ONGOING PROJECT

Is there an auditory uncanny valley for synthesised speech?

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?

An online experiment investigates listeners' reactions to realistic text-to-speech (TTS) voices, using an array of TTS samples with modified prosodic characteristics.

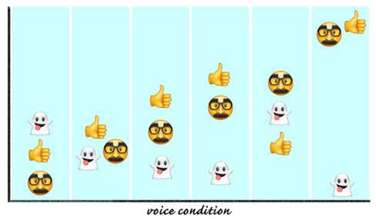
HYPOTHESES

- More realistic TTS voices are generally preferred over flat, mechanical ones.
- There is a significant dip in this correlation between voice realism and positive ratings; some users find very 'human-like' synthesised voices unpleasant.

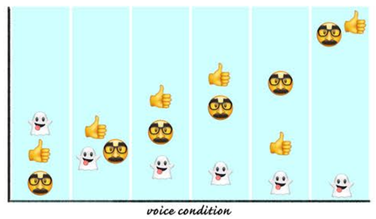
METHOD

- The same text is recorded in each of six 'voices' with pitch variation modifications, e.g:
  - Exaggerated
  - Flattened (monotone)
  - Appropriate (copied from the original human voice)
  - Inappropriate (copied from a different passage)
  - Inconsistent (alternating appropriate and monotone)
  - Control: Unmodified human voice
- Participants ($n \geq 180$) are randomly assigned to a voice condition
- They listen to a short passage and answer comprehension questions
- Finally, they rate the voice used on sliding scales (0-100) for comparisons including:
  - 🤓 mechanical - human
  - 👍 pleasant - unpleasant
  - 👻 reassuring - eerie

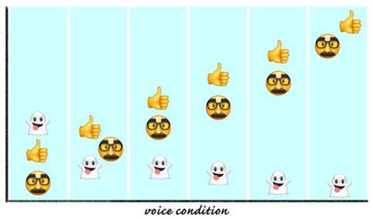
POSSIBLE RESULTS



Results support both hypotheses: evidence compatible with UVE



A dip in pleasantness, no increase in eerie or cold feelings: realistic TTS is disliked but not uncanny



No valley: a straightforward positive correlation between realism and approval

FUTURE RESEARCH QUESTIONS

- Charting the valley: replicating with wider arrays, different languages, systems trained using other voices
- Exploring context: agents using TTS with various appearances and purposes (assistive, instructive, social...)
- Real time speech processing comparisons: how do reaction time and cognitive load vary?
- User groups: TTS listeners are diverse, from different generations, L1/L2 proficiency, accessibility needs, familiarity with and attitudes to technology. Do these factors affect preferences?

Testable norming study

Alice Ross
MSc Developmental Linguistics 2021-22
University of Edinburgh

Supervisors:
Dr Catherine Lai
Prof Martin Corley