Speaker Accommodations and VUI Voices: Does Human-likeness of a Voice Matter?

Voice user interfaces (VUI) are becoming increasingly embedded in peoples' lives, as they are built into every voice assistant (e.g. Amazon's Alexa, Apple's Siri and Microsoft's Cortana) on smartphones, computers, and smart home products (Amazon Echo, Google Home), and are able to perform a large number of automated tasks. Therefore, people may be speaking to VUI with increasing ease and frequency [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]. Such increases may have an effect on the way that people perceive and interact with the devices, so that interactions with VUI become more like those with human interlocutors (human-to-animate), such as adjusting the properties of one's speech to be better understood (e.g. hyperarticulation) or to be more or less like the interlocutor (i.e. accommodation). For example, previous work has demonstrated that speakers may attempt to produce clearer or more understandable speech by changing phonetic properties of their speech, such as changing their pitch (f0) or making their voice louder [Oviatt, S., Maceachern, M., & Levow, G.-A. (1998). Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication, 24*, 87-110]. These types of phonetic adjustments have also been found as a communication repair strategy used in human-to-animate interactions and have often been described under the label of hyperarticulation [Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. *Speech Production and Speech Modeling*, *55*, 403–439]. This might suggest that human interaction with VUI are, at least, in some ways like human-to-animate interaction. However, much of the current work looking at the phonetic changes that are observed in human-to-VUI interactions was conducted at a time when VUI were newer, not as advanced in terms of uses or the way they sound, and less integrated into devices and humans' lives. The limited more recent work has generally been impressionistic and focused on how understanding phonetic changes can enhance automatic speech recognition systems (ASR) rather than what the implications are for speakers and changes in their speech behavior. Understanding more about the phonetic characteristics of human speech in human-to-VUI interaction may shed light on people's expectations of the cognitive capabilities and animacy of the devices and also the changing nature of human-to-VUI interactions, as they become more intertwined in our daily lives. Our wider project is interested in investigating whether people treat VUI as animate (more human-like) or inanimate technology (more like a tool). The current paper focuses on one part of this question by looking at whether the human-likeness of VUI's synthetic voice (text to speech - TTS) influences the extent that speaker accommodation occurs for the human interlocutor, and what this can tell us about how speakers categorise VUI and, in turn, VUI interaction as either different or the same as other spoken interactions.

A large body of research demonstrates that speakers adjust the properties of their speech to be more or less like their interlocutor depending on whether the speaker perceives the interlocutor to be part of their same in-group (e.g. from the same region, community, ethnic background, socio-economic status, etc.) or not (for work on accommodation, see Giles 1973, Pardo 2006, Giles & Ogay 2007, Babel 2012). Speakers converge (sound more like) to interlocutors that they consider to be in their group or who they would like to lessen social distance with, and they diverge (sound less like) from interlocutors that they consider to be outside of their group or with whom they would like to increase social distance from themselves. We, therefore, hypothesise that if the speaker considers a VUI voice to be in-group (i.e. group them as like a human interlocutor) then they will converge to the VUI's speech characteristics and if they do not they will either diverge from the VUI or demonstrate maintenance (i.e. no accommodation). Whether they diverge or show maintenance will also provide us with an understanding of the way that human participants perceive these voices. Divergence would specifically suggest an "othering" and wanting to create distance, demonstrating superiority/power or categorising the VUI voice as specifically non-human. On the other hand, maintenance does not necessarily suggest that the participant is distancing themselves, but rather that this is not comparable to human-to-animate interactions. As the exposure to VUI and interactions have increased most substantially over the last decade, it might also be the case that

speakers would interact with any VUI or no VUI in a similar way to human-to-animate interactions. In the current study we investigate these hypotheses by varying how human-like the VUI's synthetic voice sounds.

We will present an expected 50 participants with two female sounding synthetic voices generated with Amazon Polly's text-to-speech service (Amazon Web Services, 2020), which participants are told are VUI. 26 linguistic researchers rated the two voices as the most human-like and most robotic, while still being perceivable as synthetic voices. One voice uses Polly's Neural system (a sequence of phonemes converted to a sequence of spectrograms, subsequently converted into continuous audio signal) or Polly's Standard system (concatenated pre-recorded phonemes). In order to be able to explore how speakers accommodate to the different voices, we manipulated word-initial voiceless plosive voice onset time (VOT) duration in Praat (Boersma & Weenink, 2020), so that it's twice as long for the robotic voice and half as long as for the more human-like voice compared with the original samples. We also normalised the voices for speech rate, as it was found that there were substantial differences in speaking rate across the robotic and human-like voices.

The experiment is counterbalanced for presentation of voices to participants, so that half of the participants are presented with the more robotic voice first and the other half are presented with the more human-like voice first, and participants are randomly assigned to hear one of these voice orders. Participants are asked to use ten pre-scripted prompts with the VUI, the order of which is randomised within a block. The VUI responds to each of the prompts appropriately. For example, the participant speaks the prompt, "Send a text to Paul", to which the synthetic voice returns, "What would you like to write to Paul?".

Participants' speech are analysed acoustically within each block in order to determine whether VOT duration of word-initial voiceless plosives for the participants are adjusted over the course of the interaction with each of the VUI voices. In other words, is there evidence of accommodation by the participants to the speech characteristics of the VUI's voices and the magnitude of the change that occurs? Preliminary results suggest that half of the participants accommodate to the human-like voice in comparison to the pre-exposed and robotic data, resulting in statistically significant shorter VOT ($p<0.05$). In all participants, VOT accommodation did not occur in relation to the robotic voice. A survey administered after the study reported that the participants believed they were speaking to real virtual assistants.

As previously mentioned, this preliminary study feeds into a wider project that investigates the phonetic properties of speech during human-to-VUI interaction, adjustments to speech that occur as a result of differences in the backgrounds of the human interlocutor (e.g. familiarity with VUI, language variety) and differences in the properties of the VUI (e.g. accuracy of responses, language variety), and how these changes relate to those seen in human-to-human interactions. Ultimately, this project will provide us with insights into under what conditions, if any, humans interact with these devices in the same way they do with other humans, which may give us a better understanding of how these devices are perceived. Finally, understanding if TTS systems mold human speech. If people are accommodating to the VUI, and the VUI is learning from their speech, will ASR natural speech recognition improve, or will there be a fundamental change in human speech production with technological or animate interlocutors?