You Don't Need to Speak, You Need to Listen: Robot Interaction and Human-Like Turn-Taking

Matthew P. Aylett
Heriot Watt University and CereProc Ltd.
Edinburgh, UK
m.aylett@hw.ac.uk

Marta Romeo
Heriot Watt University
Edinburgh, UK
m.romeo@hw.ac.uk

ABSTRACT

The focus on one-to-one speak/wait conversational interaction with artificial system is partly misguided and partly cynical. Misguided because it pre-supposes that our relationship with such a system should be one-to-one and that human-like turn taking is never required. Cynical because we avoid the difficult challenge of building complex systems with a problematic route for publication. Whereas vision systems are regularly used in social robots and virtual agents to detect multiple dialogue partners and aid diarization, speech analysis and human-like turn-taking has lagged far behind. In this positional paper we make the case for focusing on human-like turn taking and multi-party interaction, discuss why realtime speech analysis and conversational management has been neglected, and put forward a program to correct this.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Computer systems organization** → *Robotics.*

KEYWORDS

social robots, multi-party dialogue, conversational user interfaces

ACM Reference Format:
Matthew P. Aylett and Marta Romeo. 2023. You Don't Need to Speak, You Need to Listen: Robot Interaction and Human-Like Turn-Taking. In *ACM conference on Conversational User Interfaces (CUI '23), July 19–21, 2023, Eindhoven, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3571884.3603750

1 INTRODUCTION

There are many things the academic community excels at but admitting failure is not one of them. This is a problem because when you try to build complex systems that no one has built before with very few resources, failure is something you have to live with. In general academics tend to convert failure into challenges, slip it into further work, or often just avoid focusing on it. It can be hard enough as it is to get funding to do interesting stuff without having to admit we failed to achieve our objectives last time. Thus, although we often learn through failure in a domain when we must promise and deliver success, there is a danger of sweeping failure under the carpet, of avoiding discussing it.

If we look at current conversational speech interfaces we see they are not very similar to human/human conversation. They are mostly two party speak-wait/speak-wait systems. Human conversation in contrast, is often multiparty, allows for fluid interruption and back channeling (See [22] for a detailed review of the phonetic and linguistic features of human-like turn-taking). Mimicking human behavior may not matter for many applications. For example, Siri, Google Assist, and Alexa function adequately without human/human style turn taking. It may also not be desireable in many contexts as it may encourage unwanted anthropomorphism. However, not being able to use effective elements of human behavior that are appropriate in an engineering design is severely limiting. But, perhaps what is more concerning, is that academics have been highlighting this deficit for many years and we haven't made any progress.

2 SOME HISTORY

Let's go back many years to a time when the default view of speech technology in interface design was that it was either inappropriate [20] or not *ready yet* which was often not directly stated. Munteanu et al. [15] pointed out in 2013 there was a *"a widespread perception that perfect domain-independent speech recognition is an unattainable goal."*

Years before HCI professionals were willing to engage with speech interfaces, dialogue engineers had been attempting to move away from a dominant approach of text in and out to using speech input and output. Two main challenges were set out: the need to deal with multiparty dialogue; the need to allow interactive conversational interfaces.

In 2004, David Traum wrote, *"Most formal and computational studies of natural language dialogue have considered only the two-party case. Eg., communication between two people, a person and a dialogue system, or a pair of agents."* [24].

So how did we do? Well, the general experience of trying to do multiparty speech interfaces has not been good. It is hard to find direct admissions of underlying problems in the literature. As we argued in the introduction, researchers prefer to focus on the positives. However, both from personal experience and anecdotally, speech recognition, diarization, conversational management and conversational speech synthesis are limited and can't offer the underlying technical support required.

Roll forwards almost two decades and in the Handbook of Social Agents in 2022 Gillet at al. write *"While most research has focused on one-on-one interactions, Socially Interactive Agents (SIAs) for multiparty interaction have received increasingly more attention in the*

CUI '23, July 19–21, 2023, Eindhoven, Netherlands

ACM ISBN 979-8-4007-0014-9/23/07.
https://doi.org/10.1145/3571884.3603750

past years for a number of reasons."[11]. The silver lining being that at least vision systems have progressed to the point where they could spot faces and work out how many people were in front of a robot. Based on this, you can at least deduce a dialogue is multiparty and use an array microphone to focus on the speaker talking and changing eye gaze (i.e. [1]).

However, that is about it. Dealing with two speakers speaking concurrently? No. Judging when to take or cede the floor (knowing when to stop or start speaking)? No. Joining a multiparty dialog? Definitely not[1].

What about human-like conversational interfaces? Over 20 years ago Allen et al [2] pointed out that speak-wait/speak-wait processing can *"... make the interaction unnatural and stilted, and will ultimately interfere with the user's ability to focus on the problem itself rather than on making the interaction work."* David Traum again wrote in 2004 [24] *"There has been a fair amount of work on turn-taking even for two-party dialogue. The basic questions are when to speak and when to stop speaking. Older dialogue systems generally force rigid turn-taking, where one party must wait until the other finishes before speaking. Many more recent systems allow* barge-in*, where a human who already understands a system query may provide the answer before the system has finished the utterance. Other systems allow interruption, by both parties, to correct or initiate something new, as well as to respond to the current utterance. Speakers can give verbal and non-verbal signals of continuation or imminent termination of the turn."*

So how did we do? We mostly ignored the problem and the people who suggested ways to solve it. In 2013 (nearly a decade later) Gabriel Skantze points out [23] *"Contrary to [human/human turn-taking], conversational systems of today typically process the dialogue one utterance at a time, and each module of the system has to complete the processing of the utterance before passing the result on to the next module."*. And in 2021, nearly two decades later, [22] *"Conversational systems (including voice assistants and social robots), on the other hand, typically have problems with frequent interruptions and long response delays, which has called for a substantial body of research on how to improve turn-taking in conversational systems."*

There is a body of previous work looking at incremental dialog processing[2] (e.g. [2, 12, 18, 19, 21, 28]). There are also toolkits available to implement incremental processing of dialog for example InproTK [6], Incremental RASA [17], Retico [14]. These systems follow a waterfall design pattern where each module can form hypotheses based on incremental input but allow replanning if these hypotheses are rejected as new data is processed.

It's just, to the author's knowledge, no one is using any of it.

One of the authors of this paper was involved in a multi-site European project with an intelligent virtual agent and we also did nothing with regards to incremental processing or multiparty dialog. When Amazon decided to build Alexa they opted for a *speak/wait* architecture and they were not short of a few dollars to spend on this problem. So there is nothing to be ashamed of. In the rest of this paper we will discuss the reasons human-like conversation has been avoided in CUIs, the challenges, the design considerations, and, in a spirit of happy optimism, suggest a way forwards.

[1]Caveat: Some of this may have been done for publication, but again and again when being presented by systems at demos and conferences none of this is working

[2]All dialog processing is incremental in some respects because you don't know what the next utterance will be. However, incremental in this context means processing before you discover the end-point of a dialog partner's current utterance.

3 HOW TO AVOID BUILDING HUMAN-LIKE CONVERSATIONAL SYSTEMS

To understand why we have made virtually no progress in 20 years let's list some of reasonable, cynical and misguided reasons to avoid human-like conversational interaction.

3.1 High Hanging Fruit

There are many interesting areas to study in CUIs. We need to publish new work, we need to get research proposals funded. Given two problems, one which requires extensive engineering knowhow, a multitude of complex systems that interact together (some difficult or impossible to source) and the requirement of a realtime multithreaded parallel architecture, and another which focuses on hot topics like personality, emotion, gender, social applications etc. It is a bit foolish to look at the first. Basically, human-like conversational processing is hard and best avoided. Also it's easy to *approximate* human-like conversation with a speak/wait system. So why not do that first, look at interesting areas and get back to it later?

3.2 Never Do Today What you Can Put Off Until Tomorrow

We have a study limitations and future work section in a paper for a reason. We can't do everything and we need to put a line under a piece of work at some point. Furthermore, the impact of realtime interruption, back channeling, human-like turn taking are hard to assess (especially as no one is doing any work in the area). So the best thing to do is to *approximate* human-like conversation with a speak/wait system and discuss issues like multiparty interaction and human-like turn taking when you outline the limitations of the study and talk about further work. Bear in mind the best thing about further work sections in a paper is that no one is going to make you do any of the work you suggest.

3.3 Mimicry is Creepy

As Aylett et al. [4] point out, just duplicating human behavior for the sake of it and regarding this as the research objective ignores the design process. For interactive systems it is easier to design your way out of it and not use human-like interaction. Winkle et al. [26] gives a good overview of the dangers of anthropormorphism and these arguments extend into the audio and conversational domains. This is quite convenient because we don't have to worry how deceptive or creepy human-like turn-taking might be if we can't build systems that do it.

4 THE CHALLENGES

Meanwhile, speech recognition has started to approach human performance, speech synthesis has become hard to distinguish from human generated speech and natural language processing is able to automatically spoof essays, chat responses and almost any other text we can think of.

Is human-like conversational interaction really that hard? Yes I'm afraid it is.

4.1 Realtime multi-threaded architectures

A speak/wait dialogue system can run on one thread/processor. Everything waits for everything to finish before doing anything. As soon as processes can be interrupted, or they must respond or send data within a fixed time window (say 20 milliseconds), a multi-threaded architecture is required. Such architectures are notoriously difficult to build, debug and guarantee stability. However we have an additional problem, this is not like a GUI where a user might interrupt an operation, these processes have to by synchronized in realtime. This is hard from a software design perspective. If you have colleagues that work in concurrency and parallel architectures give them a call.

4.2 Realtime audio and visual analysis

Lets take a fairly trivial component for audio analysis, pitch tracking. Pitch tracking analyses audio to estimate the pitch or melody in a spoken voice. Over the years, pitch trackers have become quite good at dealing with some noise and the difficulty of estimating pitch when it can be doubled or halved. But all this presupposes a single audio track of a human voice. Noise, crosstalk (when another persons voice is heard over the target voice), background music, the voice of the system you are building will wreck this data. Also, most of the mature systems are not realtime and not incremental.

This problem spreads into virtually all other speech analysis software you might use. Most of it uses batch processing or at least end-pointed data (i.e a coherent complete phrase or image sequence). But before you even adapt all these systems (speech recognition, social signal processing etc.) you will need to clean up your audio and that my friend is a challenge in itself.

Arguable, Alexa's biggest innovation was effective far-field microphones. Without being able to pick up the audio scene from multiple users and try, even in a mundane way, to work out who is talking to Alexa from the middle of a busy household, Alexa would have failed. Unfortunately both Google and Amazon systems are proprietary.

One of the biggest and most important noise cancellation processes required is to remove the sound of the system's own voice or audio output. For multiparty and human-like conversational interaction you must listen all the time which, apart from ethical considerations, requires an always on microphone. The system has to be able to continue to process audio when it is speaking. Without this functionality the system can never allow *barge in* (Allow a user to take the floor from the device by interrupting it). Ideally, this is required for adequate one-to-one spoken conversation but even more required with one-to-many where any one of the users might want to interrupt. Most current systems dodge this problem by just switching off the microphone while it makes a noise.

If this isn't challenging enough you also need realtime incremental diarization. Diarization is the term used to describe the solution to the problem of who said what when. There has been some success applying offline algorithms to fully recorded data sets but there is very little available that the authors are aware of that might be used in a realtime system.

4.3 Pollable, Interruptible, Streamable Renderers

The renderer is the device which produces the behavior and speech you design into your intelligent social agent. This could be a robot, a graphical character or a disembodied voice. In general, as researchers, we will use robots manufactured by third parties, we may also not have expert Unity programmers on hand to build graphical characters and use third party systems. For a system to be fit for purpose in terms of mimicking human-like conversational interaction it *must* be pollable, interruptible and streamable.

Pollable: You need to know what state the system is in and what it has done at any point in time. For example, if it is speaking, you need to know what audio has been output by time *t*. This is difficult to engineer because, often, systems just fling audio out to an audio device and metaphorically walk away. When this happens, you have no idea how much audio has been produced when an external event occurs, and no way of altering what it will say. This has become particularly problematic with cloud based/web based interfaces where the system may not even know *what* audio device is being used.

Interruptible: All speech synthesizers that support standard industry APIs are interruptible. This means that at any point the speech can be halted. Some systems [25] can even replan audio and insert it seamlessly into the audio stream. But for a CUI, all renderers (graphics, physical movement, sound effects) need to be interruptible and the supplier needs to give you access to this functionality.

Streamable: You need to be able to queue up a sequence of behaviors for your renderer so you can poll the system to see where its got, and interrupt the system gracefully when required. Then, you can replan, and queue up another set of behaviors given some external event. A system that will not process more input until it is finished cannot offer sufficient control.

4.4 Incremental Dialog Processing

In dialog, human participants typically respond within 200ms [7], whereas current digital systems can spend several seconds processing before saying anything. However, participants also may pause normally mid-turn for over a second. Thus it is impossible to mimic human turn-taking without knowing in advance when your dialog partner is going to finish. This is the point made by the researchers who have worked on incremental dialog processing for many years (see section 2). Why then, is there so little evidence that this work is being incorporated into modern CUIs? Aylett et al. [3] suggest the challenge of retooling completely the dialog manager, the requirement of incremental audio analysis (see section 4.2) and the massive jump in complexity has proved challenging.

4.5 Evaluating Conversational Interaction

Once you try and build a system that mimics human turn-taking you also face the difficulty of evaluation. As Clark et al. [8] point out *"A high proportion of research in our review used self-report questionnaires to measure other concepts like user satisfaction, usability, user attitudes towards speech interfaces and general user experience.*

A number of these self-report measures lacked any reliability or validity testing." Questionnaire evaluations are notoriously lacking in power and can easily be effected by factors outside experimental control. No evaluation equals no publication; building human-like conversational systems may not even measurably improve user experience in many contexts.

4.6 Lack of Novelty

If you can't publish you can't justify doing the work. Doing applied work, even if its good, hits the problem of novelty. For example, creating an incremental dialog processing system is not *novel* in itself. Also, when researchers have been saying for years that something should be done, even though virtually no one has done it, it lacks the sparkle of novelty. Without novelty you cannot publish your work.

5 DESIGNING HUMAN-LIKE MULTIPARTY CONVERSATIONAL INTERACTION

There is plenty of evidence of the importance of turn-taking behavior in human/human dialog (see [13] for a review). Furthermore, if we want to explore the use of CUIs in areas such as facilitation and mediation we need to deal with multiparty dialog. So, what advice and guidelines can we offer for designing human-like multiparty conversational systems?

Get Your Ducks in a Row: Before you can design any human-like turn-taking system you have to ensure that required third party systems can support it. There is nothing wrong with trying to design around some of these requirements (for example using a noise canceling microphone instead of incorporating an array microphone with beam forming). But this has to be worked into the design in advance. Crucially, the renderer needs to be pollable, interruptible and streamable. After that, as much incremental processing as possible is required.

Identify an Open Engineering Path: Some stuff can be added later and some can't. Once the architecture is built it is very hard to change it. If you use end-pointed cloud based speech recognition, can it be replaced or enhanced by on device streamed speech recognition or word spotting (i.e. [3])? Can you add the array microphone at a later date? When you have built the first system, how will you improve it?

Attack Concrete Measurable Use Cases: With a well specified system in mind how will it be evaluated? Can you measure latency? Can you measure interaction time? What is the envisaged interaction and how will the system achieve it?

Explore Dumb Systems that Interact Well: As Edlund points out [10], producing quite simple systems can have profound effect on the interaction. Edlund describes a *hummer* a system that reacts in a timely fashion with just a *ahuh* which has been shown to increase the speech output of the dialog partner (See also the Semaine project's Sensitive Artificial Listener [9]). Such systems allow us to explore human-like conversational processing without having to build complete systems.

6 HOW THE COMMUNITY CAN SUPPORT ADVANCES IN HUMAN-LIKE MULTIPARTY CONVERSATIONAL INTERACTION

The most important contribution the community can make to advancing human-like turn-taking for CUIs is to firstly accept that little progress has been made and secondly to respect the novelty of applied work in this area. Yes, ideas have been published, Yes, the concepts have been discussed, but without concrete examples of such systems the art will not advance.

Next, as Aylett et al. argued in 2014 [5] for the engagement of HCI with speech technology, it is critical that CUI researchers make good links with researchers and engineers in speech technology. For example, although offline diarization is a mature area of research in the speech community, realtime incremental diarization is not, *"it still remains as a challenging problem."* [16]. So pick up the phone, talk to colleagues in speech technology, they will be thrilled to support realtime work in a concrete demo-able system.

To support CUI researchers it is also important to make open source modules available for nuts and bolts requirements of human-like turn-taking systems: beam forming algorithms for array microphones, noise cancellation for listen while you speak, examples of using open source speech recognition systems like Kaldi incrementally, light weight word spotting systems like CereProc's chatty SDK which forms the basis of Honda Research's Haru realtime system. Much of this software is available but it typically assumes expertise knowledge in the area. Share you experiences and how-tos on the setup, evaluation and integration of third party open source (or proprietary systems).

Finally, we are in desperate need of realtime evaluation of user experience and engagement. Human-like turn-taking systems are by their nature realtime and many of the problems and benefits are hard to evaluate in a classic post study questionnaire. Current work in low cost EEG systems such as the Emotiv systems[3] suggest these systems could augment traditional HCI evaluation and give a deeper insight into what is happening on second by second basis (see [27] for a review).

7 CONCLUSION

We have a wonderful opportunity to design and build a new generation of delightful conversational interfaces. Advances in supporting technology have made the previously intractable, tractable. However, as a community, we have to embrace our failure in building human-like turn-taking systems. Once we can produce these systems effectively, we will be able to pick and choose what functionality is or isn't required by an application. There is nothing wrong with speak/wait interfaces in the right context. But the ability to build systems that can engage in fluid multiparty conversation opens a wealth of possibilities for CUIs to mediate and facilitate, as well as play let's pretend and perform for our enjoyment and well being.

REFERENCES

[1] Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive Behavioural Systems: COST 2102 International Training*

[3] https://www.emotiv.com/

School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers. Springer, 114–130.

[2] James Allen, George Ferguson, and Amanda Stent. 2001. An architecture for more realistic conversational systems. In *Proceedings of the 6th international conference on Intelligent user interfaces*. 1–8.

[3] Matthew P Aylett, Andrea Carmantini, and David A Braude. 2023. Why is My Social Robot so Slow? How a Conversational Listener can Revolutionize Turn-Taking. (2023).

[4] Matthew P Aylett, Benjamin R Cowan, and Leigh Clark. 2019. Siri, Echo and Performance: You have to Suffer Darling. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, alt08.

[5] Matthew P Aylett, Per Ola Kristensson, Steve Whittaker, and Yolanda Vazquez-Alvarez. 2014. None of a CHInd: relationship counselling for HCI and speech technology. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. 749–760.

[6] Timo Baumann, Okko Buß, and David Schlangen. 2010. InproTK in action: Open-source software for building german-speaking incremental spoken dialogue systems. (2010).

[7] Matthew Bull and Matthew Aylett. 1998. An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. In *Fifth International Conference on Spoken Language Processing*.

[8] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, et al. 2019. The state of speech in HCI: Trends, themes and challenges. *Interacting with computers* 31, 4 (2019), 349–371.

[9] Ellen Douglas-Cowie, Roddy Cowie, Cate Cox, Noam Amir, and Dirk Heylen. 2008. The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In *LREC workshop on corpora for research on emotion and affect*. ELRA Marrakech, Morocco, 1–4.

[10] Jens Edlund. 2019. Shoehorning in the Name of Science. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 8, 3 pages. https://doi.org/10.1145/3342775.3342794

[11] Sarah Gillet, Marynel Vázquez, Christopher Peters, Fangkai Yang, and Iolanda Leite. 2022. Multiparty interaction between humans and socially interactive agents. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*. 113–154.

[12] Helen Hastie, Oliver Lemon, and Nina Dethlefs. 2012. Incremental spoken dialogue systems: Tools and data. In *NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012)*. 15–16.

[13] Antje S Meyer. 2023. Timing in Conversation. *Journal of Cognition* 6, 1 (2023).

[14] Thilo Michael. 2020. Retico: An incremental framework for spoken dialogue systems. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 49–52.

[15] Cosmin Munteanu, Matt Jones, Sharon Oviatt, Stephen Brewster, Gerald Penn, Steve Whittaker, Nitendra Rajput, and Amit Nanavati. 2013. We need to talk: HCI and the delicate topic of spoken language interaction. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2459–2464.

[16] Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language* 72 (2022), 101317.

[17] Andrew Rafla and Casey Kennington. 2019. Incrementalizing RASA's Open-Source Natural Language Understanding Pipeline. *arXiv preprint arXiv:1907.05403* (2019).

[18] Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Multimodal continuous turn-taking prediction using multiscale RNNs. In *ICMI*. 186–190.

[19] David Schlangen and Gabriel Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse* 2, 1 (2011), 83–111.

[20] Ben Shneiderman. 2000. The limits of speech recognition. *Commun. ACM* 43, 9 (2000), 63–65.

[21] Gabriel Skantze. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *SIGDIAL*.

[22] Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language* 67 (2021), 101178.

[23] Gabriel Skantze and Anna Hjalmarsson. 2013. Towards incremental speech generation in conversational systems. *Computer Speech & Language* 27, 1 (2013), 243–262.

[24] David Traum. 2004. Issues in multiparty dialogues. In *Advances in Agent Communication: International Workshop on Agent Communication Languages, ACL 2003, Melbourne, Australia, July 14, 2003. Revised and Invited Papers*. Springer, 201–211.

[25] Mirjam Wester, David A Braude, Blaise Potard, Matthew P Aylett, and Francesca Shaw. 2017. Real-Time Reactive Speech Synthesis: Incorporating Interruptions.. In *INTERSPEECH*. 3996–4000.

[26] Katie Winkle, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. 2021. Assessing and Addressing Ethical Risk from Anthropomorphism and Deception in Socially Assistive Robots. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21)*. Association for Computing Machinery, New York, NY, USA, 101–109. https://doi.org/10.1145/3434073.3444666

[27] Tarannum Zaki and Muhammad Nazrul Islam. 2021. Neurological and physiological measures to evaluate the usability and user-experience (UX) of information systems: A systematic literature review. *Computer Science Review* 40 (2021), 100375.

[28] Lukas Zilka and Filip Jurcicek. 2015. Incremental LSTM-based dialog state tracker. In *ASRU*. 757–762.