Appropriate Voices for Artefacts: Some Key Insights

Roger K. Moore

Speech & Hearing Research Group, Dept. Computer Science, University of Sheffield, UK

r.k.moore@sheffield.ac.uk

Abstract

The 2011 release of *Siri* hailed the beginning of a sustained period of impressive advances in the capability and availability of spoken language technology. Subsequent years saw the appearance of competitors such as *Google Now*, swiftly followed by consumer products such as Amazon *Echo*. These devices are seen as the first steps towards more advanced 'conversational' artefacts (especially *robots*). However, evidence suggests that the usage of such voice-enabled devices is surprisingly low, perhaps due to noise in the environment, privacy concerns or manual alternatives.. Another possible contributing factor is that the ubiquitous deployment of *inappropriate* humanlike voices for non-living artefacts might deceive users into overestimating their capabilities, thereby creating a conflict of expectations that ultimately leads to a breakdown in communications. This paper highlights the benefits of providing an *appropriate* voice for a given artefact based on three separate studies. Results are presented that demonstrate the positive impact of a non-human voice and illustrate how 'appropriateness' might be measured objectively. Finally, a worked-example is presented of implementing an appropriate voice for the *MiRo* biomimetic robot. It is concluded that these insights could be important for the design of future generations of voice-enabled artefacts.

Index Terms: appropriate voices, robot voices, speaking artefacts

1. Introduction

After more than 40 years of research into spoken language processing, the 2011 release of *Siri* - Apple's voice-based 'personal assistant' for the iPhone - represented a significant milestone in bringing speech technology to the attention of the general public. It also hailed the beginning of a sustained period of impressive advances in the capabilities of the underlying speech technologies with dramatic improvements in the accuracy of 'automatic speech recognition' (ASR) and the quality of 'text-to-speech synthesis' (TTS). Subsequent years saw the appearance of smartphone-based competitors to *Siri* such as *Google Now* and Microsoft's *Cortana*, swiftly followed by voice-enabled consumer products such as Amazon *Echo* and Google *Home*. These latter devices are seen as the first stepping stones towards more advanced 'conversational' artefacts in the future, in particular 'automonous social agents' (such as robots) - see Fig. 1.

Notwithstanding the popularity of contemporary voice-enabled devices, it appears that actual usage is surprisingly low (see Fig. 2) [1]. Indeed, it seems that voice interfaces maintain their notoriety for "*fostering frustration and failure*" [2].

There are a number of potential explanations for this lack of genuine take-up: e.g. noise in the environment, privacy concerns or manual alternatives. However, it is argued here that another contributing factor could be the ubiquitous deployment of humanlike voices for artefacts that are clearly not human. Not only is this true of mainstream speech-based systems such as *Siri* and *Echo*, but it is also typical to find that robot research laboratories have equipped their devices with off-the-shelf humanlike speech synthesis on the basis that it's "*natural*" that people should wish to interact with a robot using 'normal' speech. The reality is that, when faced with such artefacts, users tend to be deceived into overestimating their capabilities, creating a conflict of expectations that ultimately leads to a breakdown in communications (much like the famous 'uncanny valley' in robotics [3, 4, 5]) - the opposite of what was intended.

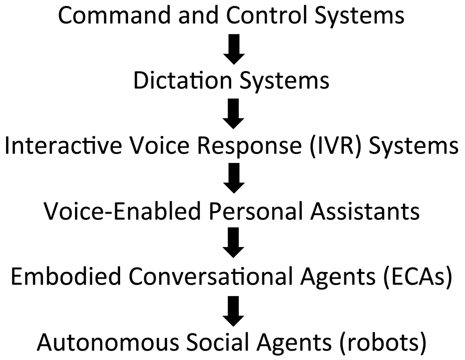
Command and Control Systems
Dictation Systems
Interactive Voice Response (IVR) Systems
Voice-Enabled Personal Assistants
Embodied Conversational Agents (ECAs)
Autonomous Social Agents (robots)

Figure 1: *The evolution of spoken language technology applications from the first 'voice command' systems of the 1970s, through contemporary smartphone-based 'personal assistants' (such as Siri) to future 'autonomous social agents' (i.e. robots).*

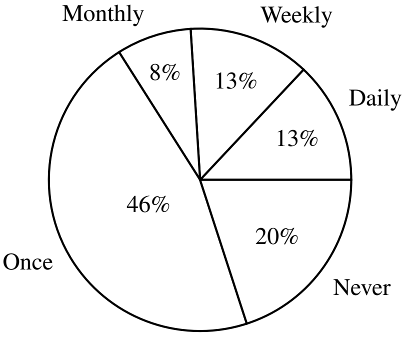

Figure 2: *Speech technology usage on smartphones [1].*

In practice, it would be relatively easy to manage users' expectations by giving artefacts an appropriate *non-human*, rather than humanlike, voice. In principle, such an approach would avoid the pitfalls of the 'uncanny valley' by aligning an artefact's visual, vocal and behavioural *affordances* [6, 7, 8], and would create a more 'habitable' interface in line with the ideas expressed in Bruce Balentine's seminal book on the usability of spoken language systems: "*It's Better to be a Good Machine than a Bad Person*" [9]. However, for one reason or another, deploying a robotic voice is still an unpopular idea, and mainstream speech technology R&D continues to strive for voices that as as humanlike as possible [10].

This paper brings together three separate studies which support the general hypothesis that there are benefits to be gained from providing artefacts with *appropriate* voices. Section 2 reprises experiences using a robotic voice in a genuine telephone-based travel planning service, Section 3 describes an experiment that was designed to measure vocal appropriateness, and Section 4 presents a worked-example of implementing an appropriate voice for a biomimetic robot. Finally, Section 5 concludes that this paper has brought together a number of important insights into the potential benefits and practical steps required to create appropriate voices for artefacts.

2. Experiences with a Genuine Telephone-based Travel Planning Service

The first study was conducted some years ago while the author was Head of the UK Government's *Speech Research Unit* (SRU). At the time, there was burgeoning interest in 'spoken language dialogue systems' (SLDS), and there was a need to collect corpora of speech-based transactions for study. Much of the SLDS research during that period was based on *simulated* applications, so a project was established at SRU to attempt to collect *real* conversations in a task-based dialogue - in this case, a telephone-based travel planning service.

2.1. The Setup

As is common in the SLDS research area, a 'Wizard-of-Oz' (WoZ) arrangement was used in which a human operator plays the role of all or part of a supposedly automated system. However, what was special about the SRU study was (i) the service was *genuine* (in that it was advertised with no mention that it was experimental or automated or connected with the SRU), and (ii) callers to the service were handled either by a human operator (in 'normal' mode) or by the same operator with a modified robotic-sounding voice (in 'WoZ' mode).

The enquiry service was configured around a commercially available route planning software package running on a PC. Its main feature was its ability to find the shortest and/or quickest routes between two locations in accordance with a range of user-specified preferences. Such software was not readily available to ordinary members of the public at the time. The call handling system was configured to operate with two incoming telephone lines - one assigned to the human operator's normal voice and one assigned to the robotic voice - and, in order for there to be minimal differences between the operator's behaviour in both conditions, the same operator was used in each case. The WoZ voice was created using a 'voice disguise' unit which changed the talker's pitch and then combined the natural and altered signals to produce a robotic, yet fully intelligible, vocal timbre. On receipt of a call, the operator (in normal or WoZ mode) always used the same introductory announcement: "*Welcome to the route planning service - how can I help you?*".

2.2. Results

The full results were published shortly after the study [11, 12], but the key outcome was the observation that the robotic voice had a dramatic effect on the behaviour of the callers (who, immediately upon hearing the robotic voice, genuinely believed that they had been connected to a fully automated system). The main effect was that callers in WoZ mode did not engage in lengthy social exchanges; they did not feel obliged to explain to the (apparently) automated system *why* they wanted to travel. As a consequence, WoZ-based transactions were considerably more efficient in terms of task completion. In particular, the average number of words spoken by each caller was reduced from 186 in response to the humanlike voice to just 31 for the robotic voice: an 83% reduction. Also, disfluencies were reduced by an order-of-magnitude (see Fig. 3).

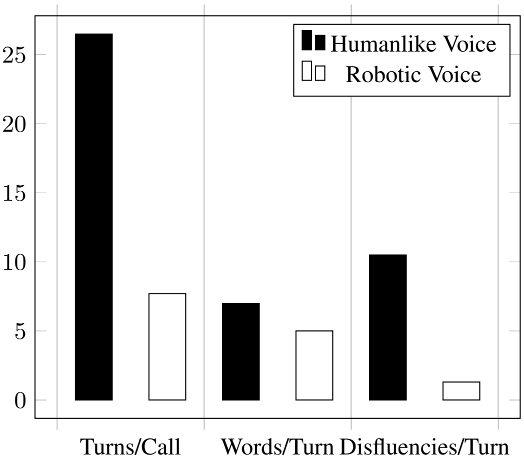

Figure 3: *The effect of the operator's voice on various measures in the telephone-based travel planning service.*

Overall, the results of this study made it clear that merely changing the timbre of a voice can have a dramatic effect on an interlocutor's interactional behaviour. In particular, an *appropriate* robotic voice can successfully reflect the limited social capabilities of an automated system, thereby facilitating efficient and successful voice-based transactions.

3. Measuring Vocal Appropriateness

The second study reported here was conducted as part of the EU-funded project *Social Engagement with Robots and Agents* (SERA). SERA was aimed at investigating the social acceptability of verbally interactive robots and agents, and it conducted long-term field trials in which a *Nabaztag* robot was placed in elderly people's homes to provide advice and encouragement about maintaining an active and healthy lifestyle.

Nabaztag is a 23 cm high WiFi-enabled highly-stylised plastic rabbit with flashing lights on its belly and nose, and rotating ears (see Fig. 4). Subjects described the robot as cute, comical and somewhat like an animation character (particularly *Pokemon*). Feedback from the initial field trials suggested that the agent must be friendly, likeable, polite and submissive, and that *its voice should be consistent with its visual appearance*.

Nabaztag's voice was generated using a state-of-the-art text-to-speech synthesiser (provided by *Loquendo*). Therefore, in order to meet the requirement that the voice should be consistent with the character of the robot, an experiment was conducted to select the most appropriate voice: the default (adult male) voice or one that was more childlike. The aim of the experiment was not simply to ask people's subjective opinions, but to attempt to measure appropriateness *objectively*.

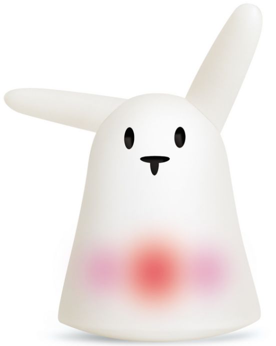

Figure 4: *The Nabaztag robot.*

3.1. The Experiment

3.1.1. Approach distance

The first part of the experiment investigated an established measure based on 'approach distance'. Previous research on 'proxemics' had suggested that the size of the space between humans reflects (and influences) their social relationships and their attitudes to each other [13, 14]. Other studies found that inanimate objects are generally approached closer than other humans [15], and that users do not always respect a robot's interpersonal space (by moving very near to it) [16]. Hence, maintaining a proper social space between a robot and a human had been hypothesised to express the acceptance of the robot as a social actor, and that the distance was influenced by the voice [17].

3.1.2. Dislocation perception

The second part of the experiment investigated a new measure based on 'dislocation perception'. Inspired by the 'ventriloquist effect' [18], it was hypothesised that an appropriate voice for an agent could be physically displaced from an artefact and yet still be perceived as emanating from it: the more appropriate the voice, the larger the displacement. In order to test this for the different synthetic voices, the Nabaztag robot was placed in front of an acoustically transparent screen, and its voice was played through a hidden loudspeaker 29 cm to the side of the robot's 'mouth'. This meant that, at a distance of 120 cm, the voice from the robot was at an angle of approximately 12°, well over the minimum audible angle (MAA) of 1-2° [19].

3.1.3. Subjects

46 normal hearing subjects were recruited for the experiment, all of whom had with little or no prior exposure to agents/robots. Each subject was exposed to one voice only, and met the robot in a specially prepared room, with the agent approximately 3.5 metres from the entrance. Once in the room, the subject was instructed to keep eye contact with the robot, and to wait for it to invite them to come closer. When it was confirmed that they were looking at the robot, it would say: "`Hello, I've been expecting you - please come closer`". The subject then moved towards the robot, and the approach distance was noted.

The robot would then ask the researcher to offer the subject a seat, and a chair was placed directly in front (120 cm from the robot). This ensured that each subject faced the agent at approximately 0° azimuth and elevation. The robot then delivered a short speech explaining its role and purpose, finishing with: "`It was so nice of you to offer to help with this, thank you - now the researcher would like to ask you a few questions`". The researcher informed the subject that the experiment was over and led him/her away from the robot, but then casually asked: "*By the way, where did you think the voice came from - the robot or somewhere else?*", and the response was noted.

3.2. Results

The results of the 'approach distance' experiment are shown in Fig. 5. As can be seen, the majority of subjects chose to occupy the robot's 'personal space' regardless of the selected voice.

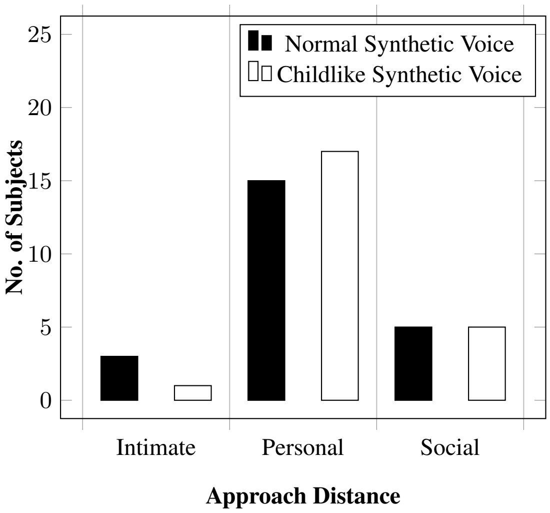

Figure 5: *The number of subjects that entered the Nabaztag robot's 'intimate space' (∼30 cm), 'personal space' (∼80 cm) or 'social space' (∼120 cm). The differences in the responses for the two voices are not statistically significant.*

The results of the 'dislocation perception' experiment are shown in Fig. 6. In this case, there is a clear (and statistically significant) difference between the subjects' responses for the two voices. As expected, the childlike synthetic voice benefitted from the 'ventriloquist effect' and was perceived by the majority of subjects to be emanating from the robot.

Overall, the results of this study suggested that, contrary to expectations, 'approach distance' is not a good objective measure of the appropriateness of a voice to an artefact, whereas 'dislocation perception' appeared to be quite effective [20].

4. A Voice for a Biomimetic Robot

The third study reported here concerns the design of a voice for *MiRo*: a highly featured, low-cost, programmable robot, with a friendly animal-like appearance, six senses, a nodding and rotating head, moveable hearing-ears, large blinking seeing-eyes, and a wagging tail. Designed and built by Consequential Robotics Ltd. in collaboration with the University of Sheffield [21], *MiRo* has been designed to look like a cartoon hybrid of a generic mammal (see Fig. 7) and is targeted at a range of applications such as assistance, companionship, pet therapy and edutainment. A unique brain-based biomimetic control system [22, 23] allows *MiRo* to behave in a life-like way: for example,

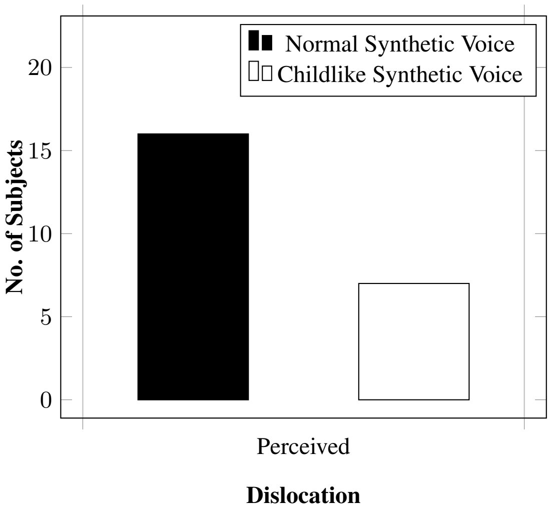

Figure 6: *The number of subjects that perceived the dislocation between the location of the Nabaztag robot and the source of its voice.*

listening for sounds and looking for movement, then approaching and responding to physical and verbal interactions.

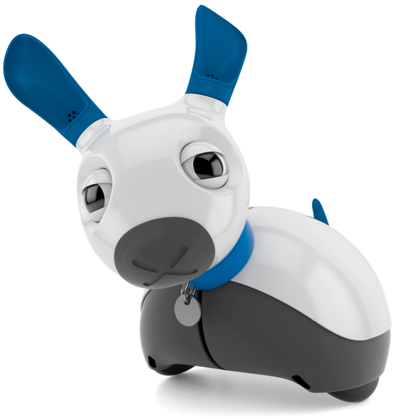

Figure 7: *The MiRo biomimetic robot.*

4.1. The Robot

MiRo is constructed around a differential drive base and a neck with three Degrees-of-Freedom (DoF). Additional DoFs include rotation for each ear, tail droop/wag, and eyelid open/close. All DoFs are equipped with proprioceptive sensors, and there is an on-board loudspeaker. The robot is equipped with stereo cameras in the eyes, stereo microphones in the ears and a sonar range-finder in the nose. Four light-level sensors are placed at each corner of the base, and two infrared 'cliff' sensors point down from its front. Eight capacitive sensors are arrayed along the inside of the body shell and over the top and back of the head. Internal sensors include twin accelerometers, a temperature sensor and battery-level monitoring.

MiRo represents its affective state (emotion, mood and temperament) as a point in a two-dimensional space covering valence (unpleasantness-pleasantness) and arousal (calm-excited) [24, 25]. Events arising in *MiRo*'s sensorium are mapped into changes in affective state: for example, stroking *MiRo* drives valence in a positive direction, whilst striking *MiRo* on the head drives valence in a negative direction. *MiRo*'s movements are modulated by its affective state, and it also expresses itself using a set of 'social pattern generators' that drive light displays, movement of the ears, tail, eyelids and *vocalisation*.

4.2. *MiRo*'s Voice

MiRo's ability to vocalise was achieved using a real-time parametric general-purpose mammalian vocal synthesiser [26] tailored to the physical and behavioural characteristics of the robot [27]. The overall structure of the synthesis software is based on a simulation of the flow of energy through a generic mammalian vocal apparatus with an appropriate body mass.

In order to allow the injection of emotion into the vocalisations, key parameters were linked to *MiRo*'s two-dimensional affect map. Arousal modulates the airflow rate and, thereby, the amplitude and tempo of the vocalisations; high arousal leads to high airflow and short vocalisations (and *vice versa*). Valence influences the variance of the fundamental frequency and the voice quality; high valence leads to expressive vocalisation whereas low valence produces more monotonic utterances. For example, stroking *MiRo*'s head increases valence, which leads to 'happier' vocalisations (and a wagging tail).

The outcome of this design approach has been the creation of an 'appropriate' voice for *MiRo* that is perfectly aligned to the physical and behavioural affordances of the robot. It thus successfully avoids the 'uncanny valley' effect mentioned in Section 1 and contributes strongly to the effectiveness of *MiRo* as an attractive interactive vocal agent.

5. Summary and Conclusion

It has been argued that that one reason users fail to engage successfully with speech-enabled devices is the ubiquitous deployment of humanlike voices for artefacts that are clearly not human. Hence, it has been hypothesised that users' expectations could be better managed by giving artefacts an *appropriate* non-human voice, e.g. a voice that is intelligible but robotic.

This paper has brought together three separate studies which support the hypothesis. First, experiences with a genuine telephone-based travel planning service confirmed that an appropriate non-human voice can have a dramatic and beneficial effect on the behaviour of naïve users. Second, results of a study to measure vocal appropriateness *objectively* revealed that 'approach distance' was not a good measure of the appropriateness of a voice to an artefact, whereas 'dislocation perception' proved to be quite effective. Third, a worked-example has been presented of implementing an appropriate voice for a biomimetic robot.

Overall, this paper has highlighted a number of important insights into the potential benefits and practical steps required to create appropriate voices for future generations of voice-enabled artefacts.

6. Acknowledgements

This work was partially supported by the European Commission [grant numbers EU-FP6-507422, EU-FP6-034434, EU-FP7-231868 and EU-FP7-611971], and the UK Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/I013512/1].

7. References

[1] R. K. Moore, H. Li, and S.-H. Liao, “Progress and prospects for spoken language technology: what ordinary people think,” in *INTERSPEECH*, San Francisco, CA, 2016, pp. 3007–3011.

[2] C. Nass and S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-computer Relationship*. Cambridge, MA: MIT Press, 2005.

[3] M. Mori, “Bukimi no tani (the uncanny valley),” *Energy*, vol. 7, pp. 33–35, 1970.

[4] W. J. Mitchell, K. A. Szerszen Sr., A. S. Lu, P. W. Schermerhorn, M. Scheutz, and K. F. MacDorman, “A mismatch in the human realism of face and voice produces an uncanny valley,” *i-Perception*, vol. 2, no. 1, pp. 10–12, 2011.

[5] R. K. Moore, “A Bayesian explanation of the Uncanny Valley’ effect and related psychological phenomena,” *Scientific Reports*, vol. 2, no. 864, 2012. [Online]. Available: http://www.nature.com/articles/srep00864

[6] J. J. Gibson, “The theory of affordances,” in *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, R. Shaw and J. Bransford, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1977, pp. 67–82.

[7] R. K. Moore, “Spoken language processing: Where do we go from here?” in *Your Virtual Butler, LNAI*, R. Trappl, Ed. Heidelberg: Springer, 2013, vol. 7407, pp. 111–125.

[8] ——, “From talking and listening robots to intelligent communicative machines,” in *Robots That Talk and Listen*, J. Markowitz, Ed. Boston, MA: De Gruyter, 2015, ch. 12, pp. 317–335.

[9] B. Balentine, *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*. Annapolis: ICMI Press, 2007.

[10] P. Taylor, *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press, 2009.

[11] R. K. Moore and A. Morris, “Experiences collecting genuine spoken enquiries using WOZ techniques,” in *5th DARPA workshop on Speech and Natural Language*, New York, 1992, pp. 61–63.

[12] R. K. Moore and S. R. Browning, “Results of an exercise to collect ’genuine’ spoken enquiries using WoZ techniques,” in *Institute of Acoustics Speech and Hearing Conference*, Windermere, 1992.

[13] E. T. Hall, R. L. Birdwhistell, B. Bock, P. Bohannan, A. R. Diebold, M. Durbin, M. S. Edmonson, J. L. Fischer, D. Hymes, S. T. Kimball, W. La Barre, J. E. McClellan, D. S. Marshall, G. B. Milner, H. B. Sarles, G. L. Trager, A. P. Vayda, and A. P. Vayda, “Proxemics,” *Current Anthropology*, vol. 9, no. 2/3, pp. 83–108, 1968.

[14] R. Sommer, *Personal Space. The Behavioral Basis of Design.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1969.

[15] M. J. Horowitz, D. F. Duff, and L. O. Stratton, “Body-Buffer Zone,” *Archives of General Psychiatry*, vol. 11, no. 6, pp. 651–656, 1964.

[16] C. L. Breazeal, *Designing Sociable Robots*. MIT Press, 2004.

[17] M. L. Walters, D. S. Syrdal, K. L. Koay, K. Dautenhahn, and R. te Boekhorst, “Human approach distances to a mechanical-looking robot with different robot voice styles,” in *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2008, pp. 707–712.

[18] D. Alais and D. Burr, “The Ventriloquist Effect Results from Near-Optimal Bimodal Integration,” *Current Biology*, vol. 14, no. 3, pp. 257–262, 2004.

[19] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Academic Press, 2003.

[20] R. K. Moore and V. Maier, “Visual, vocal and behavioural affordances: some effects of consistency,” in *5th International Conference on Cognitive Systems - CogSys 2012*, Vienna, 2012, p. 76.

[21] MiRo: The Biomimetic Robot. [Online]. Available: http://consequentialrobotics.com/miro/

[22] B. Mitchinson and T. J. Prescott, “MiRo: A robot mammal with a biomimetic brain-based control system,” in *Biomimetic and Biohybrid Systems. Living Machines 2016. Lecture Notes in Computer Science*, N. Lepora, A. Mura, M. Mangan, P. Verschure, M. Desmulliez, and T. Prescott, Eds. Springer, 2016, vol. 9793, pp. 179–191.

[23] E. C. Collins, T. J. Prescott, B. Mitchinson, and S. Conran, “MiRo: a versatile biomimetic edutainment robot,” in *Proceedings of the 12th International Conference on Advances in Computer Entertainment Technology - ACE ’15*. Iskandar, Malaysia: ACM Press, 2015, pp. 1–4.

[24] C. E. Osgood, W. H. May, and M. S. Miron, *Cross-Cultural Universals of Affective Meaning*. University of Illinois Press, 1975.

[25] E. C. Collins, T. J. Prescott, and B. Mitchinson, “Saying it with light: a pilot study of affective communication using the MiRo robot,” in *Proceedings of the 4th International Conference on Biomimetic and Biohybrid Systems - Volume 9222*. Barcelona, Spain: Springer-Verlag New York, Inc., 2015, pp. 243–255.

[26] R. K. Moore, “A real-time parametric general-purpose mammalian vocal synthesiser,” in *INTERSPEECH*, San Francisco, CA, 2016, pp. 2636–2640.

[27] R. K. Moore and B. Mitchinson, “A biomimetic vocalisation system for MiRo,” in *Living Machines 2017*, Stanford, CA, 2017. [Online]. Available: http://arxiv.org/abs/1705.05472