SPEAKER ACCOMMODATIONS TOWARDS VUI VOICES ON THE DIMENSIONS OF VOICE ONSET TIME AND PITCH RANGE

Gracellia Purnomo[*1], Chloë Farr[2], Charissa Purnomo [3], Nicole Ebbutt[3], Amanda Cardoso[3] and Bryan Gick[3,4]
[1]School of Audiology and Speech Sciences, University of British Columbia, Canada
[2]Department of Linguistics, University of Victoria, Canada
[3]Department of Linguistics, University of British Columbia, Canada
[4]Haskins Laboratory, New Haven, United States of America

1 Introduction

There is a growing presence and integration of voice-user interfaces (VUIs) in the form of virtual assistants such as Siri, Alexa, and Google Home. VUIs are inanimate objects, however they use animate (human) voices to interact with their client.

Accommodation occurs when an interlocutor adjusts their speech in relation to another interlocutor [1], either by converging (becoming more similar) to, or by diverging (becoming more different) from, the other speaker. Speakers may accommodate on any level of the hierarchy of linguistic features, including syntactic features, lexical choices, or phonetic features of their speech [2], the last of which is the focus of the current investigation. On the phonetic level, voice onset time (VOT) [3] and pitch range [4, 5] have been identified as common features in which speakers accommodate to an interlocutor.

The present paper considers whether or not interlocutors may employ the same types of speaker accommodation towards these inanimate objects. In addition, since the human-likeness, or perceived animacy of VUIs can be different amongst operating systems, the additional question arises of whether the perceived human-likeness may further increase the likelihood of the device being treated as such. The present study examines whether speakers accommodate voice onset time (VOT) and pitch to VUI voices and the extent to which the human-likeness of the voice influences accommodation.

2 Methods

2.1 "VUI" voices

Four Amazon Polly [6] synthetic voices were rated by 26 linguists for perceived human-likeness and the voices rated most and least human-like were used in the experiment. Polly's standard system was used as the robotic voice (hereafter "R"), and Polly's neural system was used as the human-like voice (hereafter "H") as a consequent of these ratings. As VOT was similar for both voices, in order to be able to see the extent of accommodation, the VOT was manipulated in Praat [7] so that the VOT of the voiceless plosive consonants of R were twice the length of the Amazon Polly output, and half the length for H. These voices were used to mimic a VUI system in that the responses would be played directly after a participant read out pre-determined prompts.

[*] gracellia.purnomo@gmail.com

2.2 Experiment

The study took place virtually via UBC-secured Zoom. Participants were asked to read two practice prompts (pre-test) presented on their screen, from which they heard no response. This was followed by thirteen prompts for which they heard a response from the VUI voice (post-test). For example, the participant read the prompt "Where can I buy pots and pans?" which the VUI responded by saying "You can buy pots and pans from Canadian Tire." This procedure was repeated with the same prompts for "R" and "H". Participants then completed a survey regarding their professional and personal experiences with VUIs, and what they believed the experiment to be about. No participants had professional experience with VUI and all participants believed that they were interacting with an authentic VUI system.

2.3 Participants

Forty-two English-speaking participants were recruited through UBC's Linguistics in the Classroom (LOC) system. Participants with poor audio or speakers who did not report English as their dominant language were omitted, leaving 25 participants. Participants were assigned to one of two counterbalanced presentations of the voices (voice order).

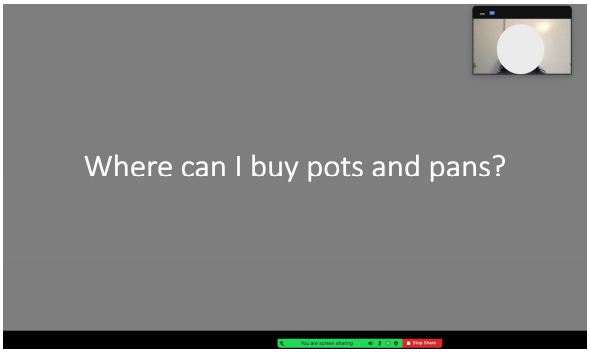

Figure 1: Screenshot of participants' view.

2.4 Measurements

Textgrids for participants' audio recordings were generated using Montreal Forced Aligner [7]. Each voiceless plosive VOT length was manually marked with high interrater reliability and then extracted using a Praat script. Pitch trajectories of acoustic syllables for each prompt sentence were extracted automatically using Prosogram [8]. We report the results for mean and median F0 in semitones and TrajPhonZ, which is essentially a Z-scored measure of how variable the pitch is.

3 Results

Linear mixed effects models were applied to the VOT and pitch results with H or R voice and voice order group as fixed effects with an interaction, and participant and prompt as random effects.

It should be noted that speakers demonstrate a wider range of pitch variation in the pre-test speech compared with the post-test (H, R) speech. None were found to be significant. In other words, there is no difference in either VOT values (Figure 2) or pitch (mean, median, variation) (Figure 3) between pre-test and post-test responses by the speakers nor between the speakers' responses to the two voice types (H, R).

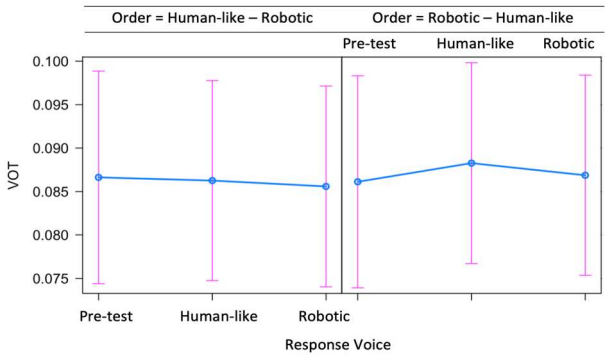

Figure 2: VOT results by voice order group and response voice.

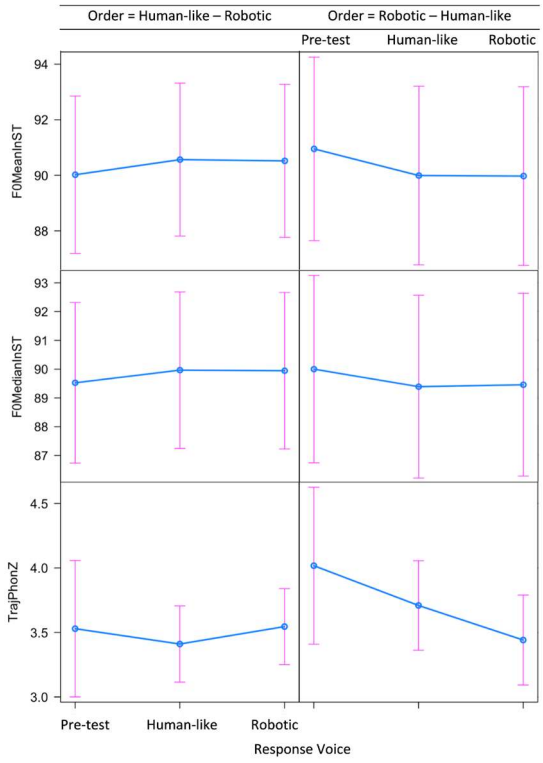

Figure 3: Pitch results by voice order group and response voice.

Furthermore, visual inspection of the individual speaker results suggest that while at a group level there were no significant differences, at an individual level, there are some speakers who are accommodating to the VUI voices.

4 Discussion

The current findings do not point to a consensus on whether and how interlocutors accommodate towards VUIs. Although some participants did show some tendency for accommodation, there was not consistency in whether speakers diverged or converged. Further investigation of who these speakers are and why they accommodate is in progress. Accommodation is in part motivated by an interlocutor's awareness of social standing. Such motivation may be less likely to exist when interacting with a VUI, and may play a part in the lack of accommodation.

Acknowledgments

The authors would like to thank Kristen Eredics for her contributions towards analysis. This work was supported by the National Institutes of Health grant number DC-002717 to Haskins Laboratories and SSHRC Insight Grant 435-2019-0426 to Bryan Gick.

References

[1] Bell A. Language Style as Audience Design. J Language in Society. 1984 Jun [cited 2021 Aug 11];13(2):145-204. Available from: https://doi.org/10.1017/S004740450001037X

[2] Lindblom B. Explaining phonetic variation: A sketch of the H&H theory. In: Hardcastle W, Marchal A, editors. Speech Production and Speech Modeling. Springer Netherlands; c1990. p. 403–439.

[3] Asano Y, Gubian M. "Excuse meeee!!": (Mis)coordination of lexical and paralinguistic prosody in L2 hyperarticulation. J Speech Communication [Internet]. 2018 May [cited 2021 Aug 11];99:183–200. Available from: https://doi.org/10.1016/j.specom.2017.12.011

[4] Babel M, Bulatov D. The role of fundamental frequency in phonetic accommodation. J Language and Speech. 2011 Sep [cited 2021 Aug 11];55(2):231-248. Available from: https://doi.org/10.1177/0023830911417695

[5] Oviatt S, Maceachern M, Levow GA. Predicting hyperarticulate speech during human-computer error resolution. J Speech Communication. 1998 May [cited 2021 Aug 11]; 24:87-110. Available from: https://doi.org/10.1016/S0167-6393(98)00005-3

[6] Amazon Web Services. Amazon Polly. 2019. Retrieved from https://aws.amazon.com/polly/

[7] McAuliffe M, Socolof M, Mihu, S, Wagner M, Sonderegger M. Montreal Forced Aligner: trainable text-speech alignment using Kaldi. Proceedings of the 18th Conference of the International Speech Communication Association. 2017.

[8] Mertens P. Polytonia: a system for the automatic transcription of tonal aspects in speech corpora. J Speech Sciences. 2014 Jan;4 (2):17-57.