Is there an auditory uncanny valley for synthesised speech?

Alice Ross
MSc Developmental Linguistics 2021-22
University of Edinburgh
Supervisors
Dr Catherine Lai
Prof Martin Corley

Alice's GitHub

BACKGROUND

The concept of the uncanny valley comes from robotics. It describes a tendency for people to feel uneasy about androids which look almost, but not quite human: realism is positively correlated with user approval only up to a certain point.

This idea has been explored thoroughly in the context of visual, but not auditory perception. TTS creators are working towards increasingly natural, human-like speech. Is this what users want?

Our online between-subjects experiment investigated listeners' reactions to an array of text-to-speech (TTS) voices with modified prosody. We obtained qualitative survey responses from 205 listeners in six randomly assigned conditions, who heard and evaluated one voice each.

HYPOTHESES

H1 Users demonstrate a general preference for more realistic TTS voices over more mechanical sounding ones.

H2 There is a significant dip in this otherwise positive correlation between voice realism and approval, indicating that some users find very realistic, 'almost human' voices unpleasant.

METHOD

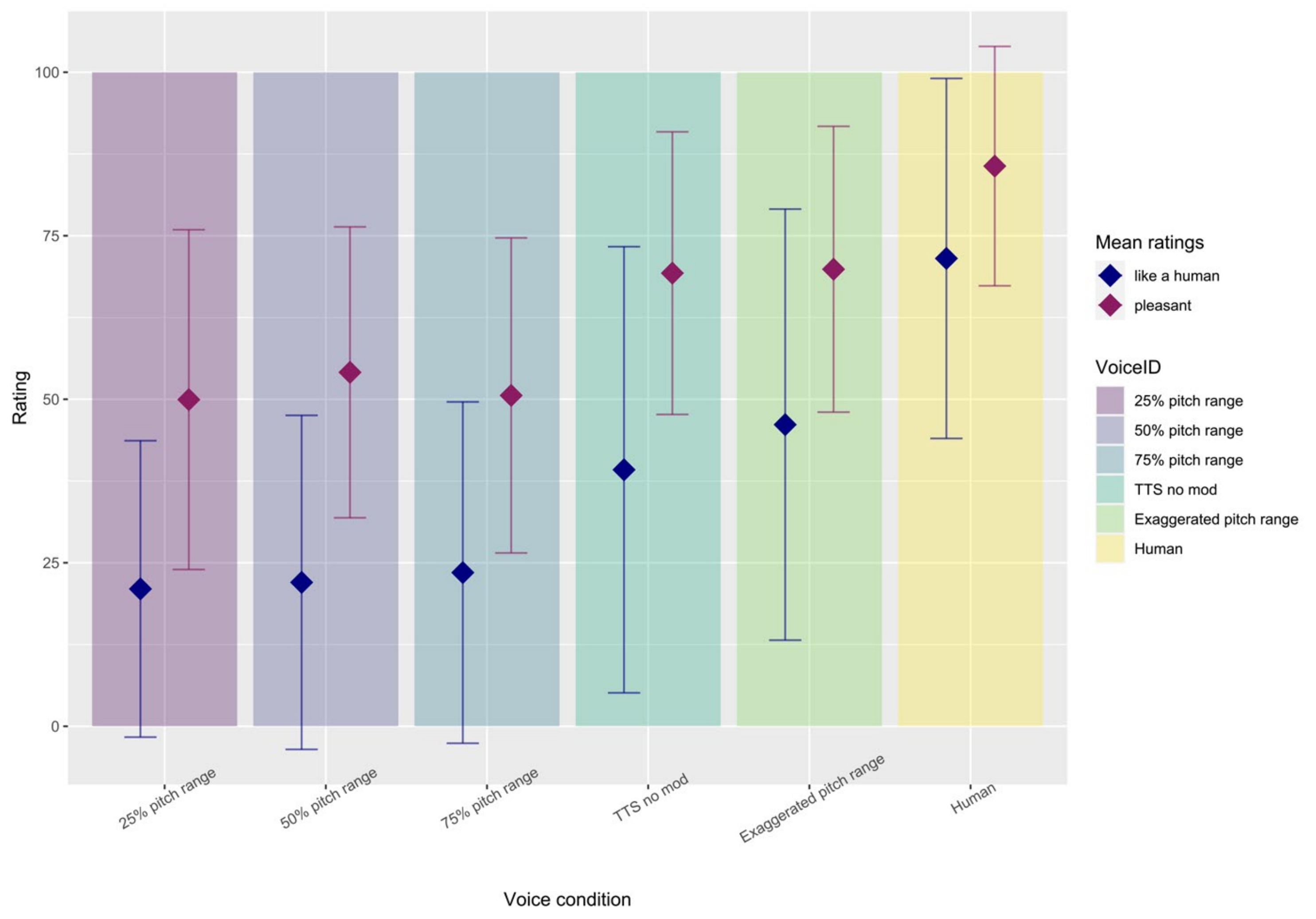
- The same text was recorded in each of six voices, all based on the same speaker:

Human: Linda Johnson's voice
TTS with pitch variation increased to 2.0 x original
Unmodified TTS voice trained on the LJ Speech dataset
TTS with pitch variation slightly decreased to 0.75
TTS with pitch variation decreased to 0.5
TTS with pitch variation heavily decreased to 0.25

- Participants (*n* = 205) were grouped into six conditions
- They listened to three short passages, answering a comprehension question after each
- Finally, they rated the voice they'd heard on sliding scales (0-100) for seven comparisons, including:

unpleasant / pleasant
cold / warm
fake / natural
like a machine / like a human

RESULTS



Consistent with H1: a positive correlation was found between ratings of realism and approval (pleasantness, friendliness, warmth) across all voices.

Our results neither support nor contradict Hypothesis 2. Absence of evidence, not evidence of absence: none of our TTS voices achieved a mean score of ≥ 50 on realism, so H2, which pertains to 'very realistic' voices, can't be fully tested using this array. The question should be revisited with new stimuli including more realistic voices.

However, there is some evidence of a potential non-linear relationship between realism and approval; at the higher end of our realism scale, there is an increase in mean realism score and no corresponding increase in positive attribute ratings.

DIRECTIONS FOR FUTURE RESEARCH

- Revisiting this study with more realistic TTS voices
- Investigating effects of exaggerated pitch variation
- Different languages, audiences, custom-made stimuli...
- Exploring context: TTS for agents with various appearances and purposes (assistive, instructive, social...)
- Real time speech processing comparisons: How do reaction time, recall accuracy, and cognitive load vary?
- Individual differences: Comparing listeners from different generations, with varying L1/L2 proficiency, accessibility needs, familiarity with and attitudes to technology