Testing for an auditory uncanny valley

[redacted]

1. Introduction

The ‘uncanny valley’ hypothesis in the design of artificial agents, first posited in 1970 by Masahiro Mori (Mori et al, 2012), predicts a tendency for humans to perceive highly realistic humanoid robots as unsettling. Replicating human-like appearance and movements in artificial agents can make them more likeable, but only up to a point: near the boundary between mechanical and human, where an android is very lifelike but still detectably artificial, people interacting with it may feel uneasy or frightened.

Researchers have investigated the uncanny valley effect in the visual realm, studying artificial faces (Chattopadhyay and MacDorman, 2016), bodies (Zlotowski et al, 2015), movement (Saygin et al, 2012), and hands (Poliakoff et al, 2013). Such work helps to inform the visual design of android robots and virtual avatars, which have gradually begun to enter various contexts in society and interact with people who are not trained roboticists (e.g. Newton and Newton, 2019).

Meanwhile, many studies in human-computer interaction focus on various aspects of communication between humans and natural language processing systems which use machine-generated voices, with or without embodiment. Linguists and psychologists have researched prosodic (Suzuki & Katagiri, 2007) and lexical alignment (Branigan et al, 2011) in human-computer dialogues; the effectiveness of synthesised voice-over narration, compared to human narration, for teaching videos (Craig and Schroeder, 2019); and the effects of such voices’ perceived gender on listeners’ attitudes (Mullenix et al, 2003; Tolmeijer et al, 2021).

One extremely fast-growing recent application of synthetic speech technology has been the introduction of non-embodied ‘voice assistants’ like Siri and Alexa, which receive natural language input from users and respond using synthesised voices. Introduced in the early 2010s, these assistants are now used in millions of personal devices. The human-like quality of machine generated speech has also rapidly progressed in recent years. A Google project, Duplex (Leviathan & Matias, 2018) produced synthesised speech so realistic that, in brief phone call interactions, listeners believed the assistant was human – and faced an immediate backlash from commenters describing the technology as unethical (O'Leary, 2019) and creepy (Applin, 2018). Machine-generated voices are also used in advanced assistive technology (AATs) for people who cannot speak due to disability or injury; in this context, users have repeatedly requested more natural-sounding voices (Kisner, 2018).

To summarise: artificially generated voices are increasingly widely used, increasingly realistic, and our attitudes and reactions to them are not fully understood. It would be timely to investigate whether uncanny valley-like effects can be detected in an auditory context.

2. Experimental design

2.1 Hypothesis

I propose a test for the presence of an auditory uncanny valley by analysing the listeners' reactions to a range of different synthesised voices.

H1: a significant 'dip' is measurable in an otherwise positive correlation between synthesised voices' similarity to a natural human voice and perceived likeability.

2.2 Procedure

I designed a simple online experiment in which participants listen to short audio clips instructing them on how to proceed to the next screen by selecting an image, pressing a specific key, or recalling a fact that they heard to answer a multiple-choice question. After completing a number of these trials, the participants are asked to rate the voice that they heard on a sliding scale for features such as 'pleasant' and 'helpful', and can make any other comments in a free text field. Responses and response times are written to a .csv file on completion of the experiment.

The complete experiment would involve an array of four or more different synthetic voices, ranging from very mechanical to very realistically human sounding, plus at least one actual human voice. A pre-test (with different participants) could be used to make sure voices were placed correctly - i.e. in an order that most listeners agree upon - on this 'human-likeness' continuum. I would create *n* sets of stimuli by having the same text 'read' by each voice, and each individual participant would be assigned to a condition determining which voice they would hear. The present version is a demonstration of a single condition, using a synthesised voice which would be somewhere around the middle of the lifelikeness range.

Ideally, this experiment would run on a crowdsourcing platform with a large number of participants, > 25 per condition. The responses to the questionnaire section are returned as numerical values between 0 and 100, and are the most important part of the data collected. The final question asks whether participants thought they had heard a human or a machine. The intention here is that each voice will then receive a score out of 100 for 'human-likeness', which should make it easy to compare and plot the other questions' scores against these values.

Fig. 1 is a mocked up example of one type of data visualisation I could generate using the questionnaire responses sorted by condition. The individual observations are pale dots; the bars show the mean and range. This example shows a dramatic valley effect, with the 'River' voice receiving a much lower average 'pleasantness' rating than the others.

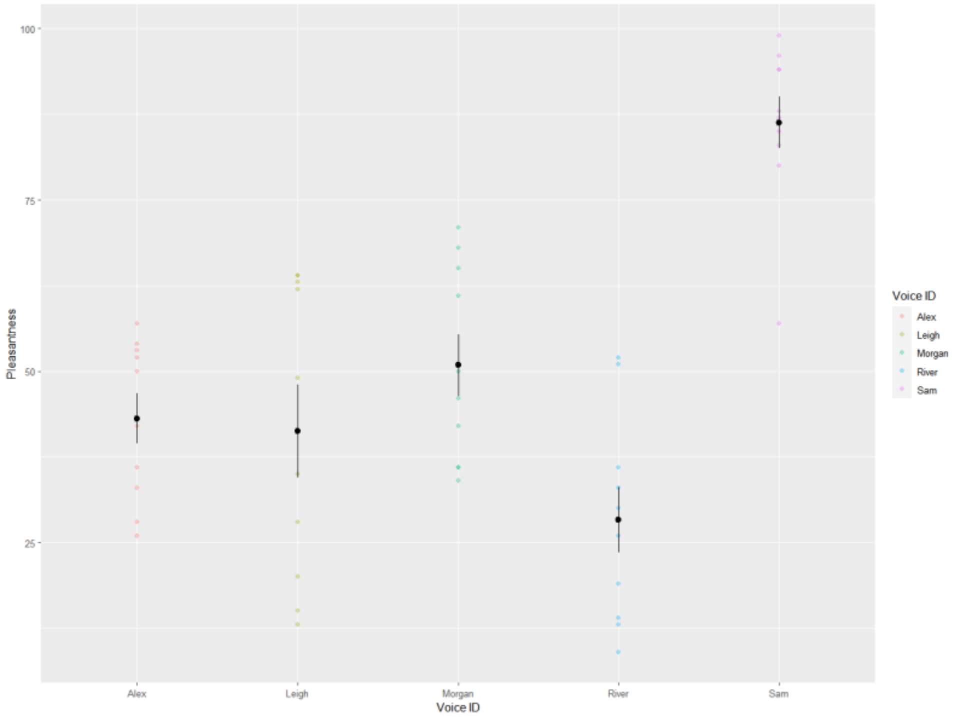

Fig. 1: EXAMPLE - mocked up results, not real data

2.3 Design decisions

Throughout most of the experiment, the prompts or instructions are delivered only via audio, without accompanying text on screen (other than the introduction screen, which has both). This is intended to ensure that participants do listen to the voice - i.e. it's hard to complete the experiment with the volume turned down, off, or while listening to something else - so their feedback about it is more likely to be valid.

The voice used is 'disembodied', with no visual avatar or agent representing the speaker, because the design of any such agent would be likely to influence people's perceptions and ratings of qualities like friendliness.

I decided to build in features that would serve as attention checks, making the experiment very easy to complete if the participant listens to the stimuli but difficult to click through randomly without listening. In this version of the experiment, I have used the *categorize-html* plugin for this reason: it requires a specific keyboard response in order to continue, meaning it is far easier to complete these 'trials' by listening to the audio than not.

By design, the image selection trials would require a correct response for this purpose, and the fact recall trials would not (because the task is less straightforward, the audio is much longer and would be annoying to repeat, and because this allows observation of the voices' effectiveness for teaching-like tasks). I attempted to use a loop node with if/else logic to repeat image trials if an incorrect image was selected, but unfortunately haven't managed to get that working, so currently the experiment continues when the participant clicks on any of the four images. A potentially positive consequence is that I would obtain data on the proportion of correct and incorrect responses in two kinds of trials for the various voices. This is not relevant to testing my hypothesis, but it could prove useful in practical terms if it turns out that a particular voice, or a specific file, is difficult for people to hear or understand. If I can get the loop node working in future, I would like to have a small group test the experiment and remove either the categorize or image trials, whichever takes people longer, as they would both be serving the same function.

Another minor coding challenge was that I wanted (preferably minimally distracting) functionality for participants to be able to replay audio instructions, in case there was some noise in their vicinity when the file played and they didn't hear it. This is especially important for any trials that do require a correct response, as it would be problematic if participants could easily get stuck half way through the experiment and miss out on getting paid for completing it.

I found two solutions for this, and tried both:

The 'categorize' trials use an HTML audio element. This is because there's no jsPsych categorize-audio plugin, but also made it very easy to include on-screen play, pause and volume functionality using the 'controls' attribute. Negative points are that this adds more visual elements on screen, means that people can do unexpected things like downloading the sound file, and I found the 'autoplay' attribute behaved oddly - pressing a key would cause the file to start playing again - so currently it's disabled and participants have to play the sound manually.

In the fact recall trials, a loop node allows participants to click a button to replay the audio once it's finished. This is preferable as it's more consistent visually, though a bit more difficult to code correctly.

The questionnaire responses use a slider, with require_movement parameter set to true to make it less likely that participants would answer randomly, e.g. repeatedly clicking on the middle button to get through the questions quickly. In previous versions I tried the Likert scale and button response plugins, but wasn't satisfied with them as I felt that using statements with 'agree/disagree' responses could be leading, especially for the human/machine question; using numeric values (e.g. "1 = human, 10 = machine") could be confusing; and using buttons labelled with lengthy strings like "neither trustworthy nor untrustworthy" was unwieldy - the buttons with the longest, or shortest, labels might also become more visually salient and more likely to be selected for that reason. The 0-100 range of the slider allows finer detail. It might prove necessary for me to standardise scores at the analysis stage, for instance if some participants only rate voices between 40 and 60 and others use the entire scale; this should be straightforward to resolve using z-scores.

2.4 Stimuli

Audio was recorded from a demo of Microsoft Azure. Like Tolmeijer et al (2021), I used a 'female' voice pitch shifted down (Sonia, pitch: 0.40), so that it was ambiguous rather than readily categorised as male or female. This is intended to minimise effects of societal bias around gender, although it's worth noting that there is a tendency to apply binary categorical perception to voices, so many listeners are likely to think it sounds like an unusual female or male voice rather than gender-neutral.

Images were sourced from historical collections of ephemera believed to be in the public domain.

2.5 Improvements

One limitation of the experiment is that it uses one-sided communication, which isn't especially naturalistic compared to the ways that people use voice assistants in everyday life. Tolmeijer et al (2021) used interactive trials in which people communicated with a voice assistant to book a particular flight; arguably a realistic task like this would allow for greater ecological validity. However, aside from being a lot more challenging to program, it would also introduce more uncontrolled variation and people might base their ratings on the AI system's speech comprehension and performance, not simply the voice, especially if they had difficulty communicating with it.

In terms of accessibility, ideally I would like to make a version of the experiment suitable for people with visual impairments; because it's best generally to be as inclusive as possible, but particularly because this group are more likely to interact frequently with text-to-speech systems. It's also not currently suitable for mobile devices, and I believe it would be if the attention checks requiring keyboard responses can be removed.

The code overall could be more streamlined, and I would like to improve the functions for saving data to cut down on data cleaning requirements later. Finally, to avoid non-naivety of participants, the file names should probably be changed so the url doesn't include 'uncanny'.

The questionnaire section is fairly basic and needs testing to see if it generates useful results, as each participant only hears one voice and their judgments of machine generated voices in general might all be similar or very different. I considered trying a different system, where many participants would each listen to two voices chosen at random from the array and choose which one they prefer; each time a voice was preferred it would be given a point and I could compare the final scores. With a sufficiently large sample size, this might be a more sophisticated method for measuring pleasantness/likeability, but it wouldn't tell us much about other features.

As stated, this experiment is intended as an initial investigation into the question of whether any 'auditory uncanny valley' effect exists; it would not allow for detailed analysis of specific dimensions of any such effect, ways of mitigating it, or fine-grained comparisons of individual differences. It would theoretically be possible to look for correlations between responses and social factors by including more survey questions and grouping participants. Mori's original description of the uncanny valley offers no hypotheses about this, but MacDorman and Entezari (2015) found some evidence of a correlation between individual differences in personality and sensitivity to the uncanny valley. In future, more detailed analyses could include grouping by factors like age, political affiliation, familiarity with or attitudes toward new technology, if there are reasons to believe that any of these are predictors of a valley effect.

3. Impact

Predicting humans' responses to artificial voices could provide valuable design guidance for programmers working on the many applications of such voices, from AATs and educational technology to GPS systems. If it is found that people generally have a strong aversion to very human-like synthetic speech, then it may be better for audio designers to aim for the ‘peak’ before the uncanny valley in voices’ realism. However, more research would be needed to discover whether any such effect is context-specific, and how it relates to wider attitudes and individual differences.

References

Applin, S. A. (2018). Google Duplex Puts AI Into a Social Uncanny Valley. *Vice*, May 2018

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, 121(1), 41-57.

Chattopadhyay, D., & MacDorman, K. F. (2016). Familiar faces rendered strange: Why inconsistent realism drives characters into the uncanny valley. *Journal of vision,* 16(11), 7.

Craig, S. D., & Schroeder, N. L. (2019). Text-to-Speech Software and Learning: Investigating the Relevancy of the Voice Effect. *Journal of Educational Computing Research*, *57*(6), 1534–1548.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1), 1-12.

Kisner, J. (2018). How a new technology is changing the lives of people who cannot speak. *The Guardian*, 23 January 2018.

Leviathan, Y. & Matias, Y. (2018). Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. *Google AI Blog*, May 2018

MacDorman, K. F., & Entezari, S. O. (2015). Individual differences predict sensitivity to the uncanny valley. *Interaction Studies*, 16(2), 141-172.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19.2 (2012): 98-100.

Mullennix, J. W., Stern, S. E., Wilson, S. J., & Dyson, C. L. (2003). Social perception of male and female computer synthesized speech. *Computers in Human Behavior*, 19(4), 407-424.

Newton, D. P., & Newton, L. D. (2019, November). Humanoid robots as teachers and a proposed code of practice. *Frontiers in education* (Vol. 4, p. 125).

O'Leary, D. E. (2019). GOOGLE'S Duplex: Pretending to be human. *Intelligent Systems in Accounting, Finance and Management*, 26(1), 46-53.

Poliakoff, E., Beach, N., Best, R., Howard, T., & Gowen, E. (2013). Can Looking at a Hand Make Your Skin Crawl? Peering into the Uncanny Valley for Hands. *Perception, 42 (9),* p. 998–1000.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social cognitive and affective neuroscience, 7(4),* p. 413–422.

Suzuki, N., & Katagiri, Y. (2007). Prosodic alignment in human–computer interaction, *Connection Science, 19:2*, 131-141

Tolmeijer, S., Zierau, N., Janson, A., Wahdatehagh, J. S., Leimeister, J. M. M., & Bernstein, A. (2021, May). Female by Default?–Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-7).

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H. (2019). Welcome to the Tidyverse. Journal of open source software, 4(43), 1686.

Zlotowski, J. A., Sumioka, H., Nishio, S., Glas, D. F., Bartneck, C., & Ishiguro, H. (2015). Persistence of the uncanny valley: the influence of repeated interactions and a robot's attitude on its perception. *Frontiers in Psychology, 6, 883.*

Source for audio stimuli:

https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/#features

Sources for image stimuli:

https://publicdomainreview.org/collection/japanese-fireworks-catalogues

https://digitalcollections.nypl.org/collections/cigarette-cards

[redacted]

GRADEMARK REPORT

FINAL GRADE

85/100

GENERAL COMMENTS

Instructor

Your lit review gives a sensible rationale for the research question, and the background is very clearly explained. Your concluding section situates the research nicely in terms of its practical applications.

The experiment runs nicely. The instructions are all very well written, the audio stimuli are well made and you've clearly gone beyond the plugins we covered on the course. The experiment looks clean and simple from a participant perspective but I can see that you're doing quite a lot behind the scenes to get it just right e.g. image buttons all the same size, images preloaded, greying out buttons until the audio has finished playing and they become clickable. Great attention to detail!

The report is excellent, demonstrating the extensive thought that has gone into your experiment design. I was particularly impressed that you had considered multiple solutions to the issue you identified around replaying audio and given detailed thoughts on the pros/cons of both. It's also good to see that you've considered accessibility issues and, specifically, how these relate to your research question.

Strengths: Very clear and detailed report; carefully constructed and professional-looking experiment combining a number of different trial types; design decisions well justified.

Weaknesses: No major weaknesses - a very strong piece of work!

PAGE 1

PAGE 2

Comment 1

A very minor point, but it might have been clearer to state this as two separate hypotheses i.e. 1. There will be a generally positive correlation between synthesised voices' similarity to a natural human voice and perceived likeability, 2. There will be a significant 'dip' in this correlation above [a certain value of similarity]. It's not a big deal in the context of this assignment but if you were running this for real I assume you would test these two things separately (firstly is there a relationship between the two variables, and secondly is it linear?) and it's always helpful for the reader if they can see the connection between the way your hypotheses are stated and the way the results are then presented.

PAGE 3

Comment 2

It's great that you've included this and are thinking about how you would analyse your data. Another very minor point but as a general rule, err on the side of more detailed figure captions so the reader can interpret your visualisations without searching in the text for more information (i.e. the last couple of sentences of the para above might be better as a caption).

PAGE 4

Comment 3

For future reference, a 'while' loop might be what you're looking for here :)

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9