EFFECTS OF DISFLUENCY ON PRAGMATIC PROCESSING IN HUMAN-ROBOT DIALOGUES

Xinyi Chen (The Hong Kong Polytechnic University) & Yao Yao (The Hong Kong Polytechnic University)
xysimba.chen@connect.polyu.hk

The occurrence of disfluencies such as “um” and “uh” in conversational speech can be interpreted as hinting unfavorable responses to requests or suggestions made by the conversational partner. For example, Schegloff (2010) noted that a “dispreferred” response to a previous turn is often preceded with “uh(m)”. In this study, we examine whether such disfluency effects also exist in human-robot dialogues. Specifically, we ask whether human listeners will process “uh(m)” markers in robot speech as a cue for disperferred responses. While there is no reason to think that robots might, like humans, hesitate or feel embarrassed for delivering an unfavorable response, previous studies suggest that humans may behave similarly in human-human and human-robot interactions (Chen et al., 2021, 2022; Cohn et al., 2019, 2020, 2021).

To test this hypothesis, we conducted two experiments, both using a cued recall task with the false memory paradigm (Brewer, 1977). In the first experiment, participants would hear multiple blocks of short, two-turn dialogues between two speakers, one male and one female. In the critical dialogues, the male speaker would make a proposal or a request for information, and the female speaker would produce a neutral response (neither favorable nor unfavorable in the verbal content), which may or may not be preceded with “uh(m)”. At the end of each block, the participant would be asked to recall the female speaker’s responses in the current block by choosing the more accurate statement for each dialogue. The second experiment adopted identical materials and procedure as the first experiment, except that participants would watch video clips of the dialogues, which would allow them to hear both speakers and to see that the female speaker is a humanoid robot, Furhat (https://furhatrobotics.com/). We predict that in both experiments, participants would remember the female speakers’ responses on critical trials **more negatively** (i.e., as dispreferred responses) when the response is preceded by “uh(m)”, compared to a completely fluent response.

A total of 123 native Mandarin Chinese speakers participated in this study (72F, 51M, aged 18-34 years old), roughly evenly split between the two experiments. The experimental materials consisted of 100 dialogues (50 critical and 50 fillers), evenly distributed in 10 blocks. The critical stimuli underwent a separate norming test to ensure that the verbal responses (not including disfluency) were indeed perceived as neutral. Auditory stimuli were recorded by two native Mandarin Chinese speakers (1F, 1M) in their 20s in a sound-proof booth. To create disfluent versions of the responses, tokens of “uh(m)” were elicited from the female speaker’s natural speech when performing a story telling task and subsequently inserted to the female speaker’s recording of the scripted dialogues at designated, utterance-initial positions. To create the video stimuli, we combined the auditory stimuli with video recordings of Furhat in the “live” mode, showing visible facial features, facial expressions and head movement while lip syncing to the female speaker’s lines in the dialogues. The video stimuli would give viewers the impression of witnessing a conversation between Furhat and a male human speaker who is not shown in the video.

Generalized linear mixed-effects models (GLMM) were built to analyze the participants’ responses in the memory recall task on critical trials. The GLMMs for both experiments revealed a significant effect of disfluency on the likelihood of remembering the response as dispreferred (Exp1: $\beta_{disfluency} = 0.52$; $p < .001$; Exp2: $\beta_{disfluency} = 0.60$; $p < .001$). When modelled together, no interactions of Experiment and Disfluency were found, showing that the disfluency effect was comparable between human-human and human-robot dialogues. These results suggest that human talkers use the same model for interacting with both humans and AIs. We discuss implications of this study in the context of prediction adjustment (or the lack thereof) based on the interlocutor.