Honda Research Institute EU
**Honda Research Institute Europe GmbH**
https://www.honda-ri.de/

# Spoken Language Interaction with Virtual Agents and Robots based on Transparency, Situatedness and Personalization

## Martin Ernst Heckmann

### 2020

**Preprint:**

This is an accepted article published in Dagstuhl Seminar Spoken Language Interaction with Virtual Agents and Robot (SLIVAR). The final authenticated version is available online at: https://doi.org/[DOI not available]

# Spoken Language Interaction with Virtual Agents and Robots based on Transparency, Situatedness and Personalization

Martin Heckmann
Honda Research Institute Europe GmbH

February 18, 2020

When two humans engage in an interaction, two independent minds with different experiences and views on the world come together. To make this joint activity a success, they have to work together. They need to align their mental representations to be able to form a common ground and define a joint goal for the interaction [1, 2, 3, 4, 5]. The communicative signals they use for such an alignment are not limited to the words they utter but encompasses a multitude of potentially multimodal signals such as prosodic variations and gestures [6, 7]. These signals also help to coordinate when the partners take their turns [8]. The necessary alignment between the partners affects a multitude of domains, the phonological, syntactic and semantic level as well as the situation model [2, 3]. With situation model, I want to refer to different aspects of the current situation such as the environment in which the interaction takes place and the progression of the interaction so far. During the interaction they also make assumptions about their partner's mental world, her ability to perceive, understand and judge, often referred to as common sense, her prior knowledge on the topic of the interaction, her goals and intentions and her current state, traits, skills and personal preferences [4]. The larger the differences between the partners' mental worlds, the more work the alignment process needs. A true alignment can never be reached but is also not necessary as long as the joint goals of the interaction are achieved.

Unfortunately, the different building blocks fundamental to human-human communication which I have outlined above, i.e. the capabilities to perceive and interpret multimodal communicative signals, build an adequate model of the environment, keep track of the history of the interaction, reason with common sense, recruit domain knowledge, interpret the partner's goals as well as model her state, traits, skills and preferences, are at best manifested in a very different form in an artificial agent. In today's artificial agents they are often not present at all or only in a rudimentary form. Furthermore, if and how these capabilities are implemented in the artificial agent is in most cases opaque to the user, i.e. the system is not transparent to the user. This means that the alignment process will take a lot of effort from the human and in the end will leave a large gap between the partners' mental models, often too large to achieve the goal of the interaction.

A common and logical approach to improve the interaction is to improve the

system's capabilities, to make them closer to that of a human. Many people have investigated the role of prosody [9, 6] as well as how to integrate the information from all available modalities [10, 11, 12, 13, 14]. We have extended this by multimodal integration and personalization for prosodic analysis [15, 16] and reference resolution [17]. In classical dialog systems, the situation model is often limited to keeping track of the dialog history [18, 19, 20]. For human robot interaction, mainly models have been developed to visually perceive the world and link these percepts to internal concepts, words in particular. This process is often referred to as grounding, sometimes also as anchoring, and can be based on pre-existing categories [21, 22, 23] or learned in interaction [24, 25]. The domain knowledge of classical dialog systems is typically very narrow, based on hand-crafted domains such as bus information [18], restaurant reservation [19], or, more recently, also technical support dialogues [20]. For the interaction with robots it is common to establish the domain knowledge by combining information from external ontologies with the learning of new representations in interaction with the user [26, 22]. Common sense reasoning is then implemented by reasoning on these ontologies. We proposed an approach to automatically acquire task-specific domain knowledge from online text-based resources, e.g. in the context of a repair task [27].

Many of the functionalities mentioned above are still nowhere near the level of a human. In my view, the most challenging and promising domains relevant to improve the interaction with artificial agents are better environment models and domain knowledge combined with an adaptation to the individual user.

To really meet human expectations will most likely require an AI with reasoning capabilities on par with a human. Yet such an AI does not seem to be around the corner [28]. Hence, instead of raising the system's perception and reasoning capabilities, I think a more promising approach is to increase its transparency. If the system states, capabilities and intentions are transparent to the human, the alignment process and hence the overall interaction can be much more efficient [29]. Humans use backchannels, facial expressions and emotional displays to give feedback on the interaction and the progress of the alignment [30, 31, 32, 33]. Backchannels, facial expressions and emotions have been also investigated in the context of human-agent interaction [34, 35, 32]. However, the research on emotions in human-agent interaction typically focuses on detecting the emotions of the human to enable emphatic responses of the agent [36, 37]. I consider using emotional displays and other non-verbal cues to provide insights into the capabilities and mental states of the agent an interesting and very relevant topic to help to align the minds of the partners and to improve the interaction. Connected to this is the need to empower the agent to make its reasoning steps transparent, often referred to as explainable AI [38, 39]. We took one step into this direction by suggesting an approach to create annotations with explanations that can be used to train a system to provide explanations for its reasoning [40]. In general, spoken feedback from the agent bears the risk of triggering very high expectations on the human's side with respect to the agent's understanding and reasoning capabilities. I consider finding good models to convey the agent's limitations via its communication an important direction to increase transparency [41].

In short, I am convinced that faster progress can be made if we focus more on making the limitations of the agent transparent then on trying to overcome them.

# References


[1] H. H. Clark and S. E. Brennan, "Grounding in communication.," in *Perspect. Soc. Shar. Cogn.*, pp. 127–149, American Psychological Association, 1991.

[2] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behav. Brain Sci.*, vol. 27, no. 02, 2004.

[3] S. Garrod and M. J. Pickering, "Why is conversation so easy?," *Trends Cogn. Sci.*, vol. 8, no. 1, pp. 8–11, 2004.

[4] K. Friston and C. Frith, "A Duet for one," *Conscious. Cogn.*, vol. 36, pp. 390–405, nov 2015.

[5] T. Scott-Phillips, *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. Red Globe Press, 2015.

[6] J. Hirschberg, "Communication and prosody: Functional aspects of prosody," *Speech Commun.*, vol. 36, no. 1-2, pp. 31–43, 2002.

[7] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Commun.*, vol. 57, pp. 209–232, 2014.

[8] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language (Baltim).*, vol. 50, no. 4, p. 696, 1974.

[9] Y. Sagisaka, N. Campbell, and N. Higuchi, eds., *Computing Prosody*. New York, NY: Springer US, 1997.

[10] M. Turk, "Multimodal interaction: A review," *Pattern Recognit. Lett.*, vol. 36, no. 1, pp. 189–195, 2014.

[11] S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, eds., *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations - Volume 1*. ACM Books, 2018.

[12] R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel, "Natural human-robot interaction using speech, head pose and gestures," in *2004 IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, vol. 3, pp. 2422–2427, 2004.

[13] K. Funakoshi, M. Nakano, T. Tokunaga, and R. Iida, "A unified probabilistic approach to referring expressions," in *Proc. 13th Annu. Meet. Spec. Interes. Gr. Discourse Dialogue*, pp. 237–246, 2012.

[14] C. Kennington and D. Schlangen, "A simple generative model of incremental reference resolution for situated dialogue," *Comput. Speech Lang.*, vol. 41, pp. 43–67, 2017.

[15] M. Heckmann, "Audio-visual word prominence detection from clean and noisy speech," *Comput. Speech Lang.*, vol. 48, pp. 15–30, 2018.

[16] A. Schnall and M. Heckmann, "Feature-space SVM adaptation for speaker adapted word prominence detection," *Comput. Speech Lang.*, vol. 53, pp. 198–216, jun 2019.

[17] D. Kleingarn, N. Nabizadeh, M. Heckmann, and D. Kolossa, "Speaker-adapted neural-network-based fusion for multimodal reference resolution," in *Proc. 20th Annu. SIGdial Meet. Discourse Dialogue*, (Stockholm, Sweden), pp. 210–214, Association for Computational Linguistics, 2019.

[18] J. Williams, A. Raux, D. Ramachandran, and A. Black, "The dialog state tracking challenge," in *Proc. SIGDIAL 2013 Conf.*, pp. 404–413, Association for Computational Linguistics, 2013.

[19] M. Henderson, B. Thomson, and S. Young, "Word-based dialog state tracking with recurrent neural networks," in *Proc. 15th Annu. Meet. Spec. Interes. Gr. Discourse Dialogue*, pp. 292–299, Association for Computational Linguistics (ACL), 2014.

[20] K. Yoshino, C. Hori, J. Perez, L. F. D'Haro, L. Polymenakos, C. Gunasekara, W. S. Lasecki, J. K. Kummerfeld, M. Galley, C. Brockett, J. Gao, B. Dolan, X. Gao, H. Alamari, T. K. Marks, D. Parikh, and D. Batra, "Dialog System Technology Challenge 7," in *Proc. NIPS2018 2nd Conversational AI Work.*, 2018.

[21] D. K. Misra, J. Sung, K. Lee, and A. Saxena, "Tell me Dave: Context-sensitive grounding of natural language to manipulation instructions," *Int. J. Rob. Res.*, 2016.

[22] S. Lemaignan, M. Warnier, E. A. Sisbot, A. Clodic, and R. Alami, "Artificial cognition for social human–robot interaction: An implementation," *Artif. Intell.*, vol. 247, pp. 45–69, jun 2017.

[23] L. Fischer, S. Hasler, J. Deigmöller, T. Schnürer, M. Redert, U. Pluntke, K. Nagel, C. Senzel, A. Richter, and J. Eggert, "Where is the tool? - grounded reasoning in everyday environment with a robot," in *Proc. Int. Cogn. Robot. Work.*, CEUR Workshop Proceedings, 2018.

[24] D. K. Roy and A. P. Pentland, "Learning words from sights and sounds: a computational model," *Cogn. Sci.*, vol. 26, pp. 113–146, jan 2002.

[25] C. Matuszek, "Grounded language learning: Where robotics and NLP meet," in *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2018-July, pp. 5687–5691, 2018.

[26] M. Tenorth and M. Beetz, "KnowRob: A knowledge processing infrastructure for cognition-enabled robots," *Int. J. Rob. Res.*, vol. 32, no. 5, pp. 566–590, 2013.

[27] N. Nabizadeh, D. Kolossa, and M. Heckmann, "MyFixit : An Annotated Dataset , Annotation Tool , and Baseline Methods for Information Extraction from Repair Manuals," in *Proc. Twelth Int. Conf. Lang. Resour. Eval.*, 2020.

[28] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, "Viewpoint: When will ai exceed human performance? Evidence from ai experts," *J. Artif. Intell. Res.*, vol. 62, pp. 729–754, 2018.

[29] R. K. Moore and M. Nicolao, "Toward a needs-based architecture for 'intelligent' communicative agents: Speaking with intention," *Front. Robot. AI*, vol. 4, p. 66, dec 2017.

[30] A. Eshghi, C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver, "Feedback in conversation as incremental semantic update," in *IWCS 2015 - Proc. 11th Int. Conf. Comput. Semant.*, pp. 261–271, 2015.

[31] K. P. Truong, R. Poppe, I. De Kok, and D. Heylen, "A multimodal analysis of vocal and visual backchannels in spontaneous dialogs," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 2973–2976, 2011.

[32] C. Lang, S. Wachsmuth, M. Hanheide, and H. Wersing, "Facial Communicative Signals - Valence Recognition in Task-Oriented Human-Robot Interaction," *Int. J. Soc. Robot.*, 2012.

[33] P. Ekman, *Emotions revealed: recognizing faces and feelings to improve communication and emotional life.* Times Books, 2003.

[34] G. Skantze, A. Hjalmarsson, and C. Oertel, "Turn-taking, feedback and joint attention in situated human-robot interaction," *Speech Commun.*, vol. 65, pp. 50–66, 2014.

[35] C. Becker, S. Kopp, and I. Wachsmuth, "Why Emotions should be Integrated into Conversational Agents," in *Conversational Informatics An Eng. Approach* (T. Nishida, ed.), pp. 49–67, John Wiley & Sons, 2007.

[36] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *2011 IEEE Int. Conf. Autom. Face Gesture Recognit. Work. FG 2011*, pp. 827–834, 2011.

[37] L. Devillers, M. Tahon, M. A. Sehili, and A. Delaborde, "Inference of Human Beings' Emotional States from Speech in Human–Robot Interactions," *Int. J. Soc. Robot.*, vol. 7, pp. 451–463, aug 2015.

[38] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda," in *Conf. Hum. Factors Comput. Syst. - Proc.*, vol. 2018-April, 2018.

[39] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[40] N. Attari, M. Heckmann, and D. Schlangen, "From Explainability to Explanation: Using a Dialogue Setting to Elicit Annotations with Justifications," in *Proc. 20th Annu. SIGdial Meet. Discourse Dialogue*, (Stockholm, Sweden), pp. 331–335, Association for Computational Linguistics, 2019.

[41] R. K. Moore, "Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction," in *Lect. Notes Electr. Eng.*, vol. 999 LNEE, pp. 281–291, 2017.