates that the agent easily engages in official and casual conversations.

3.2 Preventing Errors

Another unique feature of this agent is the ability to guide the direction of the conversation and lead the users toward desired goals without them having to struggle to explain what they need every time. During the interaction, the agent asks the user a series of multiple-choice questions, and all the user needs to do is answer with the offered choices of command. This will reduce users' habit of asking irrelevant or out-of-context questions. This design principle will be useful in a child-agent conversation where the agent assists the child with revisions on a specific topic. The main objective of the interaction will be achieved with ease. Furthermore, suppose the day's goal was not met, or the child took a short break from the interaction. In that case, the agent will retain the state of the conversation and will persuade the child to continue with their pending objectives before moving forward to other topics. Although this might be daunting to some users, if their tone seems to portray anger or frustration, the agent will guide the user on how to deal with the problem by offering solutions to similar questions or giving hints to the existing questions.

3.3 Internal Locus of Control

The agent is also designed to behave in a warm, friendly, and patient manner, allowing it to mimic the behavioral approach used by therapists. Based on the perceived emotions of the users, the agent allows the users to guide the flow of the conversation by supporting their emotional needs respectfully and encouragingly. In joyful scenarios where the user is enthusiastic and happy, the agent will respond with responses aligned with joy and happiness. On the other hand, in the scenario of pain, anguish, and hopelessness, the agent's tone, speech rate, and acoustics automatically change to depict a tender and welcoming to adapt to the scenario. Furthermore, the agent does not interrupt the user during the conversation; it encourages them to be free and hopeful while the agent maintains respect and personal boundaries. In this scenario, the user has full control of the direction of the conversation to ensure their emotional needs have been catered for in a state-of-the-art manner.

3.4 Informative Feedback

This principle is of the utmost importance during the design phase. Existing conversational agents have consistently failed to uphold this principle due to their generic and repetitive responses, which lack empathy. Unfortunately, this is why human beings have not widely embraced conversational agents. This agent has been designed to differentiate between actual conversations and system prompts to improve this drawback. The prototype will be able to offer system prompts to the user without acting arrogantly. Furthermore, error sounds would help the users differentiate feedback from the conversation and system feedback to errors. These errors exclude situations when the agent cannot comprehend what the user is saying. These errors will include system updates, sensitive topics, and low battery, among other non-conversational errors. On occasions when the user is not audible enough, instead of the agent requesting the user to be more audible, it analyses the average noise of the surroundings to the sound of the speaker. It then increases its voice by a few decibels to encourage the user to raise their voice subconsciously. A previous study by Bottalico et al. [4] has also proven that human beings will naturally increase their voice whenever interacting in a loud environment. By doing so, the agent will have avoided shifting the blame of inaudibility from the user to the surroundings, improving the user experience.

3.5 Usability Heuristics

This prototype also proposes to reduce errors during natural language processing, especially in scenarios that confuse these agents. For example, agents always struggle with acronyms and homophones as observed in the table below *Table 1*.

Table 1. Homophones that are hard to comprehend in Natural Language Processing

Homophones	Wrong Sentence Example	Correct Word
bare – bear	You should not hold hot pots with your **bear** hands.	bare
hear – here	They used to come **hear** for a cup of tea.	here
write – right	When did you **right** your essay?	write
ate – eight	**Ate** men joined the fun club this month.	eight
sea – see	I **sea** the **see** from a mile away.	interchange see and sea

4 STUDY DESIGN

To guide the study towards answering the research question, we formulate these hypotheses to ensure our research adheres to the scope of our problem statement. The hypothesis will be tested by using t-tests to compare the mean of the two groups in the study.

H1 : An emotionally intelligent conversational agent can adapt to the context and portray the right feelings during a conversation.

H2 : Unlike text-based conversational agents, speech-based conversational agents can induce natural back-and-forth conversations with the users without losing the context of the conversation.

4.1 Procedure

This study will be conducted through between-group experiments to test the prototype against existing conversational agents. There will be three sessions in this in-lab experiment. First, the participants will be divided into two groups. The first group will interact with the proposed conversational agent, which is uniquely controlled by a 'wizard,' and the other group will interact with an existing conversational agent like Siri and Alexa. This approach of dividing the participants into groups subjected to only a single conversational agent's scenarios ensures that they will not be subjected to stress, thus reducing fatigue and improving their quality of life. Each session will take 45 minutes: 30 minutes for human-to-agent interaction and 15 minutes for a short structured interview. This ensures that data distortion and participant bias are avoided, improving the accuracy and admissibility of the data.

After a thirty-minute break, there will be a brief ten-minute discussion on what needs to be improved in the prototype, and we will take notes of the overall feedback. Finally, the "wizard" will make the changes and adjustments as requested in the second session while adhering to the same procedure and duration as the first session. This study will also adopt optional feedback sessions. At the end of the experiment, close-ended questionnaires will be issued to measure the emotional prosody and anthropomorphism of the prototype.

4.2 Participants

20 participants between 10 and 35 years of age from a local community will be recruited through flyers and social media message invitations. The recruitment process will be voluntary and self-administered by the prospective volunteers. Before issuing these invitations, we will seek permission from the local authority and the organizing body of the university to conduct this experiment. Bearing in mind that all cultural norms are diverse, the sensitive opinions of the participants are respected. This study intends to onboard 4 secondary school students, 7 undergraduates, and 9 graduates who have prior experience with existing conversational agents. To ensure generalization, ethnicity, and social-cultural beliefs will be considered; furthermore, 40% of the participants will be female.

The participants will receive a comprehensive consent form explaining their role in the experiment, what the experiment is all about, how it will benefit the research community, and finally, how their data will be handled securely with concealed identity. Participants will not use their actual names and will be encouraged to use names that do not expose their identities and backgrounds. Parents will sign the consent forms on behalf of their children. The form will also inform them how long each session will be and how their behavioral logs will be captured in video form. The videos will blur out faces for confidentiality and only sound, and body movements will be used for the research. Finally, the document will highlight the non-monetary compensation of free meals during break sessions and gifts after the completion of the experiment to appreciate them for their valuable contribution.

4.3 Setup and Materials

This study will use structured interviews and close-ended questionnaires to gather feedback and data to test the research questions. This approach is chosen because of the need for more background knowledge on this topic. Both methods will contain 10 questions, each focusing on anthropomorphism, emotional prosody, and the human level perceived in the synthetic voice of the two agents. The level of the prototype acceptance will be weighed against the level of the acceptance of Siri and Alexa's naturalness. The results from these interview questions will be used to measure the proposed hypothesis. 4 questions will be directed toward the naturalness of the voice, 3 on emotional prosody, and 3 on anthropomorphism.

Table 2. Likeart Questionnaire to measure synthetic voice acceptance

Rate your experience from 1 to 7. 1 is (**STRONGLY DISAGREE**) and 7 is (**STRONGLY DISAGREE**)	1	2	3	4	5	6	7
The conversational agent understood my emotions.							
The behavior of the agent's voice seemed to contain human nature.							
The conversation was fun and pleasant.							
The likeness of the voice to the human sound made me feel irritated.							
The system prompts and notifications did not deviate from the nature of the conversation.							
The agent gave me relevant feedback on my questions.							
The agent guided me correctly when assisting me to learn a new topic.							
The agent respected my opinion.							
The agent correctly understood the emotions I was portraying during the conversation.							
The general experience with the agent was easy.							

The consent forms will be shared with participants in written and soft copy. This prototype has been designed in three phases: Traditional Text-To-Speech synthesis with Balabolka [2] and MMDAgent [14], Recording the speech into Digital Audio Workstation (DAW) and adding acoustics in the voice to mimic human breath and quality in FL Studio [17]. The final voice will be outputted using a generic speaker to avoid bias in brands and marketing. The setup of this study will be well suited to a safe environment; with consent from the authority, a small social hall with access to all amenities such as transport and cafes for breaks will enable participants to relax during breaks and not feel constricted by the experiment. To improve participants' quality of life and reduce stress, two different rooms. Furthermore, this will ensure the setup is organized and well-structured, saving time for arrangement and session allocation. Finally, two camcorders and two external microphones will record the sessions.

4.4 Measurements and Data Analysis

In this section, the study measures these variables according to the formulated hypothesis, and confounding variables will be controlled by randomly assigning participants to the two groups. We will combine the qualitative data from the Likert questionnaire and the qualitative data from the structured interview sessions. All the video files will be converted to MP4 and audio to WAV formats to retain quality and ensure the data entered into the dataset is consistent; after transcribing the media files, the variables will be coded into a list where every variable is entered into their own column. To ensure data quality, the data will be cleaned rigorously. First, missing data will not be included in the dataset, and the duplicate data will be removed. Secondly, due to the dataset's size, removing outliers would be prudent to avoid ambiguous results that might not accurately depict the actual findings. To visualize emotions in the dataset's coded voices, generate waveform and audio spectrum such as **Figure 1** and **Figure 2**.Furthermore, a plotted waveform MFFCC figure such as **Figure 3** can be used to distinguish the impact of acoustics in the synthetic voice visually.

The findings of these measurements will attempt to answer the research questions and fill the gap that has existed from previous works in this field. In every session, the agent will try to persuade the participants to engage in specific topics and subject areas. The five categories of topics will revolve around family, friends, work, jokes and games, and childhood memories. Each scenario will have 2 approaches: 1 that invokes happy feelings and one that invokes another emotion in a random order based on the direction of the conversation for the agent to adapt easily to context. Heuristic data between the existing conversational agents and the prototype will be collected and measured using the Likert scale, with the following variables divided into columns and categories.

Table 3. Variables and Categories

Category	Independent Variables	Dependent Variables
Emotional Prosody	Tone	Satisfaction, Engagement, Emotional Dependency
Emotional Prosody	Acoustics	Satisfaction, Engagement, Emotional Dependency
Anthropomorphism	Speech Rate	Satisfaction, Engagement
Anthropomorphism	Gender of the Synthetic Voice	Naturalness, Satisfaction

4.5 Design Rationale

Unlike existing CAs, the proposed design choice of the prototype in this study is to enable the agent to initiate conversations. The main advantage of doing so is evident in the study conducted by Iizuka and Mori [12] on a spontaneous agent that found that this approach will ensure a more back-and-forth natural conversation. Furthermore, subtle acoustics such as reverberation will be added to the output to ensure a smoother-sounding speech as advised in the study of [10]. The study found that acoustics improve the quality and nature of sound with more features like breath. These acoustics will be the same across all scenarios to ensure consistency in the sound feedback and avoid biased results. **Figure 1** and **Figure 2** portray the impact of acoustics on a synthetic voice.**Figure 1** conveys a longer waveform in the audio spectrum, representing the acoustics that increase emotions, which relates to the study of Parra-Gallego and Orozco-Arroyave [23]. This study captures video data to increase the accuracy of the emotions portrayed by the participants. To get a clear understanding of the participant's emotions, these behavioral logs, when measured together, offer a clearer understanding of the emotions portrayed. A conversational agent has benefited human beings in many ways so far. Through automation and personalizing, we can find great assistants in the palm of our hands. Children can learn easily by directly surfing the internet, and adults can set clock reminders without touching their devices. From the most trivial duties to the non-trivial assignments, these agents will continue to improve and reduce the human workload on repetitive tasks.

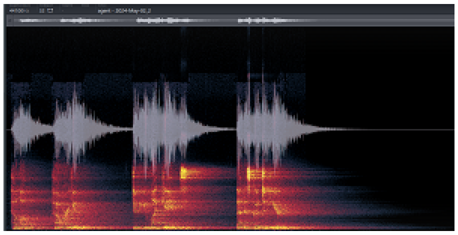

Fig. 1. Rich acoustic waveform

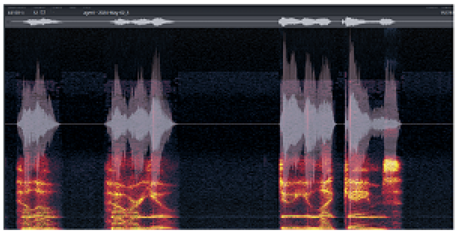

Fig. 2. Deficient acoustic waveform

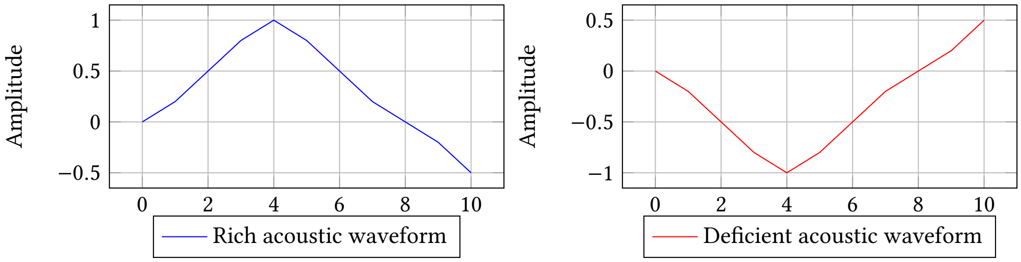

Fig. 3. Waveform Comparisson

5 LIMITATIONS AND FUTURE WORK

Since a human controls this prototype, the wizard will theoretically depict more accuracy than a machine-learning neural network. The artificially intelligent design will require a neural network that has close to 98% accuracy to match the precision in speed and spontaneity of the wizard-controlled prototype. Furthermore, this study ensures the state-of-the-art generalization of participants. Still, future studies are encouraged to replicate the same experiment in a field study and measure the hypothesis and research questions under different conditions. Future studies should create an AI model incorporating the text-to-speech approach designed in this study. To successfully replicate this study, researchers from the HCI department, psychology, and sound engineering must collaborate throughout the research process due to the broad and complex nature of anthropomorphism and emotional prosody. This study only briefly mentions successful pet-human relationships. Future research can identify traits that make pets appealing and integrate them into conversational agents.

6 CONCLUSIONS

In conclusion humans' emotions will always fluctuate, even in an interaction amongst themselves, resulting in conflict and dissatisfaction. This agent's emotional awareness will correctly tackle these situations to ensure the emotions do not go wrong. Although synthetic voice has been used as the primary modality for in-home agent conversation, this study highlights the limitations and gaps in the existing literature and makes suggestions on how to fill these gaps. TTS seems to be the most accessible and affordable way to create synthetic voices, but the users' needs have not been entirely met at the moment. This study provides a different approach through a strong emotional bond between the in-home agent and users, especially children. The design targets to lean towards casual interactions rather than traditional official conversations, making the agent more approachable and user-friendly. Anthropomorphism and emotional prosody play a significant role in the success of in-home agent interactions, and the findings of this study will give a remarkable direction toward fulfilling these considerations. The in-lab experiment aims to provide first-hand user experience with an emotionally aware conversational agent that adapts to changes in context and mood. The prototype's design ensures that sound quality and acoustics remain consistent throughout the conversation, even with different emotions. The findings of this study will be an essential revelation to researchers and producers of conversational agents. First, the researchers will be intrigued by the findings of this study and can easily follow the directions given in this paper for more rigorous research. Lastly, the creators of conversational agents can appreciate the need to embed emotional prosody in newer versions of the agents to improve the interactions between these agents and humans.

REFERENCES

[1] Amal Abdulrahman and Deborah Richards. 2022. Is natural necessary? Human voice versus synthetic voice for intelligent virtual agents. *Multimodal Technologies and Interaction* 6, 7 (2022), 51.

[2] Dirham Gumawang Andipurnama, Dikdik Mantera Wiguna, Budi Susetyo, and Ranti Novianti. 2022. BALABOLKA Software to Improve the Ability to Access Electronic Learning Resources for Visual Impairment Students. *Journal of ICSAR* 6, 2 (2022), 230–236.

[3] Matthew P Aylett, Alessandro Vinciarelli, and Mirjam Wester. 2017. Speech synthesis for the generation of artificial personality. *IEEE transactions on affective computing* 11, 2 (2017), 361–372.

[4] Pasquale Bottalico, Simone Graetzer, and Eric J Hunter. 2015. Effects of voice style, noise level, and acoustic feedback on objective and subjective voice evaluations. *The Journal of the Acoustical Society of America* 138, 6 (2015), EL498–EL503.

[5] Marialucia Cuciniello, Terry Amorese, Claudia Greco, Zoraida Callejas Carrión, Carl Vogel, Gennaro Cordasco, Anna Esposito, et al. 2023. A Synthetic Voice for an Assistive Conversational Agent: A Survey to Discover Italian Preferences regarding Synthetic Voice's Gender and Quality Level. *Human Behavior and Emerging Technologies* 2023 (2023).

[6] Mateusz Dubiel, Martin Halvey, Pilar Oplustil Gallegos, and Simon King. 2020. Persuasive synthetic speech: Voice perception and user behaviour. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–9.

[7] Jonathan Ehret, Andrea Bönsch, Lukas Aspöck, Christine T Röhr, Stefan Baumann, Martine Grice, Janina Fels, and Torsten W Kuhlen. 2021. Do prosody and embodiment influence the perceived naturalness of conversational agents' speech? *ACM Transactions on Applied Perception (TAP)* 18, 4 (2021), 1–15.

[8] Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S Bernstein. 2018. Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[9] Radhika Garg and Subhasree Sengupta. 2020. Conversational technologies for in-home learning: using co-design to understand children's and parents' perspectives. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[10] Christer Gobl and Ailbhe Ní Chasaide. 1992. Acoustic characteristics of voice quality. *Speech Communication* 11, 4-5 (1992), 481–490.

[11] Dogan Gursoy, Oscar Hengxuan Chi, Lu Lu, and Robin Nunkoo. 2019. Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management* 49 (2019), 157–169.

[12] Takahisa Iizuka and Hiroki Mori. 2022. How Does a Spontaneously Speaking Conversational Agent Affect User Behavior? *IEEE Access* 10 (2022), 111042–111051. https://doi.org/10.1109/ACCESS.2022.3214977

[13] Maki Katayama, Takatomi Kubo, Toshitaka Yamakawa, Koichi Fujiwara, Kensaku Nomoto, Kazushi Ikeda, Kazutaka Mogi, Miho Nagasawa, and Takefumi Kikusui. 2019. Emotional contagion from humans to dogs is facilitated by duration of ownership. *Frontiers in Psychology* 10 (2019), 1678.

[14] Akinobu Lee, Keiichiro Oura, and Keiichi Tokuda. 2013. MMDAgent—A fully open-source toolkit for voice interaction systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 8382–8385.

[15] Mengjun Li and Ayoung Suh. 2021. Machinelike or humanlike? A literature review of anthropomorphism in AI-enabled technology. In *54th Hawaii International Conference on System Sciences (HICSS 2021)*. 4053–4062.

[16] Hongxia Lin, Oscar Hengxuan Chi, and Dogan Gursoy. 2020. Antecedents of customers' acceptance of artificially intelligent robotic device use in hospitality services. *Journal of Hospitality Marketing & Management* 29, 5 (2020), 530–549.

[17] Image Line. 2015. Fl Studio.

[18] Yuchen Liu, Haoyu Zhang, Shichao Liu, Xiang Yin, Zejun Ma, and Qin Jin. 2023. Emotionally Situated Text-to-Speech Synthesis in User-Agent Conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5966–5974.

[19] Chunhui Lu, Xue Wen, Ruolan Liu, and Xiao Chen. 2021. Multi-speaker emotional speech synthesis with fine-grained prosody modeling. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5729–5733.

[20] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine* 19, 2 (2012), 98–100.

[21] Sara Moussawi and Marios Koufaris. 2019. Perceived intelligence and perceived anthropomorphism of personal intelligent agents: Scale development and validation. (2019).

[22] Sara Moussawi, Marios Koufaris, and Raquel Benbunan-Fich. 2021. How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents. *Electronic Markets* 31, 2 (2021), 343–364.

[23] Luis Felipe Parra-Gallego and Juan Rafael Orozco-Arroyave. 2022. Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments. *Digital Signal Processing* 120 (2022), 103286.

Received 4 May 2024; revised 5 May 2024; accepted 6 May 2024