



This is a repository copy of *A 'canny' approach to spoken language interfaces*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/153056/>

Version: Submitted Version

Article:

Moore, R.K. orcid.org/0000-0003-0065-3311 (Submitted: 2019) *A 'canny' approach to spoken language interfaces*. arXiv. (Submitted)

© 2019 The Author(s). For reuse permissions, please contact the Author(s).

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

A ‘Canny’ Approach to Spoken Language Interfaces

Roger K. Moore
University of Sheffield
Sheffield, S. Yorks., UK
r.k.moore@sheffield.ac.uk

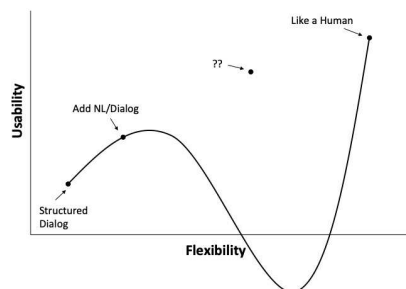


Figure 1: Increasing the flexibility of spoken language dialogue systems can lead to a ‘habitability gap’ [8].

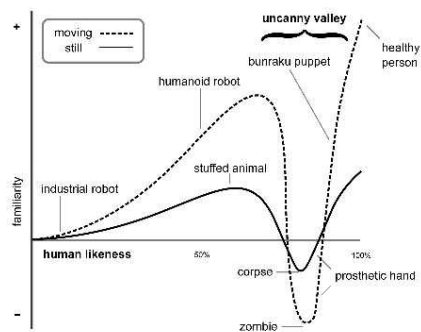


Figure 2: The ‘uncanny valley’ effect [7].

ABSTRACT

Voice-enabled artefacts such as Amazon Echo are very popular, but there appears to be a ‘habitability gap’ whereby users fail to engage with the full capabilities of the device. This position paper draws a parallel with the ‘uncanny valley’ effect, thereby proposing a solution based on aligning the visual, vocal, behavioural and cognitive affordances of future voice-enabled devices.

KEYWORDS

voice enabled devices; habitability gap; uncanny valley effect; aligned affordances

INTRODUCTION

Recent years have witnessed astonishing progress in the development of voice-enabled artefacts such as *Siri* (released by Apple in 2011) and *Alexa* (released by Amazon in 2014). For example, tens of millions of Alexa-enabled devices were sold worldwide over the 2017 Christmas holiday season, and Apple’s *Siri* had 41.4 million monthly active users in the U.S. as of July 2017 [2]. Indeed, the appearance of such ‘intelligent’ personal assistants is often hailed as a significant step along the road towards more natural interaction between human beings and future ‘autonomous social agents’ (such as robots).

CHI 2019 Workshop on Speech Interface Interactions, 4-9 May 2019, Glasgow, UK

© 2019 Association for Computing Machinery.

This is the author’s version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *Proceedings of Workshop on Mapping Theoretical and Methodological Perspectives for Understanding Speech Interface Interactions (CHI 2019 Workshop on Speech Interface Interactions)*, https://doi.org/10.475/123_4.

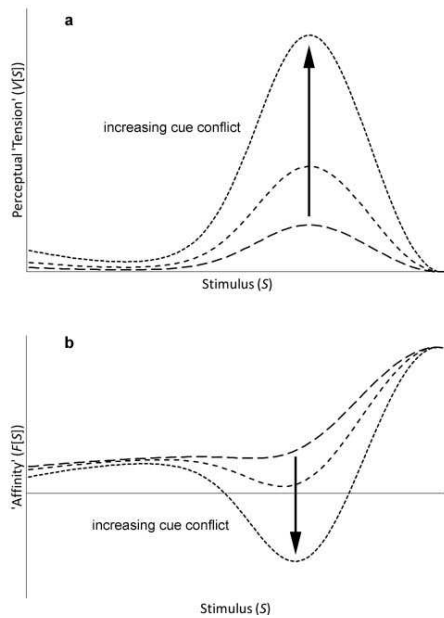


Figure 3: (a) Perceptual 'tension' increases at a category boundary as a result of misaligned perceptual cues. (b) Peaks in perceptual tension give rise to dips in 'affinity', i.e. increases in 'uncanniness' [3].

However, studies into the usage of such technology suggest that, far from engaging in a promised natural 'conversational' interaction, users tend to resort to formulaic language and focus on a handful of niche applications which work for them [6]. Given the pace of technological development, it might be expected that the capabilities of such devices will improve steadily, but evidence suggests that there is a 'habitability gap' (see Figure 1) in which usability drops as flexibility increases [8].

THE PROBLEM

It has been hypothesised that the habitability gap is a manifestation of the 'uncanny valley' effect (see Figure 2) whereby a near human-looking artefact (such as a humanoid robot) can trigger feelings of eeriness and repulsion [7]. In particular, a Bayesian model of the uncanny valley effect [3] reveals that it can be caused by *misaligned* perceptual cues (see Figure 3). Hence, a device with an *inappropriate* voice can create unnecessary confusion in a user. For example, the use of human-like voices for artificial devices encourages users to overestimate their linguistic and cognitive capabilities.

A WAY FORWARD

The Bayesian model of the uncanny valley effect suggests that the habitability gap can only be avoided if the visual, vocal, behavioural and cognitive *affordances* of an artefact are aligned. Given that the state-of-the-art in these areas varies significantly, this means that the capabilities of an artificial agent should be determined by the affordance with the lowest capability. In other words, emulating a human is a recipe for failure, rather "*it is better to be a good machine than a bad person*" [1].

So, the theoretical perspective discussed above suggests a whole-system design approach in which the characteristics of each element must be selected in accordance with the characteristics of all other elements. In particular, the voice of an artefact should be selected by taking into account all other aspects of the design. For example, the vocal tract length should be based on the physical size of the device, the vocal timbre should be based on the construction material, and the linguistic complexity of its utterances should be conditioned on its underlying cognitive abilities. Failure to follow this approach will lead to the creation of yet more voice-enabled *chimeras* sitting solidly in the habitability gap.

CONCLUSION

Notwithstanding the immense progress that has been made in voice-enabled artefacts, future progress depends on designers taking a whole-system perspective and ensuring that the visual, vocal, behavioural and cognitive affordances are aligned. Inspiration can be taken from fictional characters in cinema and television [9], and such an approach will open up a plethora of imaginative and yet appropriate voices [5] - altogether a more 'canny' approach to the development of intelligent communicative machines [4].

REFERENCES

- [1] Bruce Balentine. 2007. *It's Better to Be a Good Machine Than a Bad Person: Speech Recognition and Other Exotic User Interfaces at the Twilight of the Jetsonian Age*. ICMI Press, Annapolis.
- [2] Clark Boyd. 2018. The Past, Present, and Future of Speech Recognition Technology. <https://medium.com/swlh/the-past-present-and-future-of-speech-recognition-technology-cf13c179aaf>.
- [3] Roger K. Moore. 2012. A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. *Nature Scientific Reports* 2, 864 (2012), doi:10.1038/srep00864. <https://doi.org/10.1038/srep00864>
- [4] R K Moore. 2015. From talking and listening robots to intelligent communicative machines. In *Robots That Talk and Listen*, J Markowitz (Ed.). De Gruyter, Boston, MA, Chapter 12, 317–335.
- [5] R. K. Moore. 2017. Appropriate voices for artefacts: some key insights. In *1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017)*. VIHAR, Skovde, Sweden, 7–11.
- [6] Roger K. Moore, Hui Li, and Shih-Hao Liao. 2016. Progress and prospects for spoken language technology: what ordinary people think. In *INTERSPEECH*. ISCA, San Francisco, CA, 3007–3011.
- [7] Masahiro Mori. 1970. Bukimi no tani (the uncanny valley). *Energy* 7 (1970), 33–35.
- [8] Mike Phillips. 2006. Applications of spoken language technology and systems. In *IEEE/ACL Workshop on Spoken Language Technology (SLT)*, Mazin Gilbert and Hermann Ney (Eds.). IEEE, Aruba, 7.
- [9] S. Wilson and R. K. Moore. 2017. Robot, alien and cartoon voices: implications for speech-enabled systems. In *1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017)*. VIHAR, Skovde, Sweden, 40–44.