Is there an uncanny valley for speech?

Investigating listeners' evaluations of realistic synthesised voices

AUTHORS

Alice Ross
Prof. Martin Corley
Dr. Catherine Lai

alice-ross.github.io

AFFILIATIONS

UKRI Centre for Doctoral Training in Natural Language Processing
University of Edinburgh, School of Informatics and
School of Philosophy, Psychology & Language Sciences
The Centre for Speech Technology Research

INTRODUCTION

The uncanny valley effect (UVE) -- distaste for entities that appear almost, but not quite, human -- is often cited in human-computer and human-robot interaction research, but almost all UVE studies focus on visual perception. Meanwhile, increasingly realistic text-to-speech (TTS) voices are frequently encountered in a variety of settings. Improvements in synthetic speech's naturalness may not increase the perception of trustworthiness [1]. In some contexts and for some listeners, a machine using a very human-like voice can be undesirable [2]. We synthesise a range of TTS voices from one speaker's data, modifying the pitch range. Using a between-subjects, online listening experiment. we aim to find out how modifying the prosodic characteristics of TTS voices affects listeners' perceptions of their realism and pleasantness.

REFERENCES

[1] T. D. Do, R. P. McMahan, and P. J. Wisniewski, *A New Uncanny Valley? The Effects of Speech Fidelity and Human Listener Gender on Social Perceptions of a Virtual-Human Speaker.* Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022.
[2] P. Wagner, J. Beskow, S. Betz, et al. *Speech Synthesis Evaluation—state-of-the-art assessment and suggestion for a novel research program.* Proceedings of the 10th Speech Synthesis Workshop (SSW10). 2019.
[3] A. Łańcucki, *Fastpitch: Parallel text-to-speech with Pitch Prediction*. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.
[4] K. Ito and L. Johnson, *The LJ Speech Dataset*. keithito.com/LJ-Speech-Dataset

HYPOTHESES

H1 Listeners find more realistic TTS voices more pleasant than more mechanical-sounding ones.

H2 There is a significant dip in this otherwise positive correlation between voice realism and pleasantness, indicating that some listeners dislike very realistic, 'almost human' voices.

METHOD

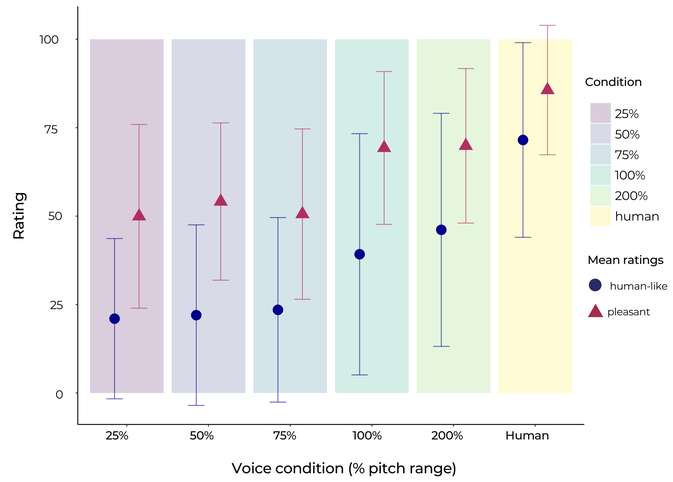
We created five synthesised voices using FastPitch [2], all trained on LJSpeech data [3]. We generated a set of utterances in each of the five voices and included Linda Johnson's unmodified recorded voice as a control, for a total of six experimental conditions:

- TTS, pitch range heavily decreased to 0.25
- TTS, pitch range decreased to 0.5
- TTS, pitch range slightly decreased to 0.75
- Unmodified LJSpeech-trained TTS voice
- TTS, pitch range increased to 2 x original
- Human: Linda Johnson's voice

Participants (n = 205) were randomly assigned to one condition each. They listened to three excerpts from a story and then rated the voice they'd heard from 0-100 on seven antonym pairs, including **unpleasant / pleasant** and **like a machine / like a human**.

RESULTS

H1 Perceived human-likeness and pleasantness were positively correlated. Pearson's product-moment correlation: r(203) = **.48**, p < .001.
H2 could not be fully tested, because all our TTS voices were rated poorly - below 50 on average - for human-likeness.
The voice with increased pitch range was rated most realistic at 46%. Ratings were consistently much higher for pleasantness than for realism, across all conditions; high realism seems not to be necessary for high approval.



Relationship between pitch variation and perceived realism: our TTS voices' mean realism scores fall into two clear groups (decreased vs normal or increased pitch range), with almost no variation between levels of flattening. Listeners' sensitivity to TTS pitch range modification warrants further research.