Natural face-to-face conversation with socially intelligent robots

Mary Ellen Foster
School of Computing Science, University of Glasgow
18 Lilybank Gardens, Glasgow G12 8RZ, United Kingdom
MaryEllen.Foster@glasgow.ac.uk

Abstract

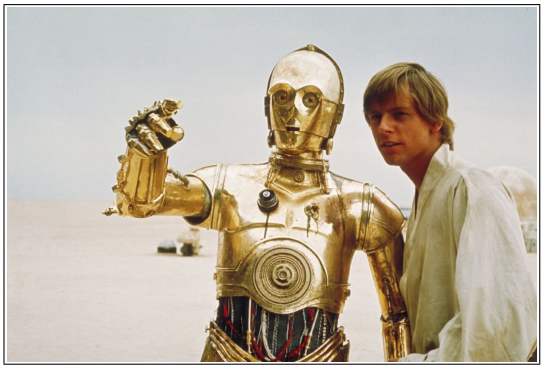
When humans engage in face-to-face conversation, they use their voices, faces, and bodies together in a rich, multimodal, continuous, interactive process. For a robot to participate fully in this sort of natural, face-to-face conversation in the real world, it must also be able not only to understand the multimodal communicative signals of its human partners, but also to produce understandable, appropriate, and natural communicative signals in response. A robot capable of this form of interaction can be used in a large number of areas: for example, it could take the role of a home companion, a museum tour guide, a tutor, or a personal health coach. While a number of such robots have been successfully deployed, the full potential of socially interactive robots has not been realised, due both to incomplete models of human multimodal communication and to technical limitations. However, thanks to recent developments in a number of areas— including techniques for data-driven interaction models, methods of evaluating interactive robots in real-world contexts, and off-the-shelf component technology—the goal of developing a naturally interactive robot is now increasingly achievable.

1. Natural face-to-face conversation ...

Face-to-face conversation can be seen as both the most basic and the richest form of human interaction. As Bavelas, Hutchinson, Kenwood *et al.* [1] point out, face-to-face conversation has three main features that other forms of communication do not. First, it permits unrestricted **verbal** expression: in principle, all participants in a conversation can speak entirely freely, unless the social situation constrains the details of the language that can be used in practice. Secondly, in a face-to-face setting, the participants have access to all of the **non-verbal** elements available in human communication, including facial displays, body language, gestures; these elements can provide redundant information to help with the interpretation of the verbal content, and can also permit participants to convey information that is difficult to communicate with words alone. Finally, face-to-face conversation is inherently a two-way process that allows for instantaneous **collaboration** among the participants: speakers may provide continuous verbal and non-verbal "back-channel" feedback during the conversation, and may even produce overlapping speech and gestures. Indeed, as has long been noted by Herb Clark and colleagues [e.g., 2], conversation is inherently a form of collaborative joint action where the participants work together to confirm that everything is mutually understood, through verbal actions such as confirmation questions as well as non-verbal actions such as gazing directly at an interlocutor.

For these reasons, Bavelas *et al.* suggest that face-to-face conversation should be used as a prototype for other communication systems; that is, that other forms of communication should be considered based on how they differ from face-to-face conversation on one or more of the above dimensions. This in turn implies that for artificial systems, the richest and most usable form of interaction is one that mimics face-to-face conversation as closely as possible.

The metaphor of face-to-face conversation has been applied to human-computer interface design for some time. However, for the most part, the influence has been simply at the level of metaphor [3]: artificial systems have not been developed with the idea that their users would literally engage them in conversation, but rather that the systems would incorporate concepts from natural conversation to inform the design of user interfaces [e.g., 4]. However, as interface technology has become increasingly sophisticated, it has become possible to incorporate more and more of the advantages of face-to-face conversation much more directly into human-computer interfaces, whether at the level of speech-based dialogue systems [5], embodied conversational agents [6], or even—as we will discuss in the remainder of this paper—socially intelligent robots [7].



Still of Mark Hamill in Star Wars: Episode IV - A New Hope (1977); image from `http://imdb.com`

Figure 1: A socially intelligent robot

2. . . . with socially intelligent robots

In 2003, Fong, Nourbakhsh and Dautenhahn [7] prepared a comprehensive survey of socially interactive robots, which they defined as "robots for which social interaction plays a key role." They focussed particularly on peer-to-peer human-robot interaction; that is, where the robot exhibits human-like social skills such as expressing and/or perceiving emotions, using natural non-verbal cues like gaze and gestures, and possibly learning or developing various social competencies. They discussed a number of issues relevant to the then-developing field: approaches and challenges in the design of such robots; issues related to the embodiment of a robot (physical shape, caricatured vs. realistic appearance, etc.); the role of emotion, personality, and user modelling in social robotics; along with approaches to human-oriented perception and learning.

In the decade since that survey was published, work in social robotics—and in human-robot interaction (HRI) generally—has advanced considerably: the field now includes two dedicated journals [8], [9] and three conference series [10]–[12], along with numerous other discussion and venues, confirming that it is an active and growing area of research. Interactive robots have been deployed and evaluated in a wide range of situations; for example, the list of Best Papers from the recent HRI 2015 conference [10] includes robots designed to be an empathy-invoking conversational companion [13], a helper in a cooperative physical task [14], a mediator for remote negotiation teams [15], and an autonomous mobile helper in a shopping mall [16]. Techniques for evaluating interactive robots have also moved forwards considerably: the Godspeed questionnaire series [17] has been widely used and validated as an instrument for user studies in HRI, while another of the HRI 2015 best papers described a large-scale study examining the moral norms for robot agents [18].

3. Building blocks for the future

Developing a robot able to engage in face-to-face conversation has been a goal of the robotics field for some time; indeed, in popular culture, the prototypical model of a robot is of a natural conversational partner (e.g., Figure 1). However, while (as described above) there has been a great deal of progress in this field, even the most advanced interactive robots are able to operate only in constrained environments, support a limited set of communicative behaviours, and are generally designed to operate in very specific interaction contexts. Also, any given robot system is usually tightly tied to its initial development context: developing a robot able to support a new interactive context has generally meant starting practically from scratch, as the hardware and software that is suitable for one setting often cannot easily be translated to a different one.

As the field moves from the current state of the art towards developing interactive robots that support natural, real-world face-to-face conversation, several factors must be taken into account. Hartholt, Traum, Marsella *et al.* [19] discussed the factors required for intelligent embodied agents to have wider applicability, which include the following: the agents must be able to reproduce human abilities in perception, reasoning, and behaviour generation; these abilities must be combined into a larger overall system; and the agents must be general and reusable, in a range of interactive situations.

Until now, the development of naturally interactive social robots has been limited by two main factors that have together imposed practical limitations on the range of interactive behaviours for such robots. On the one hand, the mechanisms underlying natural human embodied communication remain imperfectly understood, making it difficult to implement high-quality models for the artificial agents. And even where theoretical, descriptive models exist, they are often not in a state where they can be directly implemented on an artificial agent. On the other hand, the current techniques in input processing, interaction management, and output generation have also not been able to support the full range of communicative behaviours observed in natural communication, either because they cannot represent the necessary level of detail, or else because they simply cannot recognise or reproduce the behaviours at all.

Thanks to recent developments, it is now becoming possible to address all of the above factors in the course of developing robust, interactive robot agents. These developments fall into three main areas: novel techniques for data-driven interaction models, increasingly sophisticated off-the-shelf components, and more effective strategies for real-world user evaluation.

3.1. Data-driven interaction models

The earliest versions of multimodal interactive systems generally used expert-written hand-coded rules to drive their behaviour. While such hand-coded rules can work well in constrained settings, they do not scale well to more complex, real-world interactions. Also, defining the rules requires expert knowledge and high-quality models of the intended behaviour: so in areas such as fully embodied face-to-face communication, where such models do not generally exist, this also means that rules can be prohbitively difficult to develop for all but the most constrained interactive situations.

More recently, researchers in multimodal interaction have begun to explore a range of data-driven techniques for developing interaction models [20]. All of these techniques make use of recorded data from real interactions, in both human-human and human-machine contexts, but the way that they employ the data varies: for example, some researchers have had success training models directly on the data [21], [22], while others have created data-driven simulated users for use in techniques such as reinforcement learning [23]. Most existing research in this area to this point has required the use of annotated training data, which can be difficult to obtain and process. However, novel machine learning techniques such as deep learning [24]—in which representations are learned directly from the raw data—have made it possible to develop models for a wide range of applications directly from raw recorded data, with little need for annotation or feature selection.

Applying similar techniques to the problem currently under consideration can go a long way to alleviate the issues arising from the lack of fully operational models: we could begin by implementing a high-level description from the literature, and then fine-tune the details of the model through data-driven learning. Particularly if we apply the ideas from deep learning, the necessary initial models of multimodal behaviour could possibly be learned directly from un-annotated recorded interactions, using pre-existing multimodal corpora [25] or, where necessary, newly recorded data. The models would then be fine-tuned through user evaluation as described in Section 3.3.

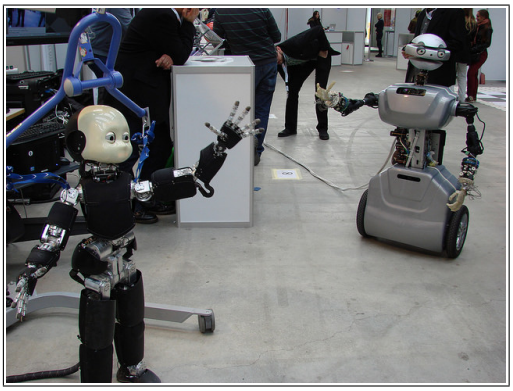
3.2. Off-the-shelf components

Previously, developing a socially interactive robot would have required significant work in a number of basic technical areas. For example, detecting and processing the multimodal social signals of a human partner would require development of technologies such as computer vision and auditory sensors, while generating the multimodal agent behaviour would require contributions from the very basic low-level areas of robotics such as motor control and path planning, as well as from fields such as computer animation to ensure that the generated behaviour is sufficiently expressive. However, recent developments in component technologies have made it possible to develop robust robots that can be deployed into a variety of real-world contexts, using off-the-shelf technology as building blocks. For example, in the area of input processing, useful components include audiovisual sensors such as the Microsoft Kinect [26], face processing libraries such as OKAO [27], along with physiological sensors such as the E3 wristband [28]. When it comes to embodied agent platforms, libraries such as the Virtual Agent Toolkit [29] have made it possible to create high-quality animated characters, while reasonably-priced, commercially available robots such as those shown in Figure 2 offer the same opportunity for physically embodied agents.

3.3. User evaluation techniques

As described in Section 3.1, we propose that a solution to the modelling problem lies in learning from data. In the context of systems that are built using data-driven techniques, a popular form of evaluation involves comparing the system behaviour to that recorded in the underlying data. However, while such techniques are useful as a "sanity check" during development, the danger in relying purely on data-driven measures in an interactive setting is that they are known to penalise output that differs in any way from the exact examples in the data. This in turn tends to favour "average" behaviour that does not make use of the full system capabilities, and that is often not preferred by actual users in practice [30], [31].

This means that the developed agents and their models must be evaluated through interactions with real users, which can be a difficult and time-consuming task. However, the development of crowdsourcing platforms such as Amazon Mechanical Turk and Crowdflower have made it possible to carry out real-user evaluations on a significantly larger scale than ever before [32], and the results of such studies have been found to correlate well with those from lab-based studies [33]. This technique has even been applied to the evaluation of interactive robots, using techniques such as robot simulators and prerecorded videos [34], [35]. Although ultimately it is necessary to test robot agents with actual humans in the real world, the addition of crowdsourcing to the process can greatly help with evaluating and improving the interaction models during the development process.

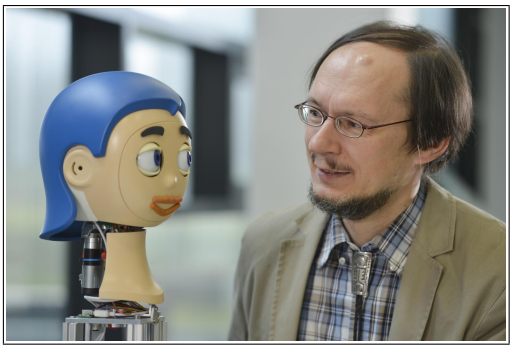


(a) iCub and Flash (Image by **flash.wrut** on Flickr)

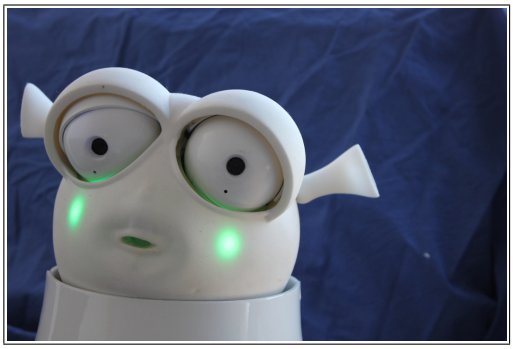
(b) Jibo (Image from `http://jibo.com`)

(c) Nao and Pepper (Image by Aldebaran)

(d) REEM-C (Image by PAL Robotics for Wikipedia)



(e) Flobi (Image by CITEC/Bielefeld University)



(f) Reeti (Image from `http://reeti.fr`)

Figure 2: A selection of modern interactive robots

4. Potential application domains

A robot able to learn to interact with humans in a natural, face-to-face setting has a wide range of potential applications, particularly in contexts where trained, skilled target users are not expected: for example, in short-term, public-space interactions, or in contexts where the robot should interact with users such as children or elderly people. The range of possible application domains includes the following:

Virtual receptionists and tour guides These agents monitor and respond to the behaviour of participants in a dynamically changing, multi-party situation, where only some of the participants need attention at any time, and where the type of response required varies widely.

Robots acting as lab demonstrators or tutors This scenario combines multi-party social interaction with task-based behaviour in the physical world. It is also an area where robots can have a real-world impact, for example by performing repetitive or dangerous demonstrations in areas such as physics or chemistry.

Assistive robots in the home environment An assistive robot must understand the needs of the user, whether explicitly or implicitly stated, and to carry out tasks in the physical world (e.g., retrieving objects or switching devices on or off). Crucially, it must also be able to do all of this while ensuring that the user understands what the robot is doing and why it is doing it.

Companion robots for the elderly A robot in this context must understand and respond with a wide range of multimodal social behaviours in order to develop an appropriate relationship with its human partner.

As the recent advances described above are applied to the active research area of social robotics, more and more robots will be able to be trained and deployed into such domains, allowing untrained real-world users to interact with and benefit from the presence and assistance of such robots.

References

[1] J. B. Bavelas, S. Hutchinson, C. Kenwood and D. H. Matheson, 'Using face-to-face dialogue as a standard for other communication systems', *Canadian Journal of Communication*, vol. 22, no. 1, 1997. [Online]. Available: http://www.cjc-online.ca/index.php/journal/article/view/973.

[2] H. H. Clark, *Using language*. Cambridge: Cambridge University Press, 1996.

[3] J. Cassell, 'Nudge, nudge, wink, wink: Elements of face-to-face conversation for embodied conversational agents', in, [6], pp. 1–27.

[4] A. M. Bueno and S. D. J. Barbosa, 'Using an interaction-as-conversation diagram as a glue language for HCI design patterns on the web', in *Task Models and Diagrams for Users Interface Design*, K. Coninx, K. Luyten and K. A. Schneider, Eds. Springer Berlin Heidelberg, 2007, pp. 122–136. DOI: 10.1007/978-3-540-70816-2_10.

[5] K. Jokinen and M. McTear, *Spoken Dialogue Systems*, 1. Morgan & Claypool, 2009, vol. 2. DOI: 10.2200/S00204ED1V01Y200910HLT005.

[6] J. Cassell, J. Sullivan, S. Prevost and E. Churchill, Eds., *Embodied Conversational Agents*. MIT Press, 2000.

[7] T. Fong, I. Nourbakhsh and K. Dautenhahn, 'A survey of socially interactive robots', *Robotics and Autonomous Systems*, vol. 42, no. 3–4, pp. 143–166, 2003. DOI: 10.1016/S0921-8890(02)00372-X.

[8] S. Kiesler, Ed., *Journal of Human-Robot Interaction*. [Online]. Available: http://humanrobotinteraction.org/journal/index.php/HRI/index.

[9] S. S. Ge and O. Khatib, Eds., *International Journal of Social Robotics*. Springer. [Online]. Available: http://www.springer.com/engineering/robotics/journal/12369.

[10] J. A. Adams and W. Smart, Eds., *HRI 2015: 10th IEEE/HRI Conference on Human-Robot Interaction*, (2nd–5th Mar. 2015). [Online]. Available: http://humanrobotinteraction.org/2015/.

[11] Y. Nakauchi, Ed., *IEEE RO-MAN 2015: The 24th International Symposium on Robot and Human Interactive Communication*, (31st Aug.–4th Sep. 2015). [Online]. Available: http://ro-man2015.org/.

[12] A. Tapus and M. Vincze, Eds., *ICSR 2015: Seventh International Conference on Social Robotics*, (26th–30th Oct. 2015). [Online]. Available: http://www.icsoro.org/icsr2015/.

[13] G. Hoffman, O. Zuckerman, G. Hirschberger, M. Luria and T. Shani Sherman, 'Design and evaluation of a peripheral robotic conversation companion', in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, Portland, Oregon, USA, 2015, pp. 3–10. DOI: 10.1145/2696454.2696495.

[14] S. Nikolaidis, R. Ramakrishnan, K. Gu and J. Shah, 'Efficient model learning from joint-action demonstrations for human-robot collaborative tasks', in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, Portland, Oregon, USA, 2015, pp. 189–196. DOI: 10.1145/2696454.2696455.

[15] C. Bevan and D. Stanton Fraser, 'Shaking hands and cooperation in tele-present human-robot negotiation', in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, Portland, Oregon, USA, 2015, pp. 247–254. DOI: 10.1145/2696454.2696490.

[16] D. Brscić, H. Kidokoro, Y. Suehiro and T. Kanda, 'Escaping from children's abuse of social robots', in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, Portland, Oregon, USA, 2015, pp. 59–66. DOI: 10.1145/2696454.2696468.

[17] C. Bartneck, D. Kulić, E. Croft and S. Zoghbi, 'Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots', *International Journal of Social Robotics*, vol. 1, pp. 71–81, 1 2009. DOI: 10.1007/s12369-008-0001-3.

[18] B. F. Malle, M. Scheutz and T. A. and John Voiklis and Corey Cusimano, 'Sacrifice one for the good of many?: People apply different moral norms to human and robot agents', in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, Portland, Oregon, USA, 2015, pp. 117–124. DOI: 10.1145/2696454.2696458.

[19] A. Hartholt, D. Traum, S. Marsella, A. Shapiro, G. Stratou, A. Leuski, L.-P. Morency and J. Gratch, 'All together now', in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, vol. 8108, Springer Berlin Heidelberg, 2013, pp. 368–381. DOI: 10.1007/978-3-642-40415-3_33.

[20] O. Lemon and O. Pietquin, Eds., *Data-Driven Methods for Adaptive Spoken Dialogue Systems*. Springer New York, 2012. DOI: 10.1007/978-1-4614-4803-7.

[21] D. Bohus and E. Horvitz, 'Learning to predict engagement with a spoken dialog system in open-world settings', in *Proceedings of SIGDIAL 2009*, London, United Kingdom, 2009, pp. 244–252.

[22] N. Dethlefs and H. Cuayáhuitl, 'Hierarchical reinforcement learning for situated natural language generation', *Natural Language Engineering*, vol. FirstView, pp. 1–45, Jun. 2014. DOI: 10.1017/S1351324913000375.

[23] S. Keizer, S. Rossignol, S. Chandramohan and O. Pietquin, 'User simulation in the development of statistical spoken dialogue systems', in, [20], pp. 39–73. DOI: 10.1007/978-1-4614-4803-7_4.

[24] Y. Bengio, A. Courville and P. Vincent, 'Representation learning: A review and new perspectives', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013. DOI: 10.1109/TPAMI.2013.50.

[25] P. Paggio, D. Heylen and M. Kipp, Eds., *Journal on Multimodal User Interfaces*, vol. 7, 1–2, 2013, *Special Issue: Multimodal Corpora*, *Special Issue: Multimodal Corpora*. [Online]. Available: http://link.springer.com/journal/12193/7/1/.

[26] Kinect for Windows Team, 'The Kinect for Windows v2 sensor and free SDK 2.0 public preview are here', *Kinect for Windows Blog*, 15th Jul. 2014. [Online]. Available: http://blogs.msdn.com/b/kinectforwindows/archive/2014/07/15/the-kinect-for-windows-v2-sensor-and-free-sdk-preview-are-here.aspx (visited on 10/08/2015).

[27] OMRON Corporation. (2015). Image Sensing Technology / Products / OMRON Electronic Components Web, [Online]. Available: http://www.omron.com/ecb/products/mobile/ (visited on 10/08/2015).

[28] Empatica S.r.L. (2015). Monitor real-time physiological signals with Empatica's EDA sensor / GSR sensor - E4 wristband, [Online]. Available: https://www.empatica.com/e4-wristband (visited on 10/08/2015).

[29] J. Gratch, A. Hartholt, M. Dehghani and S. Marsella, 'Virtual humans: A new toolkit for cognitive science research', *Applied Artificial Intelligence*, vol. 19, pp. 215–233, 2013.

[30] M. E. Foster and J. Oberlander, 'User preferences can drive facial expressions: Evaluating an embodied conversational agent in a recommender dialogue system', *User Modeling and User-Adapted Interaction*, vol. 20, no. 4, pp. 341–381, Oct. 2010. DOI: 10.1007/s11257-010-9080-6.

[31] A. Belz and E. Reiter, 'Comparing automatic and human evaluation of NLG systems', in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, 2006.

[32] A. Kittur, E. H. Chi and B. Suh, 'Crowdsourcing user studies with Mechanical Turk', in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2008, pp. 453–456. DOI: 10.1145/1357054.1357127.

[33] A. Koller, K. Striegnitz, D. Byron, J. Cassell, R. Dale, S. Dalzel-Job, J. Oberlander and J. Moore, 'Validating the web-based evaluation of NLG systems', in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 2009, pp. 301–304.

[34] C. Breazeal, N. DePalma, J. Orkin, S. Chernova and M. Jung, 'Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment', *Journal of Human-Robot Interaction*, vol. 2, no. 1, pp. 82–111, 2013.

[35] R. Toris, D. Kent and S. Chernova, 'The robot management system: A framework for conducting human-robot interaction studies through crowdsourcing', *Journal of Human-Robot Interaction*, vol. 3, no. 2, pp. 25–49, 2014. DOI: 10.5898/JHRI%2F3.2.Toris.