Impact of synthetic voice and avatar animation on the usability of a dialogue agent for digital health monitoring

Hardik Kothare, Doug Habberstad, Michael Neumann, Sarah White, David Pautler and Vikram Ramanarayanan

Abstract We evaluate the impact of avatar visual animation and synthetic voice on how participants with and without depression rate the user experience (UX) and usability of a virtual conversational agent, Tina, for digital health monitoring. We propose a novel experimental design to investigate whether: (a) avatar gesture animations improve the UX of the interactive health monitoring platform relative to a static graphic, (b) text-to-speech (TTS) voice prompts are rated as positively as human voice prompts, if not better, and (c) the UX is equally good for participants across different severity levels of depression symptoms (as measured by their self-rated Personal Health Questionnaire or PHQ-8 score). Tina engaged 458 crowdsourced participants in an interactive conversation consisting of a battery of speech tasks in one out of four study arms (corresponding to all combinations of the two experimental conditions: with and without animation, and with and without TTS). We observe that participants, especially males, rate the usability of the platform higher when the virtual agent is animated. Additionally, participants with moderate to severe depression rate the usability of the platform higher when TTS prompts are present. Our findings support the use of avatar animations (allowing for improved patient engagement) and synthetic voices (allowing for greater scalability and flexibility) in conversational dialogue-based systems for digital health monitoring.

1 Introduction

Speech-based technologies are becoming more and more ubiquitous in the world we live in. Conversational agents or dialogue systems have made their way into millions of homes through phones, computers, smart TVs, smart speakers, home assistants, wearables and appliances [1]. Healthcare is one domain that has seen an increase in

Modality.AI, Inc., San Francisco, CA, USA
e-mail: hardik.kothare@modality.ai

the use of conversational agents [2, 3, 4]; indeed, there is an increasing body of work supporting the case for speech as a biomarker of neurological and mental health [5]. The COVID-19 pandemic further underscored the utility of conversational agents in healthcare support [6] and remote patient monitoring [7]. In earlier work, we presented Tina, a conversational agent powered by the Modality platform [8, 9]. The Modality platform is a cloud-based multimodal dialogue platform that conducts automated interviews to assess neurological and mental health disorders.

Voice prompts for spoken dialogue agents are often recorded by voice actors to enhance the naturalness of the conversation. The recording, selection and integration of these voice prompts can be a time-consuming process that requires work by voice actors, speech scientists, software engineers and sometimes expert translators and external stakeholders. Any improvements or changes thus require a lengthy development process. One way to ease this process is by the adoption of text-to-speech (TTS) synthesis. TTS enables technologists to use synthesised voices as a substitute for human voice recordings, and also in real-time utterance generation. The easy generation of TTS prompts and their integration with software systems thus enables rapid development cycles. Previous work has compared human and TTS voices and the effects they have on listeners — like emotion perception and intimacy, learning outcomes, voice quality evaluation and preference — in various scenarios [10, 11, 12, 13, 14]. Some work suggests that socio-cognitive mechanisms may have a role to play during interactions between humans and voice-based virtual agents [15, 16]. When human therapists engage with depressed individuals, they adopt the speech features of the speakers with depression, perhaps to convey emotional empathy [17]. It is therefore important to consider how individuals with depression perceive voice prompts while interacting with a digital health monitoring agent used for conversational assessments.

When offered different interaction modalities (like text-only, static image plus text, animated and animated with non-verbal speech), users are observed to establish stronger social bonds with the animated options [18]. Virtual conversational agents capable of non-verbal communication have been shown to improve user perception of friendliness, trust and rapport [19]. When the conversational agent is interacting with users in an interview setting, non-verbal gestures have been shown to reduce the anxiety levels of candidates and increase speaking time [20]. However, any animation of the avatar needs to be minimal so as not to act as a distraction in the conversation [21]. Individuals with anxiety have been shown to have a reduced sense of rapport with the virtual agent when the avatar's responses are non-contingent on the speaker behaviour [22].

In this work, we present results from a crowdsourced study that was designed to answer the following three questions: (i) do avatar gesture animations improve the UX of the interactive health monitoring platform relative to a static graphic? (ii) are text-to-speech (TTS) voice prompts rated as positively as human voice prompts, if not better? (iii) is the UX equally good for participants across different severity levels of depression symptoms? To the best of our knowledge, this work is the first exploration of the perception of avatar animations and TTS voices in conversational agents for health monitoring by speakers with and without depression.

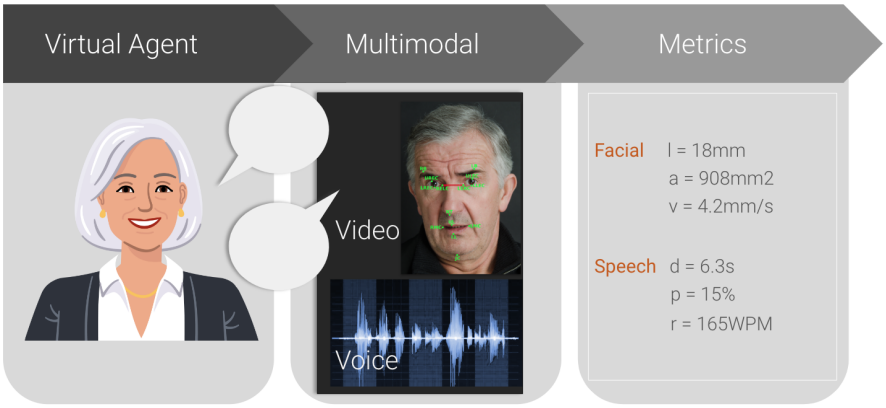


Fig. 1: Multimodal dialogue platform driven by a virtual agent, Tina.

2 Methods & Study Design

The study protocol was granted exempt status by Advarra[1], an independent external Institutional Review Board. Crowdsourced participants based in the United States of America were recruited through Prolific[2], an online platform for research participant recruitment. After filling out standard demographic questions — regarding race, ethnicity, year of birth, sex at birth [3], education level, first language, current state of residence, etc. — participants were guided by a virtual agent, Tina (see Figure 1), to complete a protocol comprising various speaking exercises that elicit speech and facial behaviours [23]. For example, Tina would ask the user to read a passage and wait for the participant to finish reading before moving on to the next task where Tina would ask the participant to describe what they see in a picture. At the end of the interactive session, participants filled out the Personal Health Questionnaire (PHQ-8) depression scale [24].

To answer the questions asked in the previous section, we designed this study to have four separate arms (number of participants in parentheses):

1. **TTS = No, Animation = No**: Tina's prompts voiced by a human voice actress and no avatar animation (n = 132)
2. **TTS = No, Animation = Yes**: Tina's prompts voiced by a human voice actress and avatar animation present (n = 110)
3. **TTS = Yes, Animation = No**: TTS voice (Elizabeth by Microsoft Azure) prompts for Tina and no avatar animation (n = 114)

[1] https://www.advarra.com

[2] https://www.prolific.co

[3] Sex at birth was collected instead of gender to account for sex-specific differences in objective metrics.

4. **TTS = Yes, Animation = Yes**: TTS voice (Elizabeth by Microsoft Azure) prompts for Tina with avatar animation (n = 102)

No participant took part in more than one arm of the study. Animation included subtle expressions like eye blinks by the avatar while listening to the participant and a smile at the end of every participant turn. It is important to note that participants did not know whether the prompts they heard were recorded by a human or were synthesised. All participants were also instructed to complete a UX survey, inspired by previous work reported in the literature [25, 26], consisting of nine Likert-scale questions, each with a possible score ranging from 0 (most negative sentiment) to 4 (most positive sentiment). A sum score of 36 thus represents total satisfaction with respect to user experience. The nine questions were as follows:

1. **UX performance**: How would you rate the system's overall performance?
2. **UX engagement**: How engaged did you feel during the interaction?
3. **UX intelligibility**: How intelligible was Tina?
4. **UX delay**: How satisfactory was the delay in Tina's response to you?
5. **UX interruption**: How often did Tina interrupt you?
6. **UX relatability**: How relatable is Tina's voice?
7. **UX understanding**: How well do you think Tina understood you?
8. **UX regularity**: How regularly would you use an app that involves interacting with Tina to monitor your health?
9. **UX experience**: How would you rate your overall experience interacting with Tina?

3 Analysis

3.1 Categorical Groups

To analyse how participants at different levels of depression symptom severity rate the system, we grouped all participants (with complete PHQ-8 survey scores) across the four arms of the study into two groups based on their depression level [24]; those with PHQ-8 scores $<=$ 9 were included in the 'None to Mild' group (n = 338) whereas those with PHQ-8 scores from 10 to 24 were grouped in the 'Moderate to Severe' group (n = 111).

To analyse how participants of different age groups rate the system, we grouped all participants (whose year of birth was available) across the four arms of the study into two groups based on their age (mean age was 39.5 years); those between 18 and 40 years in the 'Younger' group (n = 251) whereas those above 40 years of age in the 'Older' group (n = 178).

3.2 Statistics

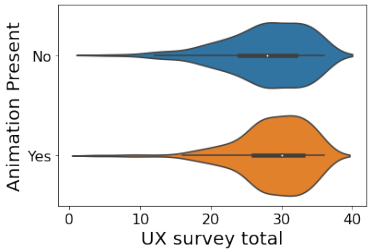
We performed an n-way ANOVA with the UX survey total as the dependent variable (with a maximum possible score of 36 and minimum possible score of 0) and the following categorical independent variables: (i) presence or absence of animation, (ii) presence or absence of TTS, (iii) Depression Group, (iv) Age Group and (v) sex at birth. To account for non-normal distributions of scores, we further ran non-parametric Kruskal-Wallis tests for the main effects to look at differences in the total score and the scores of the individual questions of the UX survey. To reduce the risk of Type I errors, we applied the Benjamini-Hochberg procedure [27] to every statistical test and only those differences that survived this correction were reported in the paper.

4 Results

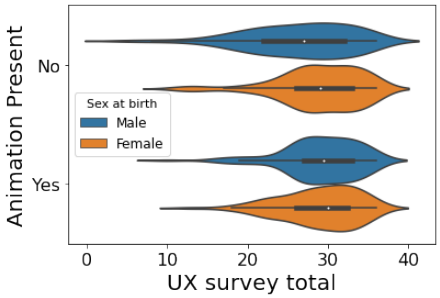
The n-way ANOVA revealed that there is a main effect of animation (see Figure 2a) on the UX survey total (F = 6.82; p = 0.0094). A Kruskal-Wallis test, however, was not significant at $\alpha = 0.05$ (H = 2.42, p = 0.1200) indicating that although the mean UX survey totals were statistically different when animation was present or absent, the median totals were not. The UX survey total was greater when animation was present (mean = 28.69, std = 5.28; median = 30, first quartile/Q1 = 26, third quartile/Q3 = 33) than when animation was absent (mean = 27.70, std = 6.04; median = 28, Q1 = 24, Q3 = 32). Kruskal-Wallis tests revealed that these differences were driven primarily by engagement (H = 4.38; p = 0.0364) and intelligibility (H = 4.88, p = 0.0272).

The interaction term for the presence/absence of animation and sex at birth was significant (see Figure 2b, F = 4.19; p = 0.0413). When animations were present, there wasn't much of a difference in the system rating by males (mean = 29.02, std = 4.82; median = 29.5, Q1 = 27, Q3 = 33) and females (mean = 28.79, std = 4.80; median = 30, Q1 = 26, Q3 = 32.5). But when animation was absent, males rated the system lower (mean = 26.34, std = 6.57; median = 27, Q1 = 22, Q3 = 32) than females (mean = 28.74, std = 5.42; median = 29, Q1 = 26, Q3 = 33).

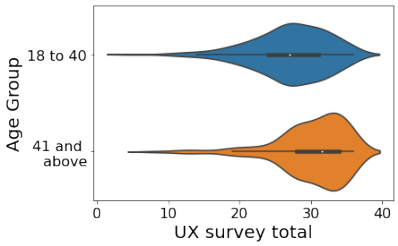
There is also a main effect of age group (see Figure 2c) on the UX survey total (F= 28.05; p = 1.96e-7). A Kruskal-Wallis test also confirmed significant differences (H = 42.97; p = 5.55e-11). Participants from the older age group (41 and above) rated the system more favourably (mean = 30.05, std = 5.18; median = 31.5, Q1 = 28, Q3 = 34) than those in the younger age group, i.e. 18 to 40 years of age (mean = 27.04, std = 5.44; median = 27, Q1 = 24, Q3 = 31). These differences appeared in eight of the nine questions of the UX survey with the older age group consistently rating every question higher than the younger age group: UX performance (H = 15.20; p = 9.7e-5), UX engagement (H = 19.86; p = 8e-6), UX intelligibility (H = 15.48; p = 8.3e-5), UX delay (H = 10.80; p = 0.0010), UX relatability (H = 45.30; p =



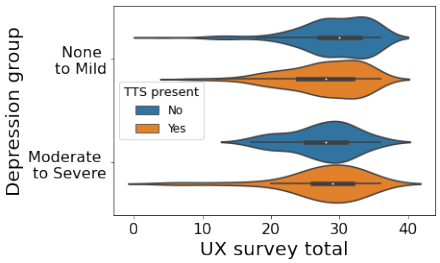
(a) Violin plots for UX survey total when animation is absent or present.



(b) Violin plots for UX survey total by sex at birth when animation is absent or present.



(c) Violin plots for UX survey total according to age group.



(d) Violin plots for UX survey total by depression group when TTS is absent or present.

Fig. 2: Violin plots for UX survey totals (x-axes) as a function of the main effects and interactions that were statistically significant in the n-way ANOVA.

1.69e-11), UX understanding (H = 12.88; p = 0.0003), UX regularity (H = 21.02; p = 5e-6) and UX experience (H = 17.65; p = 2.7e-5).

The interaction term for the presence/absence of TTS and Depression Group was significant (see Figure 2d, F = 5.73; p = 0.0171). When TTS prompts were present, the group with 'Moderate to Severe' depression rated the system slightly higher (mean = 27.57, std = 6.42; median = 29, Q1 = 26, Q3 = 32) than the 'None to Mild' group (mean = 27.60, std = 5.77; median = 28, Q1 = 24, Q3 = 32). When TTS prompts were absent and the prompts were human-voiced, the 'Moderate to Severe' depression group had lower UX survey totals (mean = 27.43, std = 4.88; median = 28, Q1 = 25, Q3 = 31) than the 'None to Mild' group (mean = 29.07, std = 5.66; median = 30, Q1 = 27, Q3 = 33).

5 Discussion and Limitations

The first question that this study aimed to answer was whether avatar gesture animations improve the UX of the multimodal health monitoring platform driven by a conversational agent. Results indicate that the presence of animations is indeed received favourably by users at large when mean UX survey totals are considered (the differences are not statistically significant when median scores are considered). In particular, engagement and intelligibility were rated higher when animation was present. These findings regarding engagement are in line with prior research in the field [18, 19]. It is not entirely clear why an avatar with non-verbal animations would be rated more intelligible than a static avatar. However, one could argue that this could potentially be a demonstration of a multisensory illusion arising from the the simultaneous perception of auditory and visual information, like the McGurk effect [28]. Alternatively, participants may have perceived the animations as an indication of correct task performance, which in turn was attributed to greater intelligibility of instructions. Another possible explanation is that the animation of the virtual agent helped establish a rapport with the users [29], thus boosting the usability scores. Interestingly, participants who reported their sex at birth to be male displayed a strong preference for animation. Prior work has shown that male users tend to be more socially attracted to and want to interact with a female virtual avatar [30]. Hence, it is not surprising that one would observe sex-based differences in UX ratings after interacting with a female-presenting virtual agent. Future studies with diverse avatars could help examine the role that sex and gender play in the usability of virtual agents.

Participants in the older age group rated the system and Tina more favourably than those in the younger age group. Although this observation may seem counter-intuitive given the perception that older users are less likely to adopt technologies (and which is likely a myth [31]), it is in line with previous findings that older participants accept virtual agents more than younger participants [32]. This finding encouragingly supports the deployment of health monitoring agents in older populations.

The second research question in this study was whether TTS prompts are received as favourably as human-voiced prompts. There was no main effect of the presence of TTS in the n-way ANOVA, indicating that the presence of TTS prompts does not hamper the UX of the system and that TTS prompts and human prompts are perceived equally favourably. Moreover, participants with moderate to severe depression liked interacting with the system better when TTS prompts were present. This observation could have multiple explanations: (i) participants with no or mild depression symptoms pay more attention to the quality of the voice prompts and can perceive the peculiarly artificial characteristics of synthesised voice *or* (ii) participants with moderate to severe depression do not pay attention to the quality of the voice prompts *or* (iii) participants with moderate to severe depression who may exhibit impaired prosody during interactions [33] perhaps prefer aligning with the synthetic prosody of the TTS prompts [34]. TTS prompts may have increased the non-human likeness of Tina allowing the participants to interact easily with the

agent [35]. It will be interesting to extend this study to participants with autism spectrum disorder who have shown a preference to interactions with technology typically falling in the 'uncanny valley' over those with humans [36].

Finally, we observed no main effect of Depression Group in the n-way ANOVA, thus answering the third research question — is the UX equally good for participants across different severity levels of depression symptoms — affirmatively. The fact that evaluations by individuals with moderate to severe depression were comparable to those who reported none to mild depression underscores the point that such a digital health monitoring agent can be used efficiently across clinical populations. One caveat to this is that there were fewer participants with moderate to severe depression than those with minimal to mild depression. Confirmatory studies with balanced designs need to be pursued in the future.

6 Conclusions

In conclusion, our findings support the use of avatar animation which can be leveraged to improve patient engagement and in ensuring regular usage of digital health monitoring platforms in collecting meaningful and objective data points to track disease progression. Our findings also indicate that users are generally open to embracing synthetic voices in conversational dialogue-based systems for digital health monitoring, thus allowing for greater scalability, flexibility and perhaps a global outreach.

References

1. R. Garg and S. Sengupta, "He is just like me: A study of the long-term use of smart speakers by parents and children," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–24, 2020.
2. L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, A. Y. Lau *et al.*, "Conversational agents in healthcare: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, 2018.
3. D. Dojchinovski, A. Ilievski, and M. Gusev, "Interactive home healthcare system with integrated voice assistant," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2019, pp. 284–288.
4. T. Dingler, D. Kwasnicka, J. Wei, E. Gong, and B. Oldenburg, "The use and promise of conversational agents in digital health," *Yearbook of Medical Informatics*, vol. 30, no. 01, pp. 191–199, 2021.
5. V. Ramanarayanan, A. C. Lammert, H. P. Rowe, T. F. Quatieri, and J. R. Green, "Speech as a biomarker: Opportunities, interpretability, and challenges," *Perspectives of the ASHA Special Interest Groups*, pp. 1–8, 2022.
6. W. L. Woo, B. Gao, R. R. O. Al-Nima, and W.-K. Ling, "Development of conversational artificial intelligence for pandemic healthcare query support," *International Journal of Automation, Artificial Intelligence and Machine Learning*, vol. 1, no. 1, pp. 54–79, 2020.
7. F. Motolese, A. Magliozzi, F. Puttini, M. Rossi, F. Capone, K. Karlinski, A. Stark-Inbar, Z. Yekutieli, V. Di Lazzaro, and M. Marano, "Parkinson's disease remote patient monitoring during the covid-19 lockdown," *Frontiers in neurology*, vol. 11, 2020.
8. D. Suendermann-Oeft, A. Robinson, A. Cornish, D. Habberstad, D. Pautler, D. Schnelle-Walka, F. Haller, J. Liscombe, M. Neumann, M. Merrill *et al.*, "Nemsi: A multimodal dialog system for screening of neurological or mental conditions," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 245–247.
9. V. Ramanarayanan, O. Roesler, M. Neumann, D. Pautler, D. Habberstad, A. Cornish, H. Kothare, V. Murali, J. Liscombe, D. Schnelle-Walka *et al.*, "Toward remote patient monitoring of speech, video, cognitive and respiratory biomarkers using multimodal dialog technology." in *INTERSPEECH*, 2020, pp. 492–493.
10. A. Abdulrahman, D. Richards, and A. Aysin Bilgin, "A comparison of human and machine-generated voice," in *25th ACM Symposium on Virtual Reality Software and Technology*, 2019, pp. 1–2.
11. M. Cohn, E. Raveh, K. Predeck, I. Gessinger, B. Möbius, and G. Zellou, "Differences in gradient emotion perception: Human vs. alexa voices," in *Proceedings of Interspeech*, 2020.
12. J. Parson, D. Braga, M. Tjalve, and J. Oh, "Evaluating voice quality and speech synthesis using crowdsourcing," in *International Conference on Text, Speech and Dialogue*. Springer, 2013, pp. 233–240.
13. S. D. Craig and N. L. Schroeder, "Text-to-speech software and learning: Investigating the relevancy of the voice effect," *Journal of Educational Computing Research*, vol. 57, no. 6, pp. 1534–1548, 2019.
14. E. Rodero and I. Lucas, "Synthetic versus human voices in audiobooks: The human emotional intimacy effect," *New Media & Society*, p. 14614448211024142, 2021.
15. G. Zellou and M. Cohn, "Top-down effects of apparent humanness on vocal alignment toward human and device interlocutors." in *CogSci*, 2020.
16. M. Cohn, M. Sarian, K. Predeck, and G. Zellou, "Individual variation in language attitudes toward voice-ai: The role of listeners' autistic-like traits," in *Proceedings of Interspeech*, 2020.
17. B. Vaughan, C. D. Pasquale, L. Wilson, C. Cullen, and B. Lawlor, "Investigating prosodic accommodation in clinical interviews with depressed patients," in *International Symposium on Pervasive Computing Paradigms for Mental Health*. Springer, 2018, pp. 150–159.
18. T. Bickmore and D. Mauer, "Modalities for building relationships with handheld computer agents," in *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, 2006, pp. 544–549.

19. I. Wang and J. Ruiz, "Examining the use of nonverbal communication in virtual agents," *International Journal of Human–Computer Interaction*, vol. 37, no. 17, pp. 1648–1673, 2021.
20. J. Thakkar, P. S. Rao, K. Shubham, V. Jain, and D. B. Jayagopi, "Understanding interviewees' perceptions and behaviour towards verbally and non-verbally expressive virtual interviewing agents," in *GENEA: Generation and Evaluation of Non-verbal Behaviour for Embodied Agents Challenge*, 2022.
21. D. Parmar, S. Ólafsson, D. Utami, P. Murali, and T. Bickmore, "Navigating the combinatorics of virtual agent design space to maximize persuasion," in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 2020, pp. 1010–1018.
22. S.-H. Kang, J. Gratch, N. Wang, and J. H. Watt, "Does the contingency of agents' nonverbal feedback affect users' social anxiety?" in *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*, 2008, pp. 120–127.
23. M. Neumann, O. Roesler, D. Suendermann-Oeft, and V. Ramanarayanan, "On the utility of audiovisual dialog technologies and signal analytics for real-time remote monitoring of depression biomarkers," in *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, 2020, pp. 47–52.
24. K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The phq-8 as a measure of current depression in the general population," *Journal of affective disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.
25. M. Heerink, B. Kröse, V. Evers, and B. Wielinga, "Assessing acceptance of assistive social agent technology by older adults: the almere model," *International journal of social robotics*, vol. 2, no. 4, pp. 361–375, 2010.
26. V. Ramanarayanan, P. Lange, K. Evanini, H. Molloy, E. Tsuprun, Y. Qian, and D. Suendermann-Oeft, "Using vision and speech features for automated prediction of performance metrics in multimodal dialogs," *ETS Research Report Series*, vol. 2017, no. 1, pp. 1–11, 2017.
27. Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.
28. H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.
29. L. Huang, L.-P. Morency, and J. Gratch, "Virtual rapport 2.0," in *Intelligent Virtual Agents*, H. H. Vilhjálmsson, S. Kopp, S. Marsella, and K. R. Thórisson, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 68–79.
30. S.-H. Kang, A. W. Feng, A. Leuski, D. Casas, and A. Shapiro, "The effect of an animated virtual character on mobile chat interactions," in *Proceedings of the 3rd International Conference on Human-Agent Interaction*, 2015, pp. 105–112.
31. J. Durick, T. Robertson, M. Brereton, F. Vetere, and B. Nansen, "Dispelling ageing myths in technology design," in *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, 2013, pp. 467–476.
32. P. Philip, L. Dupuy, M. Auriacombe, F. Serre, E. de Sevin, A. Sauteraud, and J.-A. Micoulaud-Franchi, "Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.
33. M. Garcia-Toro, J. A. Talavera, J. Saiz-Ruiz, and A. Gonzalez, "Prosody impairment in depression measured through acoustic analysis," *The Journal of nervous and mental disease*, vol. 188, no. 12, pp. 824–829, 2000.
34. M. Cohn, K. Predeck, M. Sarian, and G. Zellou, "Prosodic alignment toward emotionally expressive speech: Comparing human and alexa model talkers," *Speech Communication*, vol. 135, pp. 66–75, 2021.
35. G. M. Lucas, J. Gratch, A. King, and L.-P. Morency, "It's only a computer: Virtual humans increase willingness to disclose," *Comput. Hum. Behav.*, vol. 37, pp. 94–100, 2014.
36. Y. Ueyama, "A bayesian model of the uncanny valley effect for explaining the effects of therapeutic robots in autism spectrum disorder," *PLoS ONE*, vol. 10, 2015.