XVII AISV Conference

Speaker Accommodations and Voice User Interfaces: Does Human-likeness of a Voice Matter?

Chloë Farr, Gracellia Purnomo, Amanda Cardoso, Arian Shamei, Bryan Gick

Overview

- Voice user interfaces (VUIs):
  - Used with smartphones & smart home devices to execute automated tasks [1]
  - VUIs consist of:
    - Automatic speech recognition (ASR) - process speech
    - Artificial Intelligence (AI) - execute instructions
    - Text to [synthetic] speech - respond to speaker
- Research with earlier rudimentary VUI's found
  - To be better understood, speakers hyperarticulate; this accommodation includes converging or diverging pitch (f0), voice onset time (VOT), pause & vowel insertions [2]
- Recent research focuses on user experience:
  - Less on phonetic adjustments
  - Has targeted the advancement ASR
- Hyperarticulation in human-animate speech:
  - Relate speech to human interlocutor's [3, 4]:
    - Converge with in-group (speakers of same region or status, etc.) or diverge from out-group
  - Human-VUI speech may be distinct from human-human, more like human-animal [5]

Question

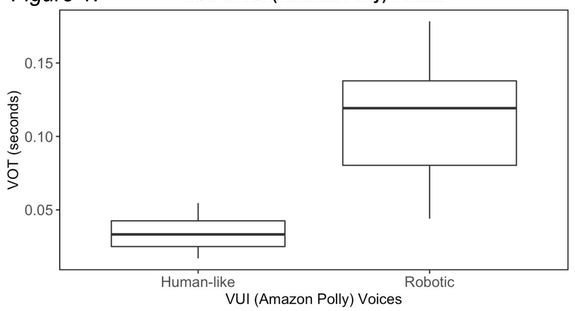
Do speakers change their speech to accommodate to a VUI?

- If considered in-group, speaker will converge to VUI, minimizing distance and treat VUI as human
- If not considered in-group, speaker will diverge from VUI, increasing distance and treat VUI as non-human
- If considered inanimate, speaker will not accommodation to the VUI

Methods

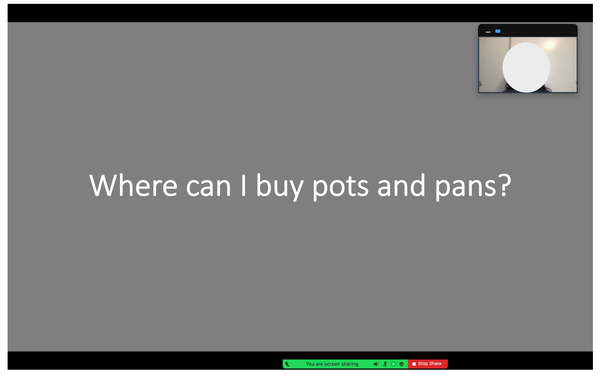
- To determine exposure effects, compared VOT pre- & post-exposure to robotic and human-like VUIs
- Materials (Fig 1):
  - 2 voices generated for VUI responses
    - In Praat, VOT of /p/ and /t/ were edited:
      - VOT shortened in Human-like VUI voice
      - VOT lengthened in Robotic VUI voice

Figure 1. **VOT of VUI (Amazon Polly) Voices**



- Participants:
  - 18 students from the University of British Columbia participated via Zoom
  - Participants counterbalanced for voice order
- Procedure (Fig 2):
  - Participants told they were speaking to virtual assistants
  - Provided sentences to read via share screen
  - Participants heard a beep (signalling speech recognition), read sentence aloud, heard the VUI response and waited for next sentence
  - Practice: 3 sentences prior to exposure to voices
  - Read 12 sentences to each VUI, received spoken response from VUI
  - Repeated with second VUI

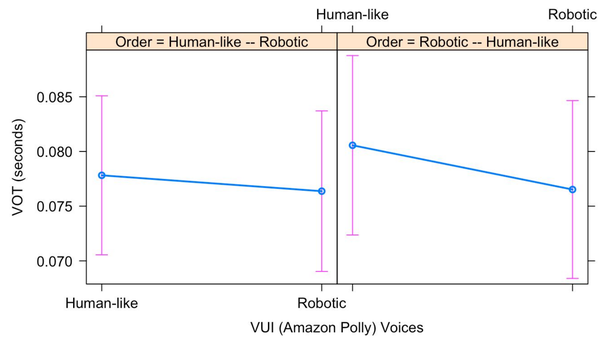
Figure 2. Screenshot of participants' view



Analysis

- Participants who didn't speak English as their dominant language, and/or had poor audio quality were excluded from the analysis
- Extracted audio was manually segmented at the sentence level in Praat, after forced alignment (Darla), manually annotated VOT in Praat, time extracted with script
- Statistical analysis: R - Linear mixed effect models across participants and T-tests within participants

Figure 3. **VOT of Speakers by Exposure to VUI Voices**



Results

- Insignificant differences in VOT between voices:
  - Within and across participants
  - Following each of the voices (Fig 3)
  - Between counterbalanced VUI order (Fig 3)

Conclusion

Results show that speakers do not accommodate to synthetic VUI voices regardless of how human-like or robotic they sound. This indicates that VUIs are treated as inanimate interlocutors and are unlikely to change the way people produce speech.

References

1. Rubio-Drosdov, E., Díaz-Sánchez, D., Almenárez, F., Arias-Cabarcos, P., & Marín, A. (2017). Seamless human-device interaction in the internet of things. *IEEE Transactions on Consumer Electronics*, *63*(4), 490–498.
2. Oviatt, S., Maceachern, M., & Levow, G.-A. (1998). Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication, 24*, 87-110
3. Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. *Speech Production and Speech Modeling, 55*, 403–439.
4. Pardo, J.. (2006). On phonetic convergence during conversation. The Journal of the Acoustical Society of America. 119. 2382-93. 10.1121/1.2178720.
5. Moore, R. K. (2020). Spoken language technology now seems to work - so what's left to be done. *UCL Speech Science Forum.*

Acknowledgement

We would like to thank Charissa Purnomo and Nicole Ebbutt for their contributions to data collection.

THE UNIVERSITY OF BRITISH COLUMBIA

Partners

Interdisciplinary Speech Research Laboratory

Haskins Laboratories

Natural Sciences and Engineering Research Council of Canada

Conseil de recherches en sciences naturelles et en génie du Canada

Canada

Speaker Accommodations and Voice User Interfaces (VUI): Does Human-likeness of a Voice Matter?

Background: Speakers often accommodate their speech to interlocutors.

Question: How do we adjust our speech to a **human-like** but **inanimate** speaker?

Methods: Participants spoke to 2 virtual assistants with either robotic and human-like voices with VOT manipulated.

Results: No change in participants' VOT to either voice.

Implications: We may treat VUIs as inanimate despite their human-like voices.