Right for the Job or Opposites Attract? Exploring Cross-Generational User Experiences with "Younger" and "Older" Voice Assistants

Yuto Sawa
Tokyo Institute of Technology, sawa.y.ac@m.titech.ac.jp
Julia Keckeis
Tu Berlin, j.keckeis.official@gmail.com
Katie Seaborn*
Tokyo Institute of Technology, seaborn.k.aa@m.titech.ac.jp

ABSTRACT

Efforts to diversify the design of virtual assistants (VAs) are inspiring new work on voice as a factor of user experience (UX). One trajectory is voice agedness. The Computers are Social Actors (CASA) paradigm suggests that humanlike computer agents are perceived in line with models of people. Notably, work on ageism and similarity attraction theory would predict that agents with "younger" and "older" voices may be treated differently. As a first effort, we explored the UX of a novel "older adult" text-to-speech (TTS) system. Younger (n=16) and older (n=18) people evaluated a VA that switched voices ("young" and "old") in a storytelling activity. While most preferred the VA's younger-sounding voices on most measures, all tended to rate the "elder" VA as more appropriate within the storytelling context, challenging expectations based on similarity-attractiveness theory. This work provides initial clues about the relationship between age characteristics and age stereotypes for voice UX.

CCS CONCEPTS

• **Human-centered computing**; • **Human computer interaction (HCI)**;

KEYWORDS

Voice assistants, Virtual assistants, Voice interface, Older adults, Ageism, Computer voice, Voice UX, User experience

ACM Reference Format:
Yuto Sawa, Tokyo Institute of Technology, sawa.y.ac@m.titech.ac.jp, Julia Keckeis, Tu Berlin, j.keckeis.official@gmail.com, Katie Seaborn, and Tokyo Institute of Technology, seaborn.k.aa@m.titech.ac.jp. 2023. Right for the Job or Opposites Attract? Exploring Cross-Generational User Experiences with "Younger" and "Older" Voice Assistants. In *Designing Interactive Systems Conference (DIS Companion '23), July 10–14, 2023, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3563703.3596642

*Corresponding author.

DIS Companion '23, July 10–14, 2023, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9898-5/23/07.
https://doi.org/10.1145/3563703.3596642

1 INTRODUCTION

Studying the user experience (UX) of voice-based virtual assistants (VAs) has become important with the update of this technology globally [17]. VAs are used in a variety of applications, including information search, entertainment, online shopping, smart home controls, e.g., lighting [13], and more recently complex, conversational interactions based on the development of natural language processing (NLP) technologies. Factors that influence voice UX include the VA's accent, appearance, and social characteristics, such as genderedness and agedness, as well as user characteristics and mental models of the world [18]. For example, studies have shown that older users may have different preferences and needs when interacting with a voice assistant compared to younger or middle-aged users [8]. Nevertheless, very little research has involved older adults [17] and even fewer have explored VAs voiced by "older adult" text-to-speech (TTS) systems [8, 17]. This is important because stereotypes about specific *human* social groups can also affect evaluations of voice-based agents [14]. With "aged" voices, ageism may play a role. Ageism refers to limited and/or negative age-based attitudes and discriminatory behaviour, especially towards older age groups; it is considered one of the most discriminatory attitudes, after racism and sexism [15]. Diversifying the voices of VAs should include diversifying the agedness of these voices, but we should also consider whether and to what degree attitudes like ageism influence UX.

To illuminate and clarify the effects of the agedness of VA voice, we conducted a comparative user study, involving a cross-generational sample of younger and older adult participants interactive with a suite of four "younger" and "older" voiced VAs. We asked: *RQ1: Does voice UX differ by voice agedness and based on participant age?* And *RQ2: Do attitudes toward age, i.e., ageism, affect the voice UX based on voice agedness?* As an in-progress work, we contribute an initial understanding of factors related to VA and user age that can influence voice UX. This is important not only for the design and development of VAs, but also for understanding human attitudes towards age.

2 THEORETICAL BACKGROUND

2.1 Vocal Computers Are Social Actors

According to the Computer Are Social Actors (CASA) paradigm, we may unreflexively apply our social responses to people to computer agents we are interacting with that have humanlike cues [14].

We unconsciously read and react to cues related to social factors embedded in agent appearance and behaviour, such as appearance, voice, and gender, and classify the agent as a social entity [14]. People can recognize different voice "ages" based on features of computer-generated speech manipulated through speech rate and pitch [11]. In this early work, we evaluate a novel TTS created from scratch with the voices of older adults, rather than modulating of existing voice stimuli.

2.2 Similarity-Attraction Theory and Voice UX

According to similarity-attraction theory (SAT) [4], we evaluate others as more attractive when they ha attributes similar to our own. This can affect UX-relevant factors such as liking [7], familiarity [1], and satisfaction [5]. In a preprint, Yücel and Rızvanoğlu[1] changed the pitch and speed of a VA's speech to create "younger" and "older" versions. They reported that older adult participants preferred the VA with high emotional empathy. However, the "older adult" VA was not significantly better than the younger-sounding VA at increasing feelings of trust and perceived usefulness. In this research, we decided to investigate the same measures of trust, likability, familiarity, satisfaction, and humanlikeness of two types of VAs: one that sounds older and one that sounds younger. We also drew on similarity-attraction theory but extended this work to include younger and older adult participants. Given this, we hypothesized that:

H1-1: The older-sounding VA will be more positively evaluated by older adults than younger adults.

H1-2: The younger-sounding VA will be more positively evaluated by younger adults than older adults.

2.3 Voice and Ageism

Ageism is a complex prejudice that includes both positive and negative stereotypes about older adults [15]. Palmore [15] defined eight positive stereotypes—Kindness, Wisdom, Reliability, Wealth, Political Power, Freedom, Eternal Youth, and Happiness—and nine negative stereotypes—Illness, Impotence, Ugliness, Mental Decline, Uselessness, Isolation, Poverty, and Depression. Ageism is also expressed in both benevolent and hostile treatment [6]. Measuring ageism attitudes from multiple perspectives is useful for studying how stereotypes affect ageist behavior [10]. Horhota et al. [10] found that level of acceptance when viewing benevolent and hostile ageism behavior with older adults changed depending on the participant's age and the relationship between the participant and the target. Based on this, we hypothesized:

H2: Ageism attitudes will be negatively correlated with positive impressions of a VA that sounds "old."

3 METHODS

In a comparative user study, each participant evaluated two voices of a VA storyteller in a within-subjects design ("younger" and "older" VAs). The between-subjects factor was participant age group (younger and older adult).

3.1 Participants

Thirty-four people participated. Sixteen were young people (9 men, 7 women, and none of another gender), all university students and working adults in their 20s, recruited by the first author. Eighteen were older adults aged 60+ (9 men, 9 women, and none of another gender) recruited from Ota-ku silver centre. Participants were compensated at 1,200 yen/hr. This study was approved by the Ethical Review Committee at Tokyo Institute of Technology.

3.2 System Design

We trained an "older adult" TTS with the phonetic features of an older woman's recorded voice. A sample is provided in the Supplementary Materials. For the "younger" voice, we used the ja-JP-Wavenet-B feminine TTS by Google[2]. Two manipulation checks were performed in-lab, confirming the expected gender and age attributions for each voice. A web-based Wizard of Oz app was created using voice clips generated by the TTS models for the activity (3.3). The VAs were identical and carried out the activity in the same exact way except for the agedness of the voice, i.e., the TTS used.

3.3 Activity

We created a storytelling activity to inspire dreams before bed. To select the "dream" stories, we conducted an online survey on people's last dream, most memorable dream, and dreams they would like to have. Forty-nine people (21 men, 28 women, none of another gender; 25 aged 18-44 and 24 aged 65+) responded. A thematic analysis [3] by the first author of the qualitative items revealed the high-level themes of "characters in the dream," "dream locations," and "own actions in the dream," with sub-themes for each theme, respectively: "I alone," "pets," "family," and "people I admire"; "sightseeing spots," "sky," and "daily life"; and "eating," "conversation," "going out," and "romantic interactions." The ranking of multiple-choice items was "flying (17)," "being in nature (16)," "traveling (15)," "food (15)," "meeting famous people (11)," "money (7)," and "interacting with animals (6)." Based on this, the first author selected six stories, ∼5 mins. each, from the Aozora Bunko (Blue Sky Library)[3]. Participants conversed with the VA to narrow the stories down in two ways: presence of family or an animal. The VA then read the story aloud. The activity was refined through lab pilot testing.

3.4 Procedure and Context

Each participant had two sessions with the VA in our lab, sitting at a table and using a PC displaying the VA system over Zoom. One researcher hosted while another controlled the VA from another PC. The voice of the VA was switched from "young" to "old" in a counterbalanced fashion across participants. Each session involved one story (3.3). Then the VA asked the participant to fill out the questionnaire. To switch voices between sessions, the VA explained that they were trying on a new story voice. The study took between 40-60 mins. A 10-15-min. break was allowed between sessions.

[1] https://doi.org/10.21203/rs.3.rs-1905540/v1

[2] https://cloud.google.com/text-to-speech/docs/voices

[3] https://www.aozora.gr.jp

Table 1: Descriptive statistics for "young" and "old" voices by participant age group (younger and older adults)

Measure	Young Adults		Older Adults	
	"Young" Voice	"Old" Voice	"Young" Voice	"Old" Voice
Satisfaction	5.4	4.8	4.7	3.7
Trust	5.6	5.0	5.1	4.9
Familiarity	4.2	4.8	4.6	4.0
Likability	3.6	3.4	3.6	3.1
Anthropomorphism	2.6	2.9	2.7	2.4

3.5 Instruments and Data Collection

After each session, participants filled out a questionnaire and were interviewed about their impressions of the VA.

3.5.1 Satisfaction, Trust, and Familiarity. Satisfaction, trust, and familiarity were measured using single-item 7-point Likert scales from work on a home VA [12].

3.5.2 Likability and Anthropomorphism. Likability (α = .982) and anthropomorphism (α = .937) were measured using the Godspeed questionnaire [2]. Each was comprised of five items measured on 5-point Likert scales. Similar to other VA research, e.g., [19], we modified the items to make them meaningful for VAs. A lab pilot survey with seven people confirmed the changes.

3.5.3 Ageism Attitudes. Ageism was measured with the AAS-JP [16], a Japanese translation of the Ambivalent Ageism Scale (AAS) [6]. 7-point Likert scales were used to capture responses to 13-item benevolent (α = .703) and 4-item hostile subscales (α = .369). Given the low Cronbach's alpha, i.e., internal consistency, for the hostile subscale, we relied on the benevolent one.

3.5.4 Brief Interview. Participants were interviewed at the end of the study. They were asked: "What is your general impression of the VA?"

3.6 Data Analysis

Descriptive statistics and normality checks via the Shapiro-Wilk test were conducted. Reflexive thematic analysis [3] was used by the first author for the qualitative data. Pearson correlations and t-tests were used to compare all measures according to the hypotheses, i.e., based on voice age group and participant age group. However, in cases of nonnormal distributions, nonparametric alternatives, including Kendall's rank correlations and Mann-Whitney U tests, were used.

4 RESULTS

Satisfaction with the "younger" voice (5.1 ± 1.4) was statistically significantly higher than the "older" voice (4.24 ± 1.71), p = .017. No other statistically significant differences were found. Thematic analysis revealed a range of impressions from the quality of the voices, characteristics of the VA, user attitudes, and activity UX. Comparisons of the quantified themes revealed statistically significant differences by voice age for speed, loudness, pronunciation, smoothness, lack of affect, coldness, compatibility with the activity, familiarity, and ease of listening (Supplementary Table 1). These results favoured the "young" voice, except lack of affect ("young" n=7, "old" n=1), coldness ("young" n=5, "old" n=0), feeling relaxed ("young" n=1, "old" n=9), and compatibility with the activity ("young" n=0, "old" n=7). Younger and older participants did not differ in their impressions of the "young" and "old" voices in terms of statistical significance (Table 1). Altogether, given the lack of statistically significant differences at the intersection of voice age and participant age, we must reject H1-1 and H1-2. Likewise, no statistically significant correlations were found for ageism by voice age, so we reject H2. In short, it appears that voice age alone and/or individual factors beyond ageism attitudes were influential.

5 DISCUSSION

Against expectations, people's impressions of and UX with the VA in the two "young" and "old" age "modes" were not linked to their own age or attitudes towards ageism. Instead, people tended to favour the "young" voice as a more familiar and higher quality TTS, while at the same time favouring the "old" voice as a warm "elder" storyteller most appropriate for the activity and somehow able to express relaxation and greater affect.

5.1 Age and Agedness in Voice and by Generation

Younger and older participants were more satisfied with the technical quality of the young voice. Yet, in contrast to the breaking results of Yücel and Rızvanoğlu[4], our older adult participants did not have a strong preference for younger-sounding VAs. Rather, participants of all ages found the "elder" VA to be more appropriate for the task of conducting the storytelling activity, challenging the expectations of similarity-attraction theory. One reason could be an effect of complementary-attraction, where we prefer others with personality cues complementary to our own [9]. Future work will have to explore how various UX factors relate to personality cues. Another possibility is a link between the activity and the age of the voice. Specifically, we may expect elders in our lives, such as a grandparent, reading a story to us before bed. This may also link to positive yet still stereotyped attitudes of age roles. Future work may include other measures of age(ist) attitudes and interview questions on the connection between task and agent.

[4]https://doi.org/10.21203/rs.3.rs-1905540/v1

5.2 Ain't Nothing But a Number? Ageism Attitudes and Older-Sounding Voices

We could not link ageism to voice UX with the older-sounding VA. Notably, we had to rely on the benevolent subscale. The AAS-JP was validated for internal consistency with older adults [16] but the original AAS involved young and middle-aged people [6]. Our results may be a matter of our sample size *or* generational differences in ageism attitudes.

5.3 Limitations and Future Work

Technical quality of the TTSs may have influenced results; we will improve the "older adult" TTS by training the models with longer voice recordings. We will also increase the sample size when conducting the full longitudinal study.

6 CONCLUSION

Voice UX can be influenced by the agedness of the VA and the user's age group. When developing and deploying VAs , we should consider the perceived age of the voice, the voice characteristics that can influence it, and the attributes of the user. Future work will need to replicate and extend these results for other scenarios and socio-linguistic contexts.

ACKNOWLEDGMENTS

This work was funded by a Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Early Career Scientists (KAKENHI WAKATE) grant (no. 21K18005) and the Tokyo Tech Young Science and Engineering Researchers Program (YSEP). We also thank the Fonobono Research Institute and all members of the Aspire Lab for supporting this work.

REFERENCES

[1] Aune, R.K. and Kikuchi, T. 1993. Effects of Language Intensity Similarity on Perceptions of Credibility Relational Attributions, and Persuasion. *Journal of Language and Social Psychology*. 12, 3 (Sep. 1993), 224–238. DOI:https://doi.org/10.1177/0261927X93123004.

[2] Bartneck, C., Kulić, D., Croft, E. and Zoghbi, S. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*. 1, 1 (Jan. 2009), 71–81. DOI:https://doi.org/10.1007/s12369-008-0001-3.

[3] Braun, V. and Clarke, V. 2012. Thematic analysis. APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological. American Psychological Association. 57–71.

[4] Byrne, D. 1969. Attitudes and Attraction. *Advances in Experimental Social Psychology*. L. Berkowitz, ed. Academic Press. 35–89.

[5] Carli, L.L., Ganley, R. and Pierce-Otay, A. 1991. Similarity and Satisfaction in Roommate Relationships. *Personality and Social Psychology Bulletin*. 17, 4 (Aug. 1991), 419–426. DOI:https://doi.org/10.1177/0146167291174010.

[6] Cary, L.A., Chasteen, A.L. and Remedios, J. 2017. The Ambivalent Ageism Scale: Developing and Validating a Scale to Measure Benevolent and Hostile Ageism. *The Gerontologist*. 57, 2 (Apr. 2017), e27–e36. DOI:https://doi.org/10.1093/geront/gnw118.

[7] Dustin, D.S. and Alfonsin, B. 1971. Similarity and liking. *Psychonomic Science*. 22, 2 (Feb. 1971), 119–119. DOI:https://doi.org/10.3758/BF03332524.

[8] Esposito, A., Amorese, T., Cuciniello, M., Pica, I., Riviello, M.T., Troncone, A., Cordasco, G. and Esposito, A.M. 2019. Elders prefer female robots with a high degree of human likeness. *2019 IEEE 23rd International Symposium on Consumer Technologies* (Ancona, Italy, Jun. 2019), 243–246.

[9] Gurtman, M.B. 2009. Exploring Personality with the Interpersonal Circumplex. *Social and Personality Psychology Compass*. 3, 4 (2009), 601–619. DOI:https://doi.org/10.1111/j.1751-9004.2009.00172.x.

[10] Horhota, M., Chasteen, A.L. and Crumley-Branyon, J.J. 2019. Is Ageism Acceptable When it Comes From a Familiar Partner? *The Journals of Gerontology: Series B*. 74, 4 (Apr. 2019), 595–599. DOI:https://doi.org/10.1093/geronb/gby066.

[11] Huff, E.W., Stigall, B., Brinkley, J., Pak, R. and Caine, K. 2020. Can computer-generated speech have an age? *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 2020), 1–7.

[12] Lee, S., Cho, M. and Lee, S. 2020. What if conversational agents became invisible? Comparing users' mental models according to physical entity of AI speaker. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3 (Sep. 2020). DOI:https://doi.org/10.1145/3411840.

[13] Moriuchi, E. 2019. Okay, Google!: An empirical study on voice assistants on consumer engagement and loyalty. *Psychology & Marketing*. 36, 5 (2019), 489–501. DOI:https://doi.org/10.1002/mar.21192.

[14] Nass, C., Steuer, J. and Tauber, E.R. 1994. Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 24 1994), 72–78.

[15] Palmore, E. 1999. *Ageism: Negative and Positive, 2nd Edition*. Springer Publishing Company.

[16] Sawa, Y. and Seaborn, K. 2022. Localizing the Ambivalent Ageism Scale for Japan. *The 8th Asian Conference on Aging & Gerontology 2022: Official Conference Proceedings* (Tokyo, Japan, Mar. 2022), 33–36.

[17] Seaborn, K., Miyake, N.P., Pennefather, P. and Otake-Matsuura, M. 2022. Voice in human-agent interaction: A survey. *ACM Computing Surveys (CSUR)*. 54, 4 (2022), Article No. 81. DOI:https://doi.org/10.1145/3386867.

[18] Seaborn, K. and Urakami, J. 2021. Measuring voice UX quantitatively: A rapid review. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan, 2021), No. 416.

[19] Seymour, W. and Van Kleek, M. 2021. Exploring Interactions Between Trust, Anthropomorphism, and Relationship Development in Voice Assistants. *Proceedings of the ACM on Human-Computer Interaction*. 5, CSCW2 (Oct. 2021), 371:1-371:16. DOI:https://doi.org/10.1145/3479515.