
Proceedings Paper:
Htet, A., Bernacka, K., Marei, O. et al. (2 more authors) (2025) Hey Miro! Multimodal interaction with an animal-like robot companion with conversational abilities. In: HRI '25: Proceedings of the 2025 20th ACM/IEEE International Conference on Human-Robot Interaction. 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 04-06 Mar 2025, Melbourne, Australia. Institute of Electrical and Electronics Engineers (IEEE) , pp. 1779-1781. ISBN 9798350378948

https://doi.org/10.1109/HRI61500.2025.10973974

Hey Miro! Multimodal Interaction with an Animal-like Robot Companion with Conversational Abilities

1st Aung Htet
Department of Computing
Sheffield Hallam University
Sheffield, United Kingdom
0009-0006-3883-7764

2nd Kinga Bernacka
Department of Computer Science
The University of Sheffield
Sheffield, United Kingdom
0009-0004-4401-6730

3rd Omar Marei
Department of Computer Science
The University of Sheffield
Sheffield, United Kingdom
0009-0002-4847-9463

4th Jake N. Holden
Department of Computer Science
The University of Sheffield
Sheffield, United Kingdom
0009-0001-5225-5107

5th Tony J. Prescott
Department of Computer Science
The University of Sheffield
Sheffield, United Kingdom
0000-0003-4927-5390

***Abstract*—To make best use of large-language models (LLMs) in social robotics it is critical that verbal interaction capabilities are suitably integrated with robot perceptual and behavioral systems, including non-verbal and emotional signaling, such that the robot can use language in a grounded and context-appropriate way. This demonstration shows the integration of a LLM with the layered control architecture of the animal-like robot platform *Miro-e* alongside deep network models for perception and spoken language recognition and generation. This system is currently being developed as a prototype companion robot for research on robot-assisted therapy.**

***Index Terms*—AI-enabled robotics; social HRI; robot companions; biologically-inspired robots; cognitive architecture**

I. Introduction

With the advent of AI large language models there is increased interest in the possibility of robot companions that we can talk to. Recent deep neural network models have also transformed the ability of robots to understand and act in the world through their perceptual and behavioral systems.

MiRo-e (Figure 1) is an animal-like robot developed by the University of Sheffield spin-out company *Consequential Robotics* and controlled by a brain-inspired control system [1]. The robot is currently used in universities and schools as a platform for robotics education and as a research tool for investigating biologically-inspired robotics and future applications of social robots in consumer electronics, health and social care.

Technology evaluation researchers have noted positive responses to MiRo-e in children similar to those seen with pet animals [2], [3]. Previous research with this platform has also identified that a language capability would be valued by older users of animal-like companion robots [4]. Here we describe a system that integrates the MiRo-e platform with off-board and cloud-based AI tools to allow for both verbal and non-verbal forms of human-robot interaction. In the remainder of this paper, we briefly summarize the capabilities of the base platform and control system and describe how this has been extended to add natural language and scene awareness capabilities that significantly enhance its interaction behavior. We conclude with a brief discussion of plans to extend this work towards applications in robot-based therapy.

Fig. 1. The MiRo-e robot developed by Consequential Robotics [1], [5].

II. About the MiRo-e robot

MiRo-e is a wheeled-robot with an animal-inspired head and body, built around a differential drive base and a three-degree-

The Wellcome Trust, through the "Imagining Technologies for Disability Futures" (ITDF) project no. 214963/Z/18/Z, and InnovateUK (grant no. 10039052) through the UK's funding guarantee scheme for the European Innovation Council Pathfinder project CAVAA (project no. 101071178).

of-freedom (DOF) neck, enabling lift, pitch, and yaw movements [1]. MiRo-e's expressive features include rotating ears, a drooping and wagging tail, embedded LED lighting, and eyelids that can open and close. These capabilities, combined with its life-like movements, allow MiRo-e to engage dynamically with its environment and users. MiRo-e interactivity is enhanced by 28 capacitive sensors across its body and head, that enable responses to touch gestures such as stroking and tapping. These interactions influence its emotional responses, expressed through the LED lighting, sound, and movement of the head and body and of cosmetic joints in the ears, tail, and eyelids. This combination of capabilities creates the potential for an engaging social robot suitable for various application scenarios. Miro-e has been widely exhibited and demonstrated including at HRI 2017 [5] and 2018 [6], the robot also received a best demonstration award at the HRI 2017 meeting.

A. MiRo-e Base System Architecture

The base version of the MiRo-e control architecture uses a layered design inspired by the vertebrate brain [1] that integrates behavioral patterns across space and time to coordinate multiple effector systems and to generate sequences of actions that satisfy drives and achieve goals. Higher-level systems can subsume lower-level operational loops in a manner inspired by the co-ordination between the forebrain, midbrain, hindbrain, and spinal cord of vertebrate animals. At the core of this architecture is a set of model midbrain/forebrain systems that generate integrated behavior through three key sub-systems:

1) A spatial behavior sub-system, modeled on the midbrain *superior colliculus*, which manages spatial orientation and attention. A critical feature of this system is the saliency map, dynamically updated with visual and auditory stimuli to highlight areas of interest, ensuring efficient spatial awareness and responsiveness.
2) An emotion sub-system, inspired by affective neuroscience, enabling MiRo-e to operate within a two-dimensional state of valence and arousal. Emotions are regulated by sensory systems and internal state and modulate robot behaviors allowing it to respond expressively during interactions.
3) An action selection mechanism, modeled on the vertebrate *basal ganglia* (a group of centralized brain structures), that selects and executes one behavioral plan at a time with persistence and pre-emption. This mechanism prioritizes actions such as "orient," "avert," "approach," and "flee," which are generated by simple hard-coded filters applied to the saliency map. These filters compute "where" as the map's maximum and "what" by integrating MiRo-e's affective state with stimulus attributes like size, location, or temporal characteristics.

These three components largely determine how MiRo-e responds to its surroundings, which includes approaching and interacting with moving or sound-generating objects (including people) when in a positive affective state. The base system also includes a biomimetic vocalization system that generates animal-like sounds that are affect-modulated.

B. Extended Architecture Integrating Language Processing

To enhance Miro-e's interaction capabilities, we have developed a new version of this control system that integrates spoken language processing alongside vision-based pattern and emotion recognition. The additions to Miro-e's control architecture are described below and illustrated in Figure 2. The new components are indicated as being "forebrain" systems since they support capacities that, in mammalian brains, involve the (forebrain) cerebral cortices.

1) Speech and Language Systems: MiRo-e's language processing pipeline begins with *Picovoice*'s *Cobra* and *Porcupine* APIs for Voice Activity Detection (VAD) and wake-word detection, respectively. Upon detecting the phrase "Hey, MiRo," the robot signals its attention by behaviors such as leaning-in, head tilting, and nodding. It then records audio until the VAD value drops below a threshold for approximately three seconds (suggesting that the interlocutor has stopped speaking). The audio is stored and transcribed using *Whisper*.

Transcribed speech is processed using OpenAI's *GPT-4o*, initialized with an engineered prompt to emulate an animal-like companion robot's personality, emotional state, and behavioral attributes. GPT-4o updates MiRo's internal state variables and generates responses, which are vocalized using *ElevenLabs* API. This API adjusts voice inflection based on punctuation and supports multiple languages. A higher-pitched synthesized voice was selected to be appropriate to the platform size and appearance. Predefined audio clips, such as hums or cries, can also trigger context-specific reactions such as when the robot's touch sensors detect interactions.

The robot repeats the listening and response process until no voice is detected within three seconds, at which point it resumes its explore behavior. During interactions, ChatGPT prompts are engineered to include emotional state information and object-detection flags derived from perceptual systems.

2) Vision and Touch Processing: MiRo's pipeline for scene awareness begins by classifying features in image frames captured by its HD cameras. This includes faces, emotional expressions, gestures, and objects. YOLO [7], a lightweight deep-neural-network object-classification model, is used to identify a limited set of objects which then act as conditioned signals, feeding into MiRo-e's action selection mechanism. In addition, MiRo-e incorporates spatial information and touch interactions into its decision-making, enabling behavior that is sensitive to human-robot physical interaction.

3) Behavior Generation: Robot behaviors can also be invoked directly via speech prompts, enabling complex actions such as spinning, imitation, and dancing. These behaviors also influence the robot's valence and arousal system, reinforcing its expressive nature. For example, the dancing behavior synchronizes with song tempo data retrieved from the *Spotify* API allowing MiRo-e to play and dance to music on request.

4) Action Selection, Motivation, and System Integration: The integration of the above novel components into the cognitive architecture is channeled through the basal ganglia-inspired action selection model (BG in Figure 2) and executed through coordinated motor actions (MPG - motor

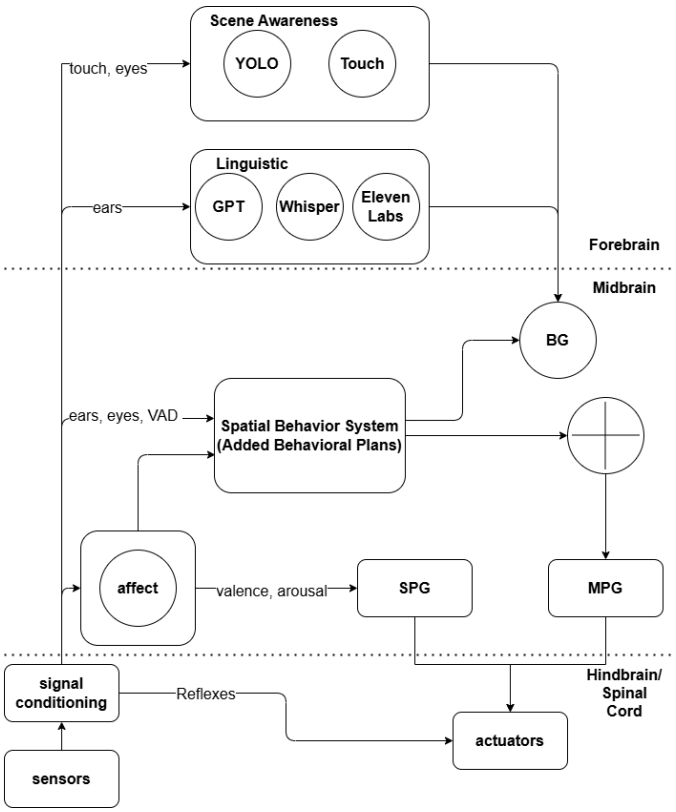

Fig. 2. The figure extends the existing MiRo-e architecture [1], all "forebrain" components are new, and have been added to complement and subsume the base architecture. See text for explanation and abbreviations.

pattern generation) and internal signaling (SPG - signal pattern generation) as described in [1]. The layered design ensures that MiRo-e can maintain goal-directed behavior while adapting dynamically to its environment, creating a robust and flexible interactive system. The full system design and code is open source and available at https://github.com/HeyMiro/Ext_Cog_Arch.

III. Interactive demo capabilities and limitations

As shown in the video figure, MiRo-e, equipped with this extended cognitive architecture, can engage in a rich array of interaction behaviors, and respond appropriately to spoken language, gesture, and physical touch. In quiet environments the speech system can use MiRo-e's built-in microphone array and loudspeaker, however, in noisy settings an external hyper cardioid microphone provides more robust speech detection and a Bluetooth speaker can provide sound amplification. The robot's in-built Raspberry Pi 3+ can support only limited processing beyond that required for the base model. Processing is therefore spread across a combination of the on-board processor, a local high-powered laptop (8Gb minimum GPU VRAM), and cloud-based processing for speech transcription (Whisper), LLM (ChatGPT), and text-to-speech (ElevenLabs). Cloud processing can lead to response delays. In good conditions response times during spoken verbal exchanges vary in the range 0.5-5s, but this can degrade to delays of up to 20s when Wi-Fi access is poor. 5G networks often provide the best option for cloud access in public settings. The current design uses verbal expression ("Hmmm") and a musical pattern to indicate that a response is being generated.

IV. Future Work

Ongoing work is focused on various improvements to the architecture including: (i) using a local CPU/GPU cluster in place of cloud processing to improve response latency (we are targeting below 0.5s delay) and enhance privacy, (ii) integration of active hearing to improve audio capture through built-in microphones, (iii) use of "Affect Control Theory" modulation to create more emotionally-aligned verbal responses (see [8]), (iv) improved scene awareness, (v) addition of episodic memory and virtualization to allow reasoning about past events and future social scenarios, (vi) increasing onboard compute power, and (vii) use of alternative LLMs including those specifically developed for robotics. In addition, we are actively working with technology evaluation partners to better understand potential applications for this system in a variety of healthcare and education settings including therapy use-cases where the robot could be employed to address mental health issues such as stress and anxiety [9].

Acknowledgment

The authors are grateful for ideas and assistance from Alex Lucas, Alejandro Jimenez, Ben Mitchinson, George Bridges, Julie Robillard, Uni. of Sheffield students on the Cognitive and Biomimetics Robotics module, members of the ITDF and CAVAA consortiums, and Consequential Robotics Ltd.

References

[1] Mitchinson, B., Prescott, T.J. (2016). "MIRO: A robot "mammal" with a biomimetic brain-based control system". In: Lepora, N. et al. (eds) Biomimetic and Biohybrid Systems. LNAI vol 9793. Springer, Cham. doi: 10.1007/978-3-319-42417-0_17.

[2] Barber, O., Somogyi, E., McBride, A.E., Proops, L. (2020). "Children's evaluations of a therapy dog and biomimetic robot: Influences of animistic beliefs and social interaction", Int. J. of Social Robotics, 13(6), pp. 1411-1425. doi: 10.1007/s12369-020-00722-0.

[3] Dosso, J.A., Kailley, J.N., Martin, S.E., Robillard, J.M.: "A Safe Space for Sharing Feelings: Perspectives of Children with Lived Experiences of Anxiety on Social Robots", Multimodal Technologies and Interaction, 2023, 7, (12), doi: 10.3390/mti7120118.

[4] Dosso, J.A., Kailley, J.N., Guerra, G.K., Robillard, J.M.: 'Older adult perspectives on emotion and stigma in social robots', Frontiers in Psychiatry, 2023, 13. doi: 10.3389/fpsyt.2022.1051750.

[5] Prescott, T.J., Mitchinson, B., Conran. S.(2017). "MiRo: An Animal-like Companion Robot with a Biomimetic Brain-based Control System." Proc. ACM/IEEE Int. Conf. on Human-Robot Interaction, 50-51. Vienna, Austria: ACM. doi: 10.1145/3029798.303666.

[6] Prescott, T.J. *et al*. 2018. "MiRo: Social interaction and cognition in an animal-like companion robot." In Companion of the 2018 ACM/IEEE Int. Conf. on Human-Robot Interaction, 41. Chicago, IL, USA: ACM. doi: 10.1145/3173386.3177844.

[7] S. H. Vemprala, R. Bonatti, A. Bucker, A. Kapoor, "ChatGPT for robotics: design principles and model abilities", IEEE Access, vol. 12, pp. 55682-55696, 2024, doi: 10.1109/ACCESS.2024.3387941.

[8] Robillard, J.M., Hoey, J. (2018). "Emotion and Motivation in Cognitive Assistive Technologies for Dementia", Computer, 51(3), pp. 24-34. doi: 10.1109/mc.2018.1731059.

[9] Kabacińska, K., Prescott, T.J., Robillard, J.M. (2020). "Socially Assistive Robots as Mental Health Interventions for Children: A Scoping Review", Int. J. of Social Robotics, 2020, 13(5), pp. 919-935, doi: 10.1007/s12369-020-00679-0.