Using Social and Physiological Signals for User Adaptation in Conversational Agents

Doctoral Consortium

Patrik Jonell
KTH Royal Institute of Technology
pjjonell@kth.se

ABSTRACT

In face-to-face communication, humans subconsciously emit social signals which are picked up and used by their interlocutors as feedback for how well the previously communicated messages have been received. The feedback is then used in order to *adapt the way the coming messages are being produced and sent* to the interlocutor, leading to the communication to become as efficient and enjoyable as possible. Currently however, it is rare to find conversational agents utilizing this feedback channel for altering *how* the multimodal output is produced during interactions with users, largely due to the complex nature of the problem. In most regards, humans have a significant advantage over conversational agents in interpreting and acting on social signals. Humans are however restricted to a limited set of sensors, "the five senses", which conversational agents are not. This makes it possible for conversational agents to use specialized sensors to pick up physiological signals, such as skin temperature, respiratory rate or pupil dilation, which carry valuable information about the user with respect to the conversation. This thesis work aims at developing methods for utilizing both social and physiological signals emitted by humans in order to *adapt the output of the conversational agent*, allowing for an increase in conversation quality. These methods will primarily be based on automatically learning adaptive behavior from examples of real human interactions using machine learning methods.

KEYWORDS

Learning agent capabilities (agent models, communication, observation); Deep learning; Single and multi-agent planning and scheduling

ACM Reference Format:
Patrik Jonell. 2019. Using Social and Physiological Signals for User Adaptation in Conversational Agents. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Speech based virtual assistants, such as Alexa, Google Assistant and Siri have established themselves as a ubiquitous part of our daily lives [6]. A similar tendency can also be noted for virtual agents and social robots (both examples of conversational agents), which are becoming increasingly prevalent in our daily environments – even more so for public environments and institutions such as schools and care facilities for the elderly, as can be seen by the increasing amount of research in those domains [9, 12, 17, 18]. In those circumstances the importance of efficient and enjoyable communication is even higher. Some elderly people have a hard time adjusting to interacting with robots. When faced with a robot which is speaking too fast in a loud environment and is completely oblivious to the confusion the users express through their non-verbal language, the forthcoming disasters are inevitable. In such cases, the elderly person (involuntarily) expresses confusion through non-verbal signals, which is something humans easily pick up and use to alter the production of the coming messages (or if noticing a complete failure, repeat or rephrase it).

The reignited developments in the field of machine learning during the last decade have made it possible to train statistical models based on large datasets from human behavior. This in turn can allow for training models based on real examples of how humans adapt their production of *system output*, while conditioned on using the above mentioned signals, producing the same sorts of adaptations humans would employ in similar situations, leading to more efficient and enjoyable interactions.

This paper attempts to outline the necessary steps toward developing and evaluating a real-time system for conversational agents which has been trained on human examples and previous interactions, in order to learn how to adapt to its interlocutor based on various input signals with the purpose of improving interaction quality in terms of efficiency and enjoyment while conversing with the system. The main research question for this thesis work is: **What methods can be used in order to create a model for multimodal conversational agents engaging in face-to-face interactions allowing adaption of conversational behavior based on social and physiological signals with the purpose of improving efficiency and enjoyment?**.

The scope of this work is constricted in several ways. It assumes that existing methods for generating multimodal system output already are available, such as speech synthesizers, and the rendering of facial expressions. The intention is to develop methods that use already existing methods in combination with each other to create a coherent output based on perceived input signals.

2 RELATED WORK

Martins et al. recently conducted a survey of the field of adaptive social robots. The authors chose to divide the literature into three main categories; (a) *adaptive systems with no user model*, (b) *systems based on static user models* and (c) *systems based on dynamic user models* [25]. These categories will also be used below.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

(a) Adaptive systems with no user model are usually reactive systems, reacting to a certain parameter in real time. For example Lubhold et al. presents a system with a robot adapting its pitch to the user's pitch, following the concept of entrainment [24]. Hirshfield et al. used neurophysiological measurements to inform the system of which gesture to use [14]. A basic method of conversational adaptation to the user is mimicry which has been explored by several researchers [7, 32, 38]. Researchers have also investigated using the *emotional state* of the user as an input parameter to adaptive social robot [5, 10].

(b) Systems based on static user models are given a model of the user at the beginning of the interaction, or creates one through a calibration processes, and adapts to the user according to this model through out the whole interaction. Torrey et al. adapted the level of detail of help provided in a cooking task based on the user's previous experience [31]. Kistler et al. explored cultural adaptation by the means of the agent's proxemics behavior [21].

(c) Systems based on dynamic user models combines the methods from the two previous categories such that it makes use of a user model, but alters it continuously during the interaction. Several researchers have been using reinforcement learning in order to achieve this [13, 22, 36]. Additionally there are a few commercial robots within this category [1–3].

This work primarily lies within category (c) as the intended system will try to build a dynamic user model from the input signals and use this to condition the generation of the output signal.

Vinciarelli et al. define social signals as "Social signals are spoken and wordless messages like head nods, winks, uh, and yeah utterances, which are sent by means of body gestures and postures, facial expressions and gaze, vocal expressions and speech". These social signals carry communicative intentions, such as affective and cognitive state, illustrators (for example pointing at a referred object), etc [34]. Cognitive load and its negative impact on conversations with dialog systems is for example an important aspect to take into consideration [23]. Various methods of estimating cognitive load from various multimodal signals has been extensively researched [4, 20, 23, 37]. Furthermore researchers have attempted to estimate user's affective state [8, 22, 29] so that it can be used as an input to conversational systems. Measuring physiological signals using non-invasive methods is an active research field. For example measuring heart rate using a consumer grade digital camera [33] or using thermal cameras in order to measure the respiratory rate by measuring the changes in temperature around the nose [16].

In order to generate facial expressions and gestures one can look into the field of motion synthesis. Pham et al. and Karras et al. have proposed methods for generating facial expressions in a virtual agent using speech as input to the model [19, 28]. Holden et al. describes how to model style variation in motion synthesis using convolutional neural networks (CNN) [15]. Several authors explore the concept of various embeddings, such as for example Wang et al. where the authors present global style tokens (GST) which encode speaking style [35]. Conditional variation autoencoders (CVAE) have been investigated in order to predict (but also generate) plausible head motion sequences based on input audio [11] Similarly Sadoughi and Busso used conditional generative adversarial networks (GAN) for the same task [30].

3 METHOD

The problem can be formulated as three distinct parts; (1) methods intended to capture multimodal human-human interactions containing both social and physiological signals, and how the receiver of these feedback signals acts on them, (2) defining which of these signals are the most relevant, and whether they provide sufficient information for the purpose of improving the current dialog and lastly (3) implementing a system which can learn from examples how humans adapt using those signals for altering the generation of actions throughout the various modalities (e.g. facial expressions, prosody, etc.).

(1) Data collection We need to be able to capture human-human interactions where these social and physiological signals are present. It is also of importance that these data capturing methods are unobtrusive and non-invasive. If the system is going to be able to be used in a public environment, it should not require the user to wear a special device on their body. To this end two methods have been developed in the scope of this project; A high-quality multi-sensory data recording framework for capturing rich synchronized multimodal data from a limited amount of participant interacting with each other [27], and a method that can be seen as the opposite to the other, as it is intended to collect multimodal data using crowdsourcing from a large amount of participants being exposed to stimuli [26].

(2) Social and physiological signal perception In order to be able to feed relevant social and physiological signals as control parameters into the systems model, these have to first be defined and evaluated making sure that they can successfully be used for generating adaptive behavior. *Facial expressions*, *gestures*, *gaze*, *body posture* and various *prosodic features* will primarily be investigated, either separately or together. As for the physiological signals, there are a number of interesting signals that will be investigated, such as respiratory rate, heart rate, skin temperature, etc. The relevant signals are those that can be used as proxies for underlying psychological states, such as cognitive load, valance or confusion.

(3) Generation of adaptive behavior The final part of this work is considered with combining the previous efforts in order to be able to generate adaptive conversational multimodal behavior based on input signals from the user. Building on previous work and by exploring the use of for example DNNs, GANs and CVAEs in the context of learning how humans produce adaptive behaviors in conversations, conditioned on the social and physiological signals, it is hopefully possible to produce output which is achieving the goal of improving the interaction with the system, i.e. increasing the efficiency and enjoyment.

An important and crucial aspect is how to evaluate the adaptive behavior generation. The final system will be compared with a baseline system which does not adapt to the user. In a series of within subject experiments, users will interact with both systems and evaluate enjoyment, social presence and other related measures. These interactions will also be evaluated by third-party observers.

4 ACKNOWLEDGMENT

I would like to extend a special thanks to Jonas Beskow, Joakim Gustafson, and André Pereira. This project is supported by Swedish Foundation for Strategic Research Grant No.: RIT15-0107.

REFERENCES

[1] 2001. PARO Therapeutic Robot. http://www.parorobots.com/. Accessed: 2018-11-26.

[2] 2016. Buddy, the emotional robot. https://buddytherobot.com. Accessed: 2018-11-26.

[3] 2017. Jibo. https://www.jibo.com/. Accessed: 2018-11-26.

[4] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. 2017. Cognitive heat: exploring the usage of thermal imaging to unobtrusively estimate cognitive load. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 33.

[5] Muneeb Imtiaz Ahmad, Omar Mubin, and Joanne Orlando. 2017. Adaptive social robot for sustaining social engagement during long-term children–robot interaction. *International Journal of Human–Computer Interaction* 33, 12 (2017), 943–962.

[6] Muhammad Raisul Alam, Mamun Bin Ibne Reaz, and Mohd Alauddin Mohd Ali. 2012. A review of smart homesâĂŤPast, present, and future. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (2012), 1190–1203.

[7] Jeremy N. Bailenson and Nick Yee. 2005. Digital Chameleons: Automatic Assimilation of Nonverbal Gestures in Immersive Virtual Environments. *Psychological Science* 16, 10 (2005), 814–819. https://doi.org/10.1111/j.1467-9280.2005.01619.x arXiv:https://doi.org/10.1111/j.1467-9280.2005.01619.x PMID: 16181445.

[8] Laura Boccanfuso, Quan Wang, Iolanda Leite, Beibin Li, Colette Torres, Lisa Chen, Nicole Salomons, Claire Foster, Erin Barney, Yeojin Amy Ahn, et al. 2016. A thermal emotion classifier for improved human-robot interaction. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on.* IEEE, 718–723.

[9] Joost Broekens, Marcel Heerink, Henk Rosendal, et al. 2009. Assistive social robots in elderly care: a review. *Gerontechnology* 8, 2 (2009), 94–103.

[10] N. Churamani, P. Barros, E. Strahl, and S. Wermter. 2018. Learning Empathy-Driven Emotion Expressions using Affective Modulations. In *2018 International Joint Conference on Neural Networks (IJCNN).* 1–8. https://doi.org/10.1109/IJCNN.2018.8489158

[11] David Greenwood, Stephen Laycock, and Iain Matthews. 2017. Predicting head pose in dyadic conversation. In *International Conference on Intelligent Virtual Agents.* Springer, 160–169.

[12] Marcel Heerink, Ben Kröse, Vanessa Evers, and Bob Wielinga. 2009. Influence of social presence on acceptance of an assistive social robot and screen agent by elderly users. *Advanced Robotics* 23, 14 (2009), 1909–1923.

[13] Jacqueline Hemminghaus and Stefan Kopp. 2017. Towards Adaptive Social Behavior Generation for Assistive Robots Using Reinforcement Learning. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17).* ACM, New York, NY, USA, 332–340. https://doi.org/10.1145/2909824.3020217

[14] Leanne Hirshfield, Tom Williams, Natalie Sommer, Trevor Grant, and Senem Velipasalar Gursoy. 2016. Workload-driven Modulation of Mixed-reality Robot-human Communication. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data (MCPMD '18).* ACM, New York, NY, USA, Article 3, 8 pages. https://doi.org/10.1145/3279810.3279848

[15] Daniel Holden, Jun Saito, and Taku Komura. 2016. A Deep Learning Framework for Character Motion Synthesis and Editing. *ACM Trans. Graph.* 35, 4, Article 138 (July 2016), 11 pages. https://doi.org/10.1145/2897824.2925975

[16] Hieyong Jeong, Yutaka Matsuura, and Yuko Ohno. 2017. Measurement of Respiration Rate and Depth Through Difference in Temperature Between Skin Surface and Nostril by Using Thermal Image. *Studies in health technology and informatics* 245 (2017), 417–421.

[17] Patrik Jonell, Joseph Mendelson, Thomas Storskog, Goran Hagman, Per Ostberg, Iolanda Leite, Taras Kucherenko, Olga Mikheeva, Ulrika Akenine, Vesna Jelic, et al. 2017. Machine Learning and Social Robotics for Detecting Early Signs of Dementia. *arXiv preprint arXiv:1709.01613* (2017).

[18] Aidan Jones and Ginevra Castellano. 2018. Adaptive Robotic Tutors that Support Self-Regulated Learning: A Longer-Term Investigation with Primary School Children. *International Journal of Social Robotics* 10, 3 (01 Jun 2018), 357–370. https://doi.org/10.1007/s12369-017-0458-z

[19] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 94.

[20] M. Asif Khawaja, Fang Chen, Christine Owen, and Gregory Hickey. 2009. Cognitive Load Measurement from User's Linguistic Speech Features for Adaptive Interaction Design. In *Human-Computer Interaction – INTERACT 2009*, Tom Gross, Jan Gulliksen, Paula Kotzé, Lars Oestreicher, Philippe Palanque, Raquel Oliveira Prates, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 485–489.

[21] Felix Kistler, Birgit Endrass, Ionut Damian, Chi Tai Dang, and Elisabeth André. 2012. Natural interaction with culturally adaptive virtual characters. *Journal on Multimodal User Interfaces* 6, 1 (01 Jul 2012), 39–47. https://doi.org/10.1007/s12193-011-0087-z

[22] Iolanda Leite, André Pereira, Ginevra Castellano, Samuel Mascarenhas, Carlos Martinho, and Ana Paiva. 2011. Modelling empathy in social robotic companions. In *International Conference on User Modeling, Adaptation, and Personalization.* Springer, 135–147.

[23] José Lopes, Katrin Lohan, and Helen Hastie. 2016. Symptoms of Cognitive Load in Interactions with a Dialogue System. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data (MCPMD '18).* ACM, New York, NY, USA, Article 4, 5 pages. https://doi.org/10.1145/3279810.3279851

[24] Nichola Lubold, Erin Walker, and Heather Pon-Barry. 2016. Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion. In *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on.* IEEE, 255–262.

[25] Gonçalo S. Martins, Luís Santos, and Jorge Dias. 2018. User-Adaptive Interaction in Social Robots: A Survey Focusing on Non-physical Interaction. *International Journal of Social Robotics* (16 Jun 2018). https://doi.org/10.1007/s12369-018-0485-4

[26] Jonell Patrik, Oertel Catharine, Dimosthenis Kontogiorgos, Beskow Jonas, and Gustafson Joakim. 2018. Crowdsourced Multimodal Corpora Collection Tool. In *LREC.*

[27] Jonell Patrik, Bystedt Mattias, Fallgren Per, Dimosthenis Kontogiorgos, Lopes Jose, Malisz Zofia, Mascarenhas Samuel, Oertel Catharine, Raveh Eran, and Shore Todd. 2018. FARMI: A Framework for Recording Multi-Modal Interactions. In *LREC.*

[28] Hai Xuan Pham, Yuting Wang, and Vladimir Pavlovic. 2018. End-to-end Learning for 3D Facial Animation from Speech. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18).* ACM, New York, NY, USA, 361–365. https://doi.org/10.1145/3242969.3243017

[29] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence* 23, 10 (2001), 1175–1191.

[30] Najmeh Sadoughi and Carlos Busso. 2018. Novel realizations of speech-driven head movements with generative adversarial networks,âĂİ. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), Calgary, AB, Canada.* 6169–6173.

[31] Cristen Torrey, Aaron Powers, Matthew Marge, Susan R. Fussell, and Sara Kiesler. 2006. Effects of Adaptive Robot Dialogue on Information Exchange and Social Relations. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction (HRI '06).* ACM, New York, NY, USA, 126–133. https://doi.org/10.1145/1121241.1121264

[32] Frank M. F. Verberne, Jaap Ham, Aditya Ponnada, and Cees J. H. Midden. 2013. Trusting Digital Chameleons: The Effect of Mimicry by a Virtual Social Agent on User Trust. In *Persuasive Technology*, Shlomo Berkovsky and Jill Freyne (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 234–245.

[33] Wim Verkruysse, Lars O Svaasand, and J Stuart Nelson. 2008. Remote plethysmographic imaging using ambient light. *Opt. Express* 16, 26 (Dec 2008), 21434–21445. https://doi.org/10.1364/OE.16.021434

[34] Alessandro Vinciarelli, Maja Pantic, and HervÃľ Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27, 12 (2009), 1743 – 1759. https://doi.org/10.1016/j.imavis.2008.11.007 Visual and multimodal analysis of human spontaneous behaviour:.

[35] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. 2018. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. *arXiv preprint arXiv:1803.09017* (2018).

[36] Klaus Weber, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser, and Elisabeth André. 2018. How to Shape the Humor of a Robot - Social Behavior Adaptation Based on Reinforcement Learning. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18).* ACM, New York, NY, USA, 154–162. https://doi.org/10.1145/3242969.3242976

[37] Elena Wolf, Manuel Martinez, Alina Roitberg, Rainer Stiefelhagen, and Barbara Deml. 2018. Estimating mental load in passive and active tasks from pupil and gaze changes using bayesian surprise. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data.* ACM, 6.

[38] Leshao Zhang and Patrick G.T. Healey. 2018. Human, Chameleon or Nodding Dog?. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18).* ACM, New York, NY, USA, 428–436. https://doi.org/10.1145/3242969.3242998