FAIM/ISCA Workshop on Artificial Intelligence for Multimodal Human Robot Interaction
14-15 July 2018, Stockholm, Sweden

Using Multimodal Information to Support Spoken Dialogue Interaction between Humans and Robots without Intrusive Language processing

Nick Campbell
Speech Communication Lab
Trinity College Dublin, The University of Dublin
Ireland
nick@tcd.ie

Abstract

This position paper expounds our recent views on autonomous spoken dialogue processing without (or with only minimal use of) automatic speech recognition (ASR). It argues that autonomous systems can be trained to read non-verbal signals which facilitate transition through a pre-prepared or stored utterance sequence so that only minimal processing of actual spoken content is needed. Of course no system, machine or human, will be able to continue an extended conversation without understanding the meaning, but we claim that it is not necessary to process each and every spoken word in order to satisfactorily complete an everyday spoken interaction.

Keywords: Autonomous Dialogue Systems, Non-verbal Signals, Engagement Sensing, Speech Processing, Social Interaction

1. Companion Robots & Call-Centres

There are many situations where robots or machines talk to humans in everyday contexts. Companion Robotics (CR) [1] is a growing area of research and the constrained but targeted social dialogues that are needed in such applications can be considered very similar in some ways to the Call-Centre dialogues that are now extensively studied and understood [2]. While the latter still (we hope) use human interlocutors whose permissible utterances are constrained by strict branding and corporate discipline, the former are becoming more autonomous and increasingly personal. While the dialogue structure is very similar, the content of CR dialogues can be much wider in scope. Recently, considerable concern has been expressed about the intrusivity of such personal conversational systems and the fear that they might 'leak' information from sensitive personal environments. Accordingly, we have been exploring ways of conducting a dialogue without understanding too much about what is being said. This may seem strange, but perhaps humans do it very often.

1.1. Scripted Interactions and Autonomous Conversational Agents

Chatbot dialogues have now become almost indistinguishable from their human equivalents [3] but there are still a large number of situations where the context constrains the content and semi-scripted interactions are sufficient. For autonomous conversational agents to successfully interact with people in everyday domestic situations it is helpful if they can maintain dominance in a conversation, and thereby have a more reliable understanding of the dialogue moves [4]. Once the human takes the initiative, the machine has to do some very complex processing to model the pragmatics of the interaction and follow the strands of intention. However, in many situations, the dialogue is constrained by context and the machine can effectively take the lead. This paper explores conversational devices whereby this is made possible, and assumes that many of these interactions follow the Call-Centre model as explained below.

1.2. On-task and Off-task Dialogues

Recent work from our lab has stressed the importance of 'off-task' dialogues for managing interpersonal relations and conducting social interactivity. As humans, we need to socialise, even during a call-centre interaction, and frequently tend to stray from the strict sequence of programmed interactions that are set by the task. A sentient autonomous dialogue agent therefore needs to be able to distinguish 'on-task' talk from 'off-task' chat and to be able to handle and switch between both efficiently and discreetly. Whether the interlocutor is 'merely' a customer, or 'simply' a companion, the agent must follow social norms and avoid potentially offensive behaviour. Chat interludes must be handled casually and the dialogue (if possible) soon brought back on task. Task-based interactions should also perhaps be 'lightened' by occasional chat-like social interludes.

The position that we are taking in this paper is based on our experience from working with three discrete research projects, the first *Herme* [5] used a LEGO device in spontaneous interaction with subjects visiting off-the-street at the Science Gallery in Dublin. The second was at a recent eNTERFACE workshop at UTwente where we took part in the design of HM^3 [6], a multimodal multiparty receptionist robot, and the third as part of the recently completed EU CHIST-ERA *Joker* project [7], using 'Miro' [8] with the 'Cara' dialogue system [9] to interact socially with people as if in a CR situation.

Our first conversational agent 'Herme' had no ASR component but apparently most people didn't notice, and they chatted happily with the robot for up to 5 minutes when she jokingly terminated the conversation. 'HM^3' did have limited speech recognition but relied primarily on multimodal information to progress through the situation, directing and dismissing customers according to their needs. With Miro, we focussed more on the biomimetic component to make sense of non-verbal cues and relate them to the progress of spoken dialogue, and it is this multimodal aspect which inspired us to consider the need for 'ethical social robotics'

10.21437/AI-MHRI.2018-2

Figure 1: The robot Miro - from Consequential Robotics – Attentive, observant, and biomimetically programmed.

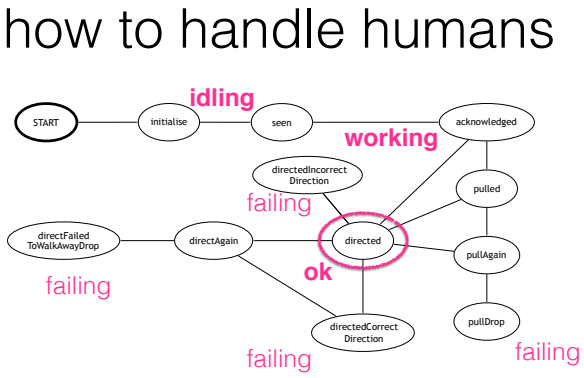

Figure 2: Idling versus serving a customer – the sequence of tasks for a receptionist robot prepared for failure

more deeply. If the robot can process enough non-verbal information to progress through a task-based or chat-based conversation, then there will be less need for potentially intrusive recording or transcription of the spoken component, and the conversation can be 'ethically' guaranteed, its privacy assured.

2. Modes of Dialogue Interaction

People also talk to dogs, and cats, and maybe even goldfish! They may not expect the animal to understand, but dogs in particular are very effective in letting people think that they do. Mankind's relationship with domestic animals goes back to pre-history and their co-existence has proved to be mutually beneficial. Some argue that our relationship with dogs should be the model for future human-robot interaction [10]. These animals have learnt to 'read our signals' and to process human non-verbal information with remarkable accuracy. Some dogs have actually been shown to have a working vocabulary for differentiating over a hundred spoken words [11], but their prime signal is the non-verbal; visual information combined with expressive tone-of-voice. Robots can be programmed to read similar information and thereby perhaps even to develop similar relationships with their companion humans. If they are good enough at reading the non-verbals, then they may be able to do much less processing of the linguistic content, and thereby become much safer machines to live with.

2.1. Multimodal Dialogue Management

In many common situations, a dialogue can be thought of as a pre-ordained sequence of moves that are enacted through speech [4]. In the Call-Centre example, after a brief social introductory phase, the problem is discerned, a solution found, and often a contract sold, then the process is socially terminated. The Herme dialogues followed a similar pattern; initiating a bond, exchanging personal information, telling and eliciting a joke, and then politely closing and dismissing the interlocutor. Each step was carried out through triads of brief utterances that were designed to elicit an appropriate response from the visitor to the Gallery. Laughter was a novel but lubricating component in these dialogues. The robot was principally only aware of facial presence, articulation gestures (both audio and visual) and length of utterance - longer talk from the visitor was greeted with appropriate back-channel responses, and briefer exchanges were taken as back-channel or acknowledgement. No speech processing occurred and the robot didn't listen to the speech of the visitor, although many acted as if they thought she did [12].

Similarly for the receptionist robot; a person appearing on the scene was assumed to have a goal in mind, and the job of the robot was to direct them accordingly (to a doctor A or B, or by giving over keys to Room A or B). The interesting complication to this situation is the queueing and combined handling of multiple customers on the scene. Here a social element or subroutine becomes essential, as in the 'off-task' behaviour mentioned above.

The biomimetic engine that is part of the Consequential Robots 'Miro' machine provided us with the processing capacity that we needed to continue in this line of research; the robot includes one processing board expressly devoted to spatial attention, forming a continuous impression of salient events in the surroundings, and it is programmed to respond to them according to its emotional states (short and long-term).

Miro's Basal Ganglia model (BG) [13] acts as a centre for action selection with hysteresis and persistence. Action sub-systems compute a priority (between zero and one), and select, at each time step, which action sub-system is disinhibited, and thereby able to take control of the robot's kinematic chain. We were able to tap into this processing stream, adding speech as a sub-system, and use it to help making inferences about the state of the interlocutor at each stage of the conversation for the selection of an appropriate utterance.

2.2. Taking the Lead, Monitoring the Effect

The art of good CR conversation design is to ensure that the robot is at all times taking the initiative in the conversation and that if it ever loses that initiative, then it has repair routines at hand to regain it.

The elicitation sequence "My name is Herme", "What's your name?" ensures that the interlocutor responds either as expected (with their name which the robot may either store or ignore) or with a sign indicating uncertainty of response.

"Jack" and "Eh??" are significantly different responses in

this case, while ”What?” and Uh?” can be considered equivalent, as are ””Mary” and ”St John the Baptist”, the former pair indicating that a repair is needed, and the latter being accepted as appropriate indicators that the conversation is still ‘on-track’.

The role of the robot is to serve appropriate utterances while comparing their reaction to a small set of nonverbal response template behaviours. It is a behavioural expectation, not a linguistic process that enables the conversation to continue.

2.3. Chunking the Interaction

Robot speech can be difficult to follow, so shorter chunks with plenty of interaction are better than longer ‘instructive’ sentences when rendered in synthesised speech.: ”Henry, did you remember to take your medicine?” would perhaps be better rendered as ”Henry?”, – pause – ”Your medicine . . . ”, – pause – ”Did you remember?” The verb ‘take’ is probably unnecessary, being too explicit for this style of interaction. The pauses between the simple chunks are required though, to ensure that Henry (or whoever the robot is talking to) has a chance to acknowledge the utterance and signal understanding, and provide the robot a chance to sense their continued engagement.

Herrme’s utterances were in triads of this form [12] and I take the present opportunity to suggest some changes to the HM^3 dialogue snippets reproduced in [12] so that they conform to the above pattern. Figure 2 shows the simple case of a robot interacting with a single customer. It is clear from the figure that there are many opportunities for failure even in this simple primitive directing operation.

The dialogue moves in this case are: Acknowledge – Greet – Instruct – InstructAgain – DismissAfterInstruct – Direct – DirectAgain – DismissAfterDirect – DirectAfterIncorrectDirection – Farewell. The figure shows how and in which order they may be used. The xxx-Again operations are included to ensure that the robot has a chance to repeat (up to twice) before giving up on the task. The Dismiss operations enable the robot to fail gracefully by acknowledging its failure to complete the conversation (or task).

All utterances can be pre-stored in an interaction template. The task of the robot is to sense the reactions of the customer, given its knowledge of the expected non-verbal resonse behaviour (which can be coded with each utterance in context), in order to select the most appropriate next utterance from the store.

Some original HM^3 utterances are reproduced here (note their complexity): *Greet:* “Hello, my name is Zeno.” *Instruct:* “Please point at the sign with your doctor’s name on it. Either this sign (1) on your left, or this sign (2) on your right.” *InstructAgain:* “My apologies, maybe I was not clear. Please point at the sign with your doctor’s name on it. Either this sign (1) on your left, or this sign (2) on your right.” *DismissAfterInstruct:* (1) “I’m sorry, I’m not able to help you out. My capabilities are still limited, so I was not able to understand you.” (2) “Please find a nearby human for further assistance.” *Farewell:* “Please go to the left/right for doctor Vanessa/Dirk.”

If the room is noisy, or the speech synthesis indistinct, then these long and explicit ‘sentences’ are likely to be misheard

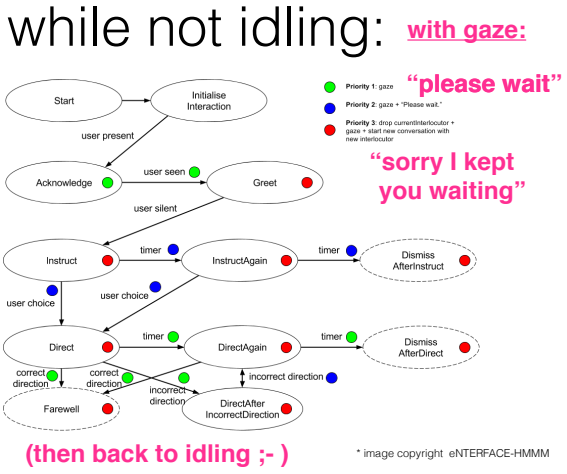

Figure 3: Multimodal handling of more than one customer for the receptionist robot. Note the constant use of gaze.

or misunderstood. They are ‘engineer-speak’, not representative of actual human-human conversational interaction. To break them down into pragmatic triads, ripe for multimodal reaction sensing, we might use something like the following:

Greet: “Hi!.”, “I’m Zeno.”, “Hello.”, or “Who’s next?”.

Instruct: “Ermm,”, “Who’s your doctor?”, “do you see the name?”, with a repair subroutine for the potential failure of communication ready: “See the signs?”, “Which one?”

InstructAgain: “Oops”, “sorry”, “Let’s try again” . . . and then once contact has been re-established, the previous subroutine (or one of its equivalents) can be re-used.

DismissAfterInstruct: (1) “Oh dear,”, “I’m not getting through . . .”, “sorry” (2) “Can you ask someone else?”

Farewell: “Thanks,”, “Take care.”, “See you again soon”.

2.4. Managing interruptions

Figure 3 shows the complicated case of more than one customer being present. This is closer to the real-world situation where people must be acknowledged and managed but not necessarily immediately. Social handling becomes essential. The green and blue circles mark the importance of ‘gaze’ in this process, which is otherwise the same as that shown in Figure 2. Red circles mark failure points. Idling is the default activity.

When handling multiple interlocutors, the eyes play an even more important role, and face-detection at a distance becomes crucial to the spoken interaction. Herme was able to detect the engagement of her interlocutors by watching them (instead of listening), and handled groups of people with the aid of a large-screen display placed behind the robot, with coloured circles round their faces on the display, showing off her ability to zoom in on the closest (largest) face by moving according to simple OpenCV [17] operations. By showing that she could see the people in front of her, Herme caught and held their attention. It was understood that the robot was talking to the person whose face she had centred.

3. 'Scripted' Situations

Before considering how successful CR might best handle interactive dialogues by multimodal techniques, it will help if we consider the nature of such dialogues. They are not bus-stop chats, nor are they theoretical discussions. They most probably relate to the immediate situation and can be determined largely from context: who is addressing whom and where. The 'why' should follow naturally, but the 'how' and 'when' are what most concern us here.

The Herme dialogues modelled casual meetings between strangers. CR, on the other hand, is by definition not with strangers, or not for long, anyway. We can assume teaching contexts, supervisory roles (caring for the elderly, for example), and personal assistants as typical examples. In each case, the 'text' of the conversation is constrained by the situation. Each dialogue session can be considered as a route to be followed from beginning to end; where fixed stages are negotiated along the way, like a Map-Task or Call-Centre interaction. At any point, the next utterance can be selected from a small group of candidates from a pre-determined list, or set of lists, provided by the Dialogue Manager, independent of the robot.

3.1. Robot as MP3 player

As a delivery device, all the robot really needs to decide at each point is which utterance to serve next. This decision is based on monitoring the engagement state of the interlocutor, using his or her expressive behaviour to make inferences about his or her level of understanding or agreement after each turn. No complex dialogue history or linguistic processing is needed so long as the utterance sequences are appropriately designed as part of a flow. The decision reduces to three options: *a)* to continue with the next pre-ordained utterance in the pattern sequence (i.e., to increment the list), *b)* to repair, repeat or re-construct parts of its previous utterance, or *c)* to perform a social repair strategy by opening up a new sub-dialogue (i.e., to deviate temporarily from the current list, and push the index for a later return).

The first of these is simple - if we think of the robot as equivalent to an mp3 player, then the command becomes "play next file". The second is more complicated, but requires a "execute repair" command (a subroutine depending on the pragmatic and linguistic content of the immediate situation but predictable from the system's current utterance. The repair options can be easily stored together with each utterance, in anticipation, in the dialogue list. The third is most complicated - "back to beginning" (reboot!) in Herme's case; when it has become clear that the engagement or attention of the interlocutor is broken or lost. In all three cases, the essential task for the robot is simply to discern whether the dialogue is still 'on-track'. The easiest way to do this is by monitoring the nonverbal behaviour through multimodal sensing. The hard way is through language understanding, which this paper argues the robot should not need.

3.2. Inline Data Collection and Processing

Recording personal interactive conversations has become very difficult due to ethical issues, so the traditional way of collecting large corpora of interactive personal speech must be replaced with lifelong learning for interactive dialogue systems. Whereas in the past we recorded and transcribed many hours of conversations, we presently argue for the avoidance of intrusive methods and are actively developing technology that allows for reduced use of ASR in conversational systems. This can be achieved through increased use of multimodal nonverbal behaviour sensing.

4. Discussion and Conclusion

In this position paper; we presented no formal evaluation of the work or specific results *per se*, but offer suggestions based on our experience developing social robots and interactive dialogue systems in light of the soft side of engineering. We are concerned about the social and ethical implications of AI-based personal dialogue systems and hope to encourage work that will reduce invasive recording, storage, or transcription of speech in order to proceed through an interactive dialogue.

Our work with Interactive Social Robots and spoken dialogues for HRI, avoids or reduces the need for ASR and natural language processing, by focussing on mutual adaptation and engagement sensing. This improves machine ethics, and human-machine trust. The paper has covered several rich communication capabilities: multimodal interaction, nonverbal communication, social signals, challenges for machine learning, and briefly touched on availability of suitable data. We did not explicitly discuss planning for interaction, and knowledge representation as required for dialogue design, but are learning more about these as we adapt our work to new tasks, new robots, and new users.

Acknowledgements

We thank Trinity College and Science Foundation Ireland for support of the Speech Communication Lab where these ideas came to fruition, and the IRC for funding the Joker Project which enabled the work.

References

[1] R. Reddy, "Robotics and Intelligent Systems in Support of Society," IEEE Intelligent Systems, vol. 21, pp. 24-31, 2006.

[2] M Koutsombogera, D Galanis, MT Riviello, N Tsere "Conflict cues in call centre interactions"- Conflict and Multimodal Communication, 2015

[3] Turing, A. (1950). Computing machinery and intelligence. Mind, vol. 59, no. 236, pp. 433

[4] D Schlangen, G Skantze "A general, abstract model of incremental dialogue processing" - Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 2009

[5] JG Han, E Gilmartin, N Campbell "Herme, yet another interactive conversational robot" - Affective Computing and Intelligent Interaction (ACII), 2013

[6] Davison, D., Gorer, B., Kolkmeier, J., Linssen, J. M., Schadenberg, B. R., van de Vijver, B., ... Reidsma, D. "Things that Make Robots Go HMMM : Heterogeneous Multilevel Multimodal Mixing to Realise Fluent, Multiparty, Human-Robot Interaction". In K. P. Truong, & D.

Reidsma (Eds.), Proceedings of eNTERFACE '16 (pp. 6-20). Enschede: Telematica Instituut / CTIT. 2017
[7] JOKe and Empathy of a Robot/ECA: Towards social and affective relations with a robot (call IUI 2012) http://www.chistera.eu/projects/joker
[8] MIRO: The First Robot That Thinks Like an Animal – http://consequentialrobotics.com/miro/
[9] Gilmartin, Emer, Brendan Spillane, Maria OÕReilly, Ketong Su, Christian Saam, Benjamin R. Cowan, Nick Campbell, and Vincent Wade. "Dialog Acts in Greeting and Leavetaking in Social Talk." In Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents, 29-30. ACM, 2017.
[10] Adam Miklosi, Peter Baranyi, Personal Communication
[11] Kaminski, J., et al. 2004. Word learning in a domestic dog: evidence for "fast mapping." Science 304: 1682-1683.
[12] JingGuang Han and Emer Gilmartin and Celine DeLooze and Brian Vaughan and Nick Campbell "The Herme Database of Spontaneous Multimodal Human-Robot Dialogues" 2012/5 Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) 978-2-9517408-7-7 European Language Resources Association (ELRA)
[13] K Gurney, TJ Prescott, P Redgrave "A computational model of action selection in the basal ganglia. I. A new functional anatomy" - Biological cybernetics, 2001
[14] OpenCV : Open Source Computer Vision Library - https://github.com/opencv