Users' attitudes to realistic TTS voices:

Is there an auditory uncanny valley?

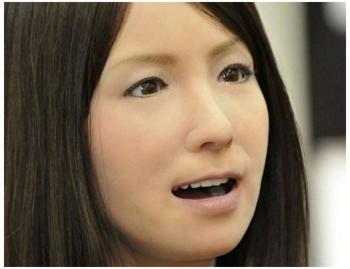
Alice Ross
0236853@ed.ac.uk
MSc Developmental Linguistics 21-22

Supervisors:
Dr Catherine Lai
Prof Martin Corley

Uncanny valley effect

- The uncanny valley effect (UVE) was first proposed by roboticist Masahiro Mori in 1970.

- It predicts a tendency for humans to perceive highly realistic humanoid robots as unsettling; a dip or ‘valley’ in an otherwise positive correlation between realism and pleasantness or likeability.

			
Not creepy	Not creepy	Creepy	Not creepy

- Although the hypothesis isn’t new, UVE is still a current topic with plenty of ongoing research. It’s been studied in artificial faces, bodies, movements, hands…

- People who aren’t roboticists are now increasingly likely to encounter AI agents, ‘virtual humans’, and especially natural language interfaces like Siri or Alexa in everyday contexts.

My research question:
Is there an auditory uncanny valley?

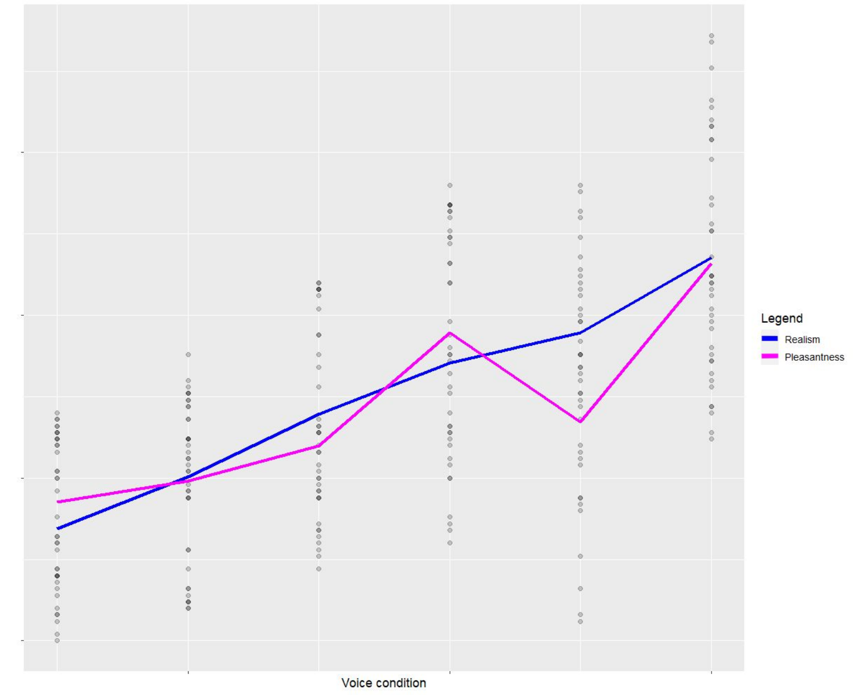
I want to compare users' perceptions of an array of voices, asking about (at least) two dimensions: realism and pleasantness.

- H1a: Users demonstrate a general preference for more realistic TTS voices over older, more 'mechanical' sounding ones.

- H1b: 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.

I plan to have participants listen to a short passage of audio (or possibly watch a video with voiceover narration) then answer questions about their recall of what they heard. This requires them to evaluate the voice in a given context.

My research question:
Is there an auditory uncanny valley?



If there's a significant uncanny valley effect, I would expect results to look something like this.

Please note this is NOT real data!

Problem: Which voices to compare?

What would an array from ‘mechanical’ to ‘almost human’ look (sound) like?

🔊 KlattTalk ‘Perfect Paul’, 1984	?	🔊 CereProc ‘Stuart’, 202?

There are obvious problems with comparing these two:

- People might prefer the Scottish accent (or not!)
- Or some other dimension, like fundamental frequency, rate of speech…
- Familiarity: one of the voices is pretty famous

Existing studies on TTS

Researchers have investigated different dimensions of TTS:

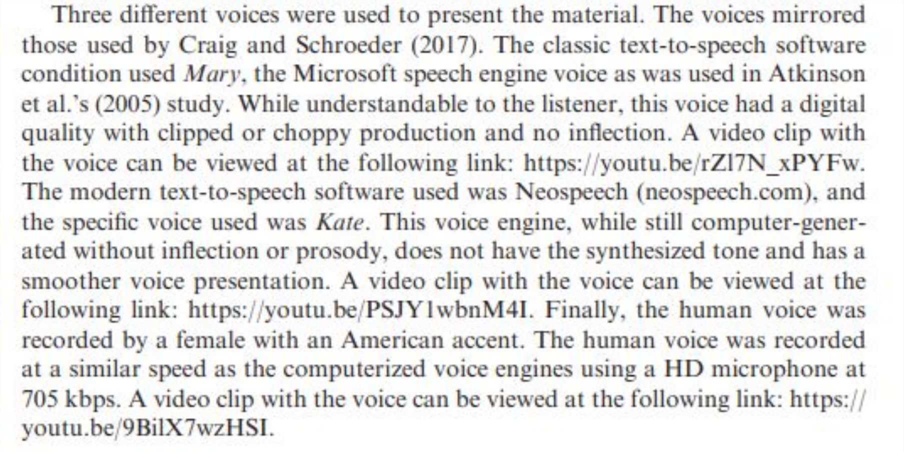
- Learning outcomes
  - Craig, S. D., & Schroeder, N. L. (2019). Text-to-Speech Software and Learning: Investigating the Relevancy of the Voice Effect. *Journal of Educational Computing Research*, 57(6), 1534–1548.
- 'Gender'/pitch
  - Tolmeijer, S., Zierau, N., Janson, A., Wahdatehagh, J. S., Leimeister, 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*.
  - 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.
- English-speaking accents
  - Tamagawa, R., Watson, C. I., Kuo, I. H., MacDonald, B. A., & Broadbent, E. (2011). The effects of synthesized voice accents on user perceptions of robots. *International Journal of Social Robotics*, 3(3), 253-262.

I looked at how they selected their voices.

Existing studies on TTS

Learning outcomes

Craig, S. D., & Schroeder, N. L. (2019). Text-to-Speech Software and Learning: Investigating the Relevancy of the Voice Effect.



Three different voices were used to present the material. The voices mirrored those used by Craig and Schroeder (2017). The classic text-to-speech software condition used *Mary*, the Microsoft speech engine voice as was used in Atkinson et al.'s (2005) study. While understandable to the listener, this voice had a digital quality with clipped or choppy production and no inflection. A video clip with the voice can be viewed at the following link: https://youtu.be/rZl7N_xPYFw. The modern text-to-speech software used was Neospeech (neospeech.com), and the specific voice used was *Kate*. This voice engine, while still computer-generated without inflection or prosody, does not have the synthesized tone and has a smoother voice presentation. A video clip with the voice can be viewed at the following link: https://youtu.be/PSJY1wbnM4I. Finally, the human voice was recorded by a female with an American accent. The human voice was recorded at a similar speed as the computerized voice engines using a HD microphone at 705 kbps. A video clip with the voice can be viewed at the following link: https://youtu.be/9BilX7wzHSI.

TLDR: Craig and Schroeder picked an ‘old’ voice (Microsoft Mary) and a ‘new’ one (Neospeech Kate)

But both voices were introduced in the early 2000s

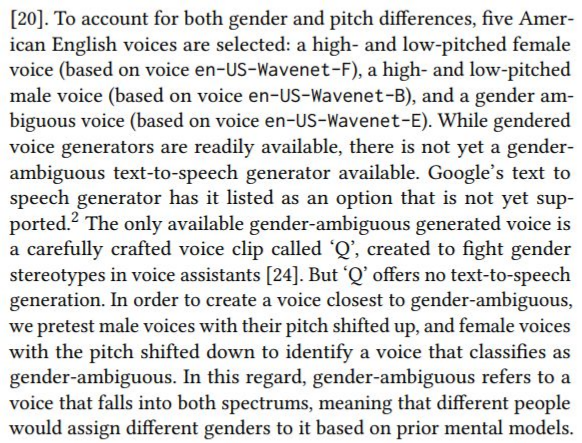
The human control voice was recorded by ‘a female with an American accent’

Not the best in terms of controlling for variation.

Existing studies on TTS

‘Gender’/pitch

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.



[20]. To account for both gender and pitch differences, five American English voices are selected: a high- and low-pitched female voice (based on voice `en-US-Wavenet-F`), a high- and low-pitched male voice (based on voice `en-US-Wavenet-B`), and a gender ambiguous voice (based on voice `en-US-Wavenet-E`). While gendered voice generators are readily available, there is not yet a gender-ambiguous text-to-speech generator available. Google’s text to speech generator has it listed as an option that is not yet supported.[2] The only available gender-ambiguous generated voice is a carefully crafted voice clip called ‘Q’, created to fight gender stereotypes in voice assistants [24]. But ‘Q’ offers no text-to-speech generation. In order to create a voice closest to gender-ambiguous, we pretest male voices with their pitch shifted up, and female voices with the pitch shifted down to identify a voice that classifies as gender-ambiguous. In this regard, gender-ambiguous refers to a voice that falls into both spectrums, meaning that different people would assign different genders to it based on prior mental models.

Tolmeijer et al. picked three different US English Google WaveNet voices

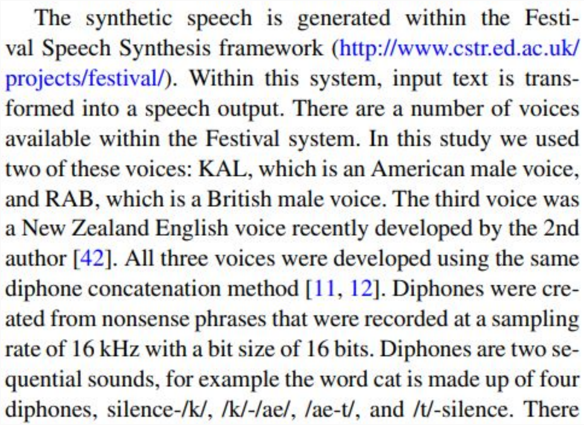
Then modified the pitch to produce high-pitched and low-pitched versions of the ‘male’ and ‘female’ voices and ensure their fifth voice was as gender-ambiguous as possible

(Most listeners tend to categorise voices as either male or female, even when it’s not intuitive)

Existing studies on TTS

Accents

Tamagawa, R., Watson, C. I., Kuo, I. H., MacDonald, B. A., & Broadbent, E. (2011). [The effects of synthesized voice accents on user perceptions of robots]().



The synthetic speech is generated within the Festival Speech Synthesis framework (http://www.cstr.ed.ac.uk/projects/festival/). Within this system, input text is transformed into a speech output. There are a number of voices available within the Festival system. In this study we used two of these voices: KAL, which is an American male voice, and RAB, which is a British male voice. The third voice was a New Zealand English voice recently developed by the 2nd author [42]. All three voices were developed using the same diphone concatenation method [11, 12]. Diphones were created from nonsense phrases that were recorded at a sampling rate of 16 kHz with a bit size of 16 bits. Diphones are two sequential sounds, for example the word cat is made up of four diphones, silence-/k/, /k/-/ae/, /ae-t/, and /t/-silence. There

This study was on New Zealanders' preferences for the voice of a healthcare robot

Tamagawa et al. used two Festival voices (UK and US accents), and created a third NZ accented one using the same method

They 'ensured the sampling rate and bit-size for all three voices was the same. Therefore, the output synthesized speech was of the same quality in terms of how easy the speech was to listen to'.

Existing studies on UVE

Researchers have studied UVE in the visual realm

Broadly speaking, there are two approaches to selecting/creating stimuli arrays:

A. Selecting some existing items (robots, animations, CG faces etc), then using norming studies to arrange them along the ‘realism’ dimension

B. Using photomanipulation to create the stages ‘in between’ an artificial and real human stimulus

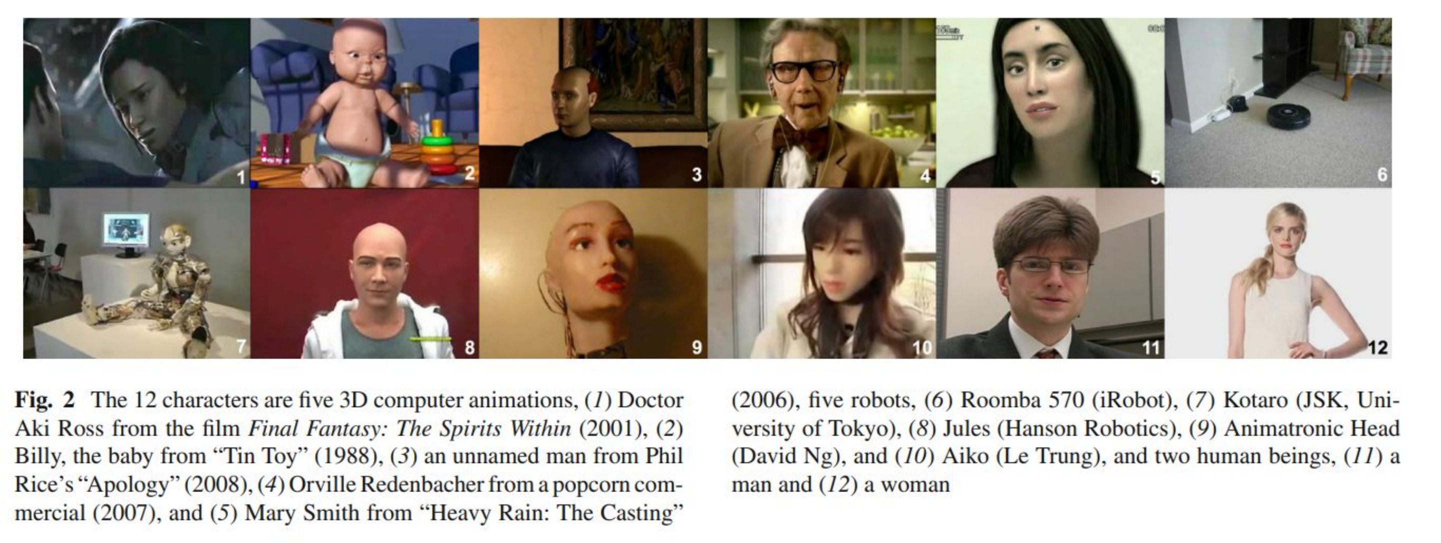
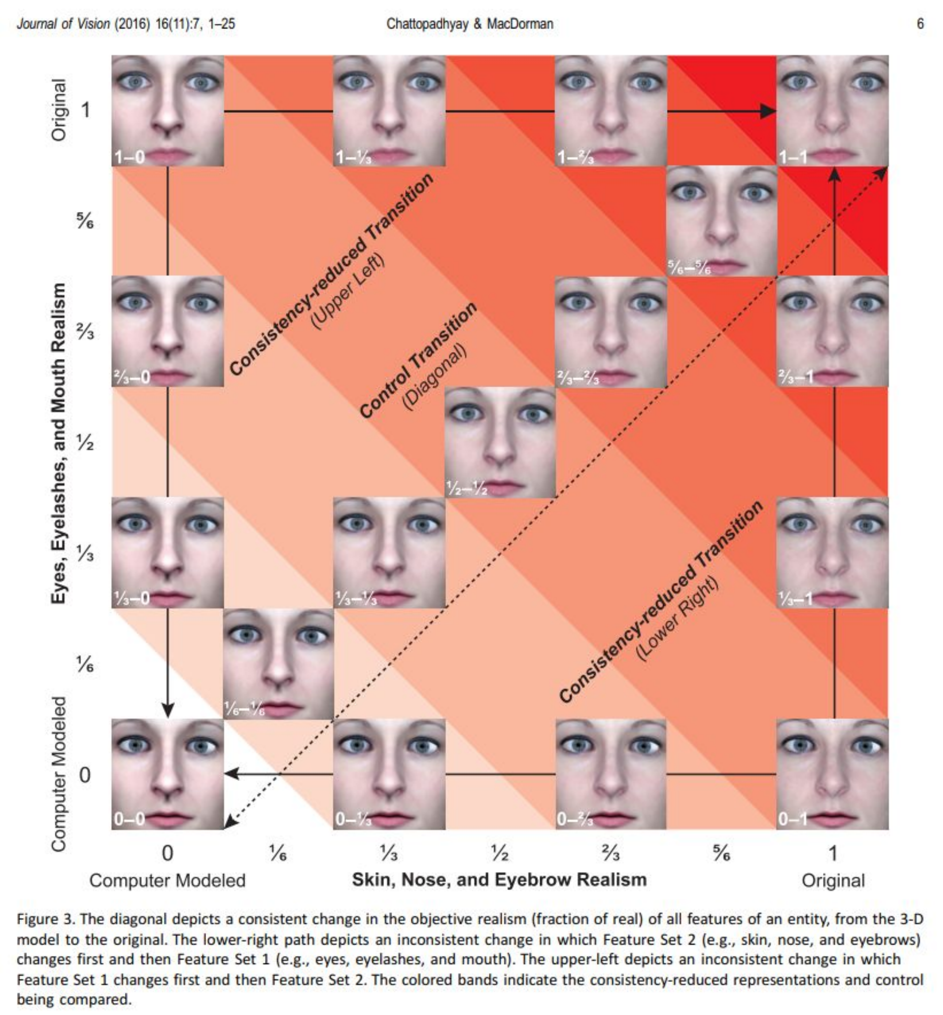


Fig. 2 The 12 characters are five 3D computer animations, (*1*) Doctor Aki Ross from the film *Final Fantasy: The Spirits Within* (2001), (*2*) Billy, the baby from "Tin Toy" (1988), (*3*) an unnamed man from Phil Rice's "Apology" (2008), (*4*) Orville Redenbacher from a popcorn commercial (2007), and (*5*) Mary Smith from "Heavy Rain: The Casting" (2006), five robots, (*6*) Roomba 570 (iRobot), (*7*) Kotaro (JSK, University of Tokyo), (*8*) Jules (Hanson Robotics), (*9*) Animatronic Head (David Ng), and (*10*) Aiko (Le Trung), and two human beings, (*11*) a man and (*12*) a woman

A.
Ho, C. C., & MacDorman, K. F. (2017). Measuring the uncanny valley effect. *International Journal of Social Robotics*, *9*(1), 129-139.

"The aim was to select robots from typical demonstration settings and 3D computer models from a variety of genres—short films, machinima, advertisements, and videogames—in addition to feature-length films. Two humans were added to extend the range of humanness."



Journal of Vision (2016) 16(11):7, 1–25 Chattopadhyay & MacDorman 6

Figure 3. The diagonal depicts a consistent change in the objective realism (fraction of real) of all features of an entity, from the 3-D model to the original. The lower-right path depicts an inconsistent change in which Feature Set 2 (e.g., skin, nose, and eyebrows) changes first and then Feature Set 1 (e.g., eyes, eyelashes, and mouth). The upper-left depicts an inconsistent change in which Feature Set 1 changes first and then Feature Set 2. The colored bands indicate the consistency-reduced representations and control being compared.

B.

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.

Back to my study

So, two possible strategies for my voice array:

- Pick some existing TTS voices and arrange them using a norming study
  - Keep them as similar as possible on all but one dimension: they should have the same fundamental frequency, accent, loudness, rate of speech, etc.
  - How many?

- Take one or two voices (e.g. a very ‘robotic’ voice and a real human one), then modify them to create ‘in between’ versions
  - How do we define ‘realism’ acoustically?
  - How do we modify them?
  - How many?

Over to you

What makes TTS voices sound good or bad? Are there any common problems you've faced when trying to make them sound more 'real'?

Another potential concern is about familiarity and novelty. I suspect many of my participants will likely be used to a given voice from using voice assistants. Is this something people have tackled before in TTS evaluations?

Perhaps it'd be better to have them explicitly compare two stimuli and say which one's better. That might help to situate the voices in a context so they're not implicitly comparing whatever they hear to, e.g., Siri?

Thanks for listening!

Please feel free to get in touch with me:

github.com/alice-ross

Teams - Alice Ross

or 🔗 email