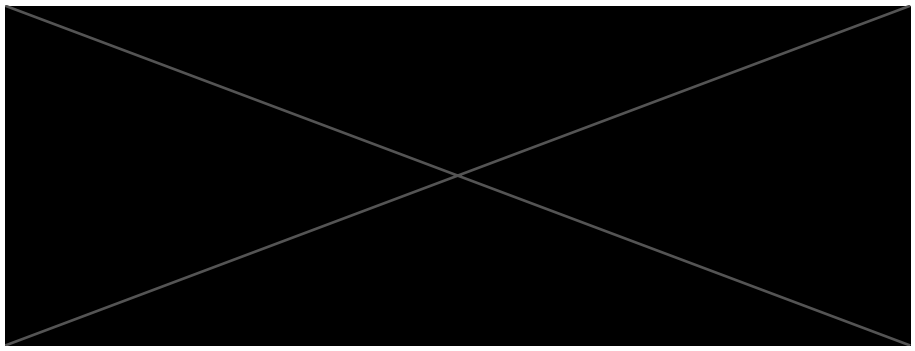

[redacted]

Online Experiments for Language Scientists: Annotated Bibliography

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.

In this experiment, participants watched a short presentation about lightning formation with voice-over narration by either a relatively old and unnatural-sounding text-to-speech synthesised voice (Microsoft Mary), a more modern synthesised voice (Neospeech Kate), or a human. They completed pre-test and post-test surveys to assess how successfully they had learned and retained information from the video, and a questionnaire on their judgments about the voice that they heard.

The authors conclude that there were minimal differences in learning results between participants in the newer machine-generated voice and the human voice conditions, and so their evidence does not support the 'voice principle' (Mayer, 2014), which states that human voices should be more effective than synthesised ones in learning environments. Although, unsurprisingly, the human voice was rated as more human-like, engaging, and credible than machine voices on the Likert questionnaire, the participants who learned from it did not show significantly better performance on any learning measures than those in the 'modern', Neospeech condition.

The discussion in this paper is very interesting and leaves several questions open for further research, particularly on personalisation, social agency and embodiment, and how learners perceive the sources of audio narration - a relevant point to investigating the voice principle, as an important objection in Mayer (2014) was that synthesised voices 'may not strongly convey the idea that someone is speaking directly to you.'

One weak point of the experiment is that a lot of data were lost as the final questionnaire, on perceptions of the voices, seems to have been skipped or incorrectly saved for 40% of the participants. The authors do not discuss their decision to use exclusively 'female' voices, or any potential effects of this on attitudes. Finally, I checked the example videos provided and although the two synthesised voices used certainly contrast, I would note that, for 2019, the 'modern' voice isn't particularly smooth or lifelike compared with other recently available TTS voices - the Wavenet ones used by Google Assistant, for example. Brief research suggests that it's also not much newer technology than Microsoft Mary, which was introduced with Windows 2000 in 1999; Neospeech introduced Kate in 2001, although it's possible that the voice has been upgraded since then.

While reading this paper, I wondered how listeners might rate the different machine-generated voices on other measures such as likeability and trustworthiness; whether this would vary according to the context and purpose of the voice (artificially intelligent agents can increasingly play assistive, instructive or even authoritative roles in humans' lives); and whether it might be possible to observe an auditory 'uncanny valley' effect, whereby synthesised voices which are almost - but not quite - indistinguishable from human ones might evoke feelings of unease. This is what I intend to investigate in my experiment for Assessment 2.

Mumm, J., & Mutlu, B. (2011). Designing motivational agents: The role of praise, social comparison, and embodiment in computer feedback. *Computers in Human Behavior, 27*(5), 1643–1650.

The effect explored in this paper is psychological, but still language-related as it involves a reading task and text feedback. Mumm and Mutlu used a 'chat' interface to present task instructions and verbal feedback on the participant's performance; feedback appeared as text in a speech bubble 'spoken' by a static image of a robot, in the agent conditions. The Wakamaru robot was chosen because it has a roughly humanoid appearance, enabling an embodiment effect on user motivation, but has no specific race or gender characteristics so participants' responses to it are less likely to be shaped by social biases. Having the robot character give instructions and feedback is potentially a good strategy for having participants attend to input from an AI agent. For my own experiment this input would be presented through audio, and I would aim to make voices race- and gender-ambiguous.

Participants completed a minimum of five rounds of a speed counting task which required them to correctly count the number of instances of a given letter in a nonsensical sentence, and were told that the experiment was on font size readability: this is a plausible 'cover' which may have helped to avoid participant non-naivety about the true purpose of the study, examining the effects of praise and social comparison on motivation. Participants could not submit inaccurate counts and would have to retry each round until they answered correctly. This could be a useful method to ensure that participants were paying attention - it seems impossible to quickly button-mash through this task by guessing.

One critique is that all participants only received $0.30 but the experiment could go on apparently indefinitely, as participants chose when to stop. This is an ethical problem, and can also be considered a practical one in the context of the authors' interest in task motivation - after a certain point they are looking at people's motivation to do unpaid work (as opposed to very low-paid work) which may have an effect. It is also noteworthy that because this paper compares twelve different conditions, there were only 16 participants per condition, and some of the statistical effects reported are much stronger than others, so the results' replicability may warrant further investigation.

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).

Tolmeijer et al examined perceptions of different synthesised voices in the context of their growing use as voice assistants, such as Siri. They used Google's WaveNet technology to make an array of five American English voices: high-and low-pitched variations of voices readily classified as male or female, and one 'gender ambiguous' voice, pitch-shifted such that listeners asked to identify the gender of the speaker were split roughly 50/50 (actually 58% 'male'). The experiment compared participants' attitudes to these voices in a 5 x 2 design, where the interaction either involved using a voice assistant to book a flight, or being asked personal questions in the context of a customer survey.

234 participants completed the experiment (345 before data cleaning); they were all US nationals, aged between 19 and 74, recruited using Prolific. The authors do not disclose details of payment. If over 100 people were paid for taking part, but their data were excluded due to being incomplete, then it might be worth checking if the efficiency of the experiment design could have been improved to avoid losing data; alternatively, if those participants were not paid despite completing some of the task, this could be problematic from an ethics standpoint. More positively, the sample size and range of ages is relatively large.

A number of interesting effects are seen in the results. The investigation of stereotypical masculine and feminine traits ascribed to the voices is somewhat inconclusive: scores are quite low across the board, so all the statistically significant results are negative - e.g. the low male voice was the only one to receive middling rather than low

scores as 'dominant'; all voices had slightly positive scores for 'friendly' but the female ones scored higher than the others. Task context affected participants' rating of trust, but notably, there was no significant difference in trust rating for the gender-ambiguous voice against the others. This is important to the authors - and to me - because most existing voice assistant systems have been coded as female by default, and concerns have been raised about this tendency potentially reproducing and reinforcing societal biases, possibly even inducing younger generations who are growing up with this technology to associate stereotypical traits like subservience and limited agency with women (and feminine men and non-binary persons). Evidence showing that this design trend is not led by some overwhelming user preference for female synthesised voices - in fact, it may have more to do with disproportionately male teams of software developers - is therefore informative for future designs.

A final pertinent point raised by Tolmeijer et al is that producing an 'ambiguous' voice, with pitch in between typical male and female ranges, does not guarantee that listeners will experience it as genderless or gender-neutral; it seems that most people still classify voices using binary categorical perception, which may complicate any effort to eradicate gender bias in attitude surveys.

Belin, P., Boehme, B., & McAleer, P. (2017). The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PloS one*, *12*(10), e0185651.

Belin et al examined people's first impressions of personality based on voice. They created nine tokens of a male voice saying 'hello' with varying intonation (slightly rising through to falling-rising f0 contour), by using computational modelling on the acoustical averages of several human voices rated as more or less trustworthy in a previous study. 500 participants rated these new synthesised 'hellos' on their speakers' perceived trustworthiness; the trustworthiness rating z-scores were found to be highly consistent and could be mapped to a neat continuum from lowest to highest rated, correlated with the increasing variation in pitch contour.

As Belin et al's findings are limited to first impressions formed after hearing only one word, it's probably unwise to extrapolate to more sustained interactions or other contexts; however, it is of interest that the voice rated least trustworthy is relatively flat and monotonous in pitch, which is also a feature of older machine-generated voices. This could be a good starting point for a more detailed examination of how synthesised voices can be

manipulated for more positive user attitudes, potentially using much more complex experiments such as the wagering game implemented by Mathur & Reichling to investigate trust in robot faces. A strong point of this paper is that the stimuli and raw data are all available to download so that readers can investigate, analyse and potentially replicate the work.

For my planned work, the most relevant part is the practical information on how to produce synthesised voice samples by using STRAIGHT in Matlab, to manipulate different parameters of a recorded voice. It's also interesting that the experiment used such a straightforward paradigm with no context or 'cover task' involved, making it a lot simpler to implement than the other experiments discussed above.

Additional references

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition*, *146*, 22-32.

Mayer, R. (2014). Principles Based on Social Cues in Multimedia Learning: Personalization, Voice, Image, and Embodiment Principles. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (Cambridge Handbooks in Psychology, pp. 345-368). Cambridge: Cambridge University Press.

3 [redacted]

GRADEMARK REPORT

FINAL GRADE

88/100

GENERAL COMMENTS

Instructor

Entry 1: 10/10

Entry 2: 8/10

Entry 3: 8/10

Entry 4: 7/10

TOTAL: 33/40 = 83%

+5% mark adjustment (applied to all assignments)

PAGE 1

QM

QM

Comment 1

Good link

Comment 2

Excellent summary of both the methodology and results

PAGE 2

QM

Comment 3

Excellent discussion of connections to other literature, thoughtful comments about the authors' methodological decisions and interesting thoughts about Assignment 2.

Comment 4

Interesting idea

QM

QM

Comment 5

Unclear what this condition is or what other conditions there were

Comment 6

Good point

Comment 7

Important point for online experiments so good to discuss this

QM

Comment 8

Good to make the connection between rate of pay and the validity of results

Comment 9

Important point

QM

QM

Comment 10

Good summary - some very specific details (e.g. participants' responses to the gender-ambiguous voice, ages and locations of participants) could have been omitted to make it a bit sharper

Comment 11

Useful to get a brief description of what the authors found, but it's not clear how some of the results you've presented relate to your evaluation of the methodology

Comment 12

Good connection to wider practical/ethical issues

QM

QM

QM

Comment 13

Not entirely clear why this is relevant

Comment 14

Good point

Comment 15

Good to make connections to other literature but this is a little hard to interpret without knowing the work you're referring to

QM

Comment 16

Open science - good to be aware of these issues

Comment 17

Good connection to the next assignment but this is perhaps slightly too detail-oriented and lacking in critical evaluation - how does the method of producing synthesised voice samples relate to what your planned experiment might be able to tell you about your research question(s)?