
Proceedings Paper:
Rodriguez Munoz, T. (2022) Incremental Dialogue Modelling for Embodied Systems and Effort-based Modelling. In: Proceedings of 18th Workshop on Spoken Dialogue Systems for PhDs, Postdocs & New Researchers. 18th Workshop on Spoken Dialogue Systems for PhDs, PostDocs & New Researchers (YRRSDS 2022), 05-06 Sep 2022, Edinburgh, Scotland. , YRRSDS 2022 Website, pp. 33-35.

Teresa Rodríguez Muñoz

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello
Sheffield, S14DP

trodriguezmunoz1@sheffield.ac.uk
www.linkedin.com/in/teresarodriguezm

1 Research interests

My research interests include **incremental dialogue modelling** for **embodied systems**, capable of engaging in **multimodal, multi-party interactions**, and **social robotics**. My hypothesis is that an incremental spoken dialogue system (SDS) may be enhanced using **effort-based models**, which would allow for the regulation of conversational effort between interlocutors. I am currently in my first year of doctoral studies, exploring multiple research avenues, while keeping up to date with recent advances in the field.

1.1 Incremental Systems

Nowadays, voice-controlled applications are used for casual information retrieval or as smart-home devices, and users cherish the fast accessibility to data. However, these conversational agents behave like question & answer systems, engaging in non-fluid turn-taking interactions that resemble a table tennis game. Furthermore, users have become accustomed to two practices: (i) using wake words like "Hey Siri" or "Alexa" to initiate a dialogue and (ii) waiting for a short period of time until the voice-enabled device replies back. These adaptations are clear indicators that existing human-machine interactions (HMI) are far from "conversational".

As described in Skantze (2021), current turn-taking dialogue systems can be classified as either been *reactive* or *predictive*. While reactive models, like silence-based methods, rely upon past cues in dialogue, predictive models aim to continuously analyse speech and potentially project turn-completions. Moreover, voice-enabled artefacts would be able to plan what to say to the user in advance, barge-in or employ cooperative backchannels, and even reply back to the user as soon as they finish their turn. However, one must be aware that achieving natural timing is key; it is not a matter of quick or slow responses, but rather appropriately timed responses depending on the type of dialogue being exchanged.

Current work aims at coupling an incremental dialogue architecture with an effort-based model (Lindblom, 1990; Moore and Nicolao, 2017). This is because human beings display regulatory behaviour in everyday speech, and take into account what the speaker and the listener(s) share in common to adjust that effort. For that reason, if the agent could also identify the user's abilities and adjust its mode of communication, then different adaptability modes may arise. Effort-based principles apply at all levels, as autonomous, progressive learning is incremental and it is driven by an *intrinsic motivation*, which varies according to the urgency/importance of a given theme or task (Oudeyer et al., 2007; Oudeyer and Kaplan, 2009; Moulin-Frier et al., 2014).

1.2 Social Robots and Multimodality

As defined in Duffy et al. (1999), there are two important concepts in cognitive science that are highly relevant for defining the idea of a *social robot*: (i) physical embodiment and (ii) being situated in a social context. Embodied talking heads may be classified as social robots because they are *situated* in a particular environment and are capable of engaging in physical face-to-face interactions. Additionally, they can employ multimodal cues such as gaze and other non-verbal cooperative overlaps (i.e., nodding or various facial expressions) to show continued attention towards the user, prosodic realisation (Axelsson et al., 2022) and to convey a certain attitude.

During a conversation, humans use a variety of multimodal cues to signal turn-completion or willingness to speak next, amongst other things. Therefore, to enhance human-robot interactions (HRI), voice-enabled systems must be able to understand and employ these signals as well. Situated social robots prove to be an interesting experimental platform for conversational dialogue modelling, as they might be able to engage in more meaningful interactions relating to specific real-world settings.

1.3 System Evaluation

Several metrics are available to evaluate the performance of systems and compare them to existing models, which helps with reproducibility. However, what are the most suitable methods for assessing incremental dialogue systems? Subjective measures like user satisfaction and likelihood of future use are difficult to measure (Hastie, 2012), time-consuming and unsuitable for rapid prototyping of systems. This is mainly because individual users

have different goals and values, thus they hold different expectations of these systems.

Standard objective metrics like Word Error Rate (WER) are unsuitable for incremental systems (Köhn, 2018), as they fail to evaluate the intermediate behaviour of these systems. Metrics like latency-BLEU, to measure timeliness (Grissom II et al., 2014), or relative-correctness (*r-correctness*), to measure incremental quality (Baumann et al., 2009), have been proposed in previous work. Nevertheless, does the community have a clear, reference evaluation framework for incremental systems? Can these evaluation methods be used to assess the effectiveness, usability and performance of situated social robots?

1.4 Cultural Aspects to Dialogue Modelling

General patterns in turn-taking differ between countries and cultures, for instance, in the turn-taking latency between speakers (in milliseconds) (Stivers et al., 2009), the placement of backchannels and other cooperative feedback signals (Axelsson et al., 2022), and the timing of turn transitions (Dingemanse and Liesenfeld, 2022). Culturally-dependent SDS appears to be an interesting research area and not much work has been done on this domain. While some of the research highlighted above focuses on time, it would be interesting to explore the types of multimodal cues used across languages and how these differ depending on the language being spoken.

2 Spoken dialogue system (SDS) research

Despite their flaws, speech technologies like Apple's Siri and Amazon's Alexa are still very popular, with millions of sales every year. Therefore, the community would benefit from improving current spoken language interfaces to achieve enriched interactive behaviours.

Future SDS research should involve shifting from reactive, end-of-turn detection models to continuous dialogue modelling, which allows for processing language *incrementally* and achieving real-time processing. Moreover, applying these incremental dialogue architectures to situated social robots (e.g., talking heads) may be the next step towards developing more conversational dialogue systems. Instead of being restricted to a single vocal mode of communication, physical agents have the ability to employ multimodal cues such as facial expressions, gestures and prosody to realise multimodal dialogues.

Finally, as mentioned in Section 1.4, language diversity is a hot topic within the community, and I believe that SDS research will shift towards creating linguistically diverse data and language- or cultural-specific design principles. These will allow for model adaptability to different countries and cultures, helping to design SDS and speech technologies for a broad range of communities.

3 Suggested Topics for Discussion

The following three topics are suggested as potential discussion themes for this year's *YRRSDS* workshop panel discussion:

- What are the key differences between SDS research in academia and industry? How do the different motivating factors, if any, affect the overall research landscape?
- What are current opinions on efforts to support and generate SDS technologies for low-resource and endangered languages? Are we promoting language diversity enough?
- Where are we heading on our journey to achieve true social intelligence and fluent, conversational dialogues in HRI and SDS research?

Acknowledgements

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

References

Agnes Axelsson, Hendrik Buschmeier, and Gabriel Skantze. 2022. Modeling feedback in interaction with conversational agents—a review. *Frontiers in Computer Science* 4.

Timo Baumann, Michaela Atterer, and David Schlangen. 2009. Assessing and improving the performance of speech recognition for incremental systems. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*. pages 380–388.

Mark Dingemanse and Andreas Liesenfeld. 2022. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pages 5614–5633.

Brian R Duffy, Colm Rooney, Greg MP O'Hare, and Ruadhan O'Donoghue. 1999. What is a social robot? In *10th Irish Conference on Artificial Intelligence & Cognitive Science, University College Cork, Ireland, 1-3 September, 1999*.

Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Proceedings of the 2014*

Conference on empirical methods in natural language processing (EMNLP). pages 1342–1352.

Helen Hastie. 2012. Metrics and evaluation of spoken dialogue systems. In *Data-driven methods for adaptive spoken dialogue systems*, Springer, pages 131–150.

Arne Köhn. 2018. Incremental natural language processing: challenges, strategies, and evaluation. *arXiv preprint arXiv:1805.12518* .

Björn Lindblom. 1990. Explaining phonetic variation: A sketch of the h&h theory. In *Speech production and speech modelling*, Springer, pages 403–439.

Roger K Moore and Mauro Nicolao. 2017. Toward a needs-based architecture for 'intelligent'communicative agents: Speaking with intention. *Frontiers in Robotics and AI* 4:66.

Clément Moulin-Frier, Sao Mai Nguyen, and Pierre-Yves Oudeyer. 2014. Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. *Frontiers in psychology* 4:1006.

Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. 2007. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation* 11(2):265–286.

Pierre-Yves Oudeyer and Frederic Kaplan. 2009. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics* 1:6.

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

Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106(26):10587–10592.

Biographical sketch

Teresa Rodríguez Muñoz is a PhD student at the UKRI Centre for Doctoral Training (CDT) in Speech and Language Technologies (SLT) and their Applications. The CDT is based in the Department of Computer Science at the University of Sheffield (UoS), UK. She is currently beginning her research on incremental dialogue modelling for embodied systems and HRI with situated social robots. She is working under the supervision of Prof Roger K. Moore, Chair of Spoken Language Processing at UoS. Prior to starting her PhD, she received her integrated master's degree at Cardiff University. In her free time, she enjoys travelling, scuba diving and playing video games.