Universität Augsburg
Fakultät für Angewandte Informatik

Dissertation zur Erlangung des akademischen Grades
Doktor der Informatik

COOPERATIVE AND TRANSPARENT MACHINE LEARNING FOR THE CONTEXT-SENSITIVE ANALYSIS OF SOCIAL INTERACTIONS

Universität
Augsburg
University

eingereicht von
M.Sc. Tobias Baur
im Mai 2018

Erstgutachterin:
Prof. Dr. Elisabeth André
Zweitgutachter:
Prof. Dr. Björn Schuller
Drittgutachterin:
Prof. Dr. Cristina Conati

ERSTGUTACHTERIN: Prof. Dr. Elisabeth André
ZWEITGUTACHTER: Prof. Dr. Björn Schuller
DRITTGUTACHTERIN: Prof. Dr. Cristina Conati
MÜNDLICHE PRÜFUNG: 24. Juli 2018

ZUSAMMENFASSUNG

Das Forschungsgebiet der Sozialen Signalverarbeitung eröffnet neue Wege virtuelle Agenten und soziale Roboter mit einem besseren Verständnis menschlicher Verhaltensweisen und der damit einhergehenden impliziten Nachrichten auszustatten. Damit Maschinen solche Verhaltensmerkmale verstehen und interpretieren können, kommen Techniken des Maschinellen Lernens zum Einsatz. In vielen Bereichen des Maschinellen Lernens werden statistische Modelle auf einer Vielzahl von annotierten Beispielen trainiert und ein Algorithmus angewandt um Muster zu erkennen und diese bestimmten Klassen oder Werten zuzuordnen. ML Techniken wie Künstliche Neuronale Netze haben sich hierbei als höchst effektiv für das Zuordnen und Identifizieren von Merkmalen zur Erkennung spezifischer Erkennungsprobleme herausgestellt. Jedoch haben diese auch einige Nachteile. Vor allem sind die Entscheidungen, die solche sogenannten "Black-Box" -Ansätze treffen, für Menschen nicht oder nur sehr eingeschränkt nachvollziehbar. Oft kann es bei solchen Modellen auch zu Problemen führen, wenn diese in einem anderen Kontext eingesetzt werden. Eine neue Forschungsrichtung, die sogenannte "eXplainable Artificial Intelligence" (XAI), beschäftigt sich damit, KI Systeme in der Lage zu versetzen, getroffene Entscheidungen erklären zu können. In dieser Arbeit werden Strategien untersucht, um die Erkennung und Interpretation von komplexen sozialen Signalen transparenter zu gestalten und dem Mensch mehr Kontrolle im Prozess des maschinellen Lernens zu ermöglichen.

Um ein besseres Verständnis zu bekommen, wie Menschen komplexe Soziale Signale interpretieren, wird zunächst ein Überblick über Erkenntnisse der Verhaltenspsychologie gegeben. Anschließend wird die Aufzeichnung unterschiedlicher Datenkorpora in verschiedenen Kontexten beschrieben, die gezielt Beispiele für solche natürlichen Verhaltensweisen in sozialen Interaktionen beinhalten sollen. Um zu verstehen, wie eine Maschine aus solchen Rohdaten Verhaltensweisen lernen kann, werden im Anschluss typische Techniken der Sozialen Signalverarbeitung und des Maschinellen Lernens eingeführt. Große und kontinuierliche Datenkorpora müssen annotiert und verwaltet werden. Hierfür wird ein neuartiges Tool mit dem Namen NOVA eingeführt, das es erlaubt, den Annotationsprozess auf mehrere Personen zu verteilen und Daten möglichst flexibel und effektiv zu annotieren. Durch das Tool wird bereits während des Annotationsvorgangs der Einsatz Maschinellen Lernens ermöglicht, um simultan mit der Maschine Daten zu annotieren. Durch diesen Ansatz, der als Kooperatives Maschinelles Lernen bezeichnet wird, wird der Prozess der Annotation beschleunigt und für Laien transparenter gestaltet. So können schnell viele Trainingsbeispiele generiert werden, so dass Modelle für einzelne Verhaltensmerkmale intuitiv mittels NOVA trainiert und verbessert werden können. Um komplexeres Verhalten, wie etwa Interesse einer Person während einer Unterhaltung oder emotionale Regulationsstrategien zu erkennen, wird ein hybrider Ansatz aus theoriebasierten, sowie datengetriebenen Verfahren angewandt. Hierbei werden die Ergebnisse einzelner Erkenner aus verschiedenen Modalitäten sowie die Dynamik der Interaktion und des Gesprächskontexts berücksichtigt. Abschließend werden die wichtigsten Beiträge der Arbeit zusammengefasst und ein Ausblick auf zukünftige Forschungsarbeiten gegeben.

ABSTRACT

The research area of Social Signal Processing paves the way for conversational companions, such as virtual agents or social robots, to become aware of nuances in our behaviours and implicit messages that come along with them. For machines to understand and interpret such behavioural cues, the state-of-the-art procedure is the application of various machine learning techniques. In many ML tasks, statistical models are trained on a large amount of annotated samples and an algorithm aims to match patterns that represent specific classes or values. ML tehniques, such as artificial neural networks, nowadays do pretty well in mapping and even identifying low level features to a specific recognition problem. A large drawback here is that the decisions they are making are not comprehensible and understandable to humans and that their assumptions are often wrong in changing contexts. Therefore a new research direction -"eXplainable Artificial Intelligence" (XAI)- identified the need of AI systems to be able to explain their decisions. In this thesis we investigate strategies to make the recognition and interpretation of complex social signals more transparent and explore ways to empower the human in the machine learning loop. To gain a better understanding of how humans interpret social cues, we first introduce an overview on results of behavioural psychology. We then describe the creation of various multi-person and muli-modal corpora in varying contexts that aim to induce multiple aspects of such behaviours. Next, we briefly introduce common techniques used in the area of social signal processing and machine learning. To successfully annotate and manage large continuous databases, a novel tool, named NOVA is presented. It allows to distribute the annotation task on multiple labellers and supports various types of annotations. NOVA further allows to take advantage of ML techniques already during the annotation process (a concept named cooperative machine learning). By employing CML, data is annotated simultaneously with the machine, which speeds up the annotation process and gives a more transparent idea of a machine's decisions. For inferring more complex behaviours, such as a person's conversational engagement or emotion regulation strategies, an approach is introduced that considers the predictions of multiple social cue recognisers and various types of context information. Finally, an outlook on future research directions is given.

PUBLICATIONS

Some ideas, examples and figures included in this thesis have appeared previously in parts of the following peer-reviewed publications:

Anderson, Keith, Elisabeth André, Tobias Baur, Sara Bernardini, Matthieu Chollet, Evi Chryssafidou, Ionut Damian, et al. (2013). "The TARDIS Framework: Intelligent Virtual Agents for Social Coaching in Job Interviews." In: *Advances in Computer Entertainment - 10th International Conference, ACE 2013, Boekelo, The Netherlands, November 12-15, 2013. Proceedings*. Ed. by Dennis Reidsma, Haruhiro Katayose, and Anton Nijholt. Vol. 8253. Lecture Notes in Computer Science. Springer, pp. 476–491.

Baur, Tobias, Ionut Damian, Patrick Gebhard, Kaska Porayska-Pomsta, and Elisabeth André (2013). "A Job Interview Simulation: Social Cue-Based Interaction with a Virtual Character." In: *International Conference on Social Computing, SocialCom 2013, SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013, Washington, DC, USA, 8-14 September, 2013*. IEEE Computer Society, pp. 220–227.

Baur, Tobias, Ionut Damian, Florian Lingenfelser, Johannes Wagner, and Elisabeth André (2013). "NovA: Automated Analysis of Nonverbal Signals in Social Interactions." In: *Human Behavior Understanding - 4th International Workshop, HBU 2013, Barcelona, Spain, October 22, 2013. Proceedings*. Ed. by Albert Ali Salah, Hayley Hung, Oya Aran, and Hatice Gunes. Vol. 8212. Lecture Notes in Computer Science. Springer, pp. 160–171.

Baur, Tobias, Gregor Mehlmann, Ionut Damian, Florian Lingenfelser, Johannes Wagner, Birgit Lugrin, Elisabeth André, and Patrick Gebhard (2015). "Context-Aware Automated Analysis and Annotation of Social Human-Agent Interactions." In: *TiiS* 5.2, 11:1–11:33.

Baur, Tobias, Dominik Schiller, and Elisabeth André (2017). "Modeling User's Social Attitude in a Conversational System." In: *Emotions and Personality in Personalized Services - Models, Evaluation and Applications*. Ed. by Marko Tkalcic, Berardina De Carolis, Marco de Gemmis, Ante Odic, and Andrej Kosir. Human-Computer Interaction Series. Springer, pp. 181–199.

Cafaro, Angelo, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres, Catherine Pelachaud, Elisabeth André, and Michel F. Valstar (2017). "The NoXi database: multimodal recordings of mediated novice-expert interactions." In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, Glasgow, United Kingdom, November 13 - 17, 2017*. Ed. by Edward

Lank, Alessandro Vinciarelli, Eve E. Hoggan, Sriram Subramanian, and Stephen A. Brewster. ACM, pp. 350–359.

Damian, Ionut, Tobias Baur, and Elisabeth André (2013). "Investigating Social Cue-Based Interaction in Digital Learning Games." In: *Proceedings of the 1st International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI 2013) held in conjunction with the 8th Foundations of Digital Games 2013 (FDG), ACM, SASDG Digital Library, Chania, Crete, Greece*.

Damian, Ionut, Tobias Baur, and Elisabeth André (2016). "Measuring the impact of multimodal behavioural feedback loops on social interactions." In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016*. Ed. by Yukiko I. Nakano, Elisabeth André, Toyoaki Nishida, Louis-Philippe Morency, Carlos Busso, and Catherine Pelachaud. ACM, pp. 201–208.

Damian, Ionut, Tobias Baur, Patrick Gebhard, Kaśka Porayska-Pomsta, and Elisabeth André (2013). "A Software Framework for Social Cue-Based Interaction with a Virtual Recruiter." In: *Intelligent Virtual Agents*, p. 444.

Damian, Ionut, Tobias Baur, Birgit Lugrin, Patrick Gebhard, Gregor Mehlmann, and Elisabeth André (2015). "Games are Better than Books: In-Situ Comparison of an Interactive Job Interview Game with Conventional Training." In: *Artificial Intelligence in Education - 17th International Conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings*. Ed. by Cristina Conati, Neil T. Heffernan, Antonija Mitrovic, and M. Felisa Verdejo. Vol. 9112. Lecture Notes in Computer Science. Springer, pp. 84–94.

Damian, Ionut, Tobias Baur, Chiew Seng Sean Tan, Johannes Schöning, Kris Luyten, and Elisabeth André (2014). "Towards Peripheral Feedback-based Realtime Social Behaviour Coaching." In: *NordiCHI 2014 workshop on Interactions and Applications on See-through Technologies (Helsinki, Finland)*.

Damian, Ionut, Chiew Seng Sean Tan, Tobias Baur, Johannes Schöning, Kris Luyten, and Elisabeth André (2014). "Exploring social augmentation concepts for public speaking using peripheral feedback and real-time behavior analysis." In: *IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2014, Munich, Germany, September 10-12, 2014*. IEEE Computer Society, pp. 261–262.

Damian, Ionut, Chiew Seng Sean Tan, Tobias Baur, Johannes Schöning, Kris Luyten, and Elisabeth André (2015). "Augmenting Social Interactions: Realtime Behavioural Feedback using Social Signal Processing Techniques." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*. Ed. by Bo Begole, Jinwoo Kim, Kori Inkpen, and Woontack Woo. ACM, pp. 565–574.

Gebhard, Patrick, Tobias Baur, Ionut Damian, Gregor Mehlmann, Johannes Wagner, and Elisabeth André (2014). “Exploring interaction strategies for virtual characters to induce stress in simulated job interviews.” In: *International conference on Autonomous Agents and Multi-Agent Systems, AAMAS ’14, Paris, France, May 5-9, 2014*. Ed. by Ana L. C. Bazzan, Michael N. Huhns, Alessio Lomuscio, and Paul Scerri. IFAAMAS/ACM, pp. 661–668.

Gebhard, Patrick, Tanja Schneeberger, Elisabeth André, Tobias Baur, Ionut Damian, Gregor Mehlmann, Cornelius König, and Markus Langer (2018). “Serious Games for Training Social Skills in Job Interviews.” In: *IEEE Transactions on Games*.

Gebhard, Patrick, Tanja Schneeberger, Tobias Baur, and Elisabeth André (2018). “MARSSI: Model of Appraisal, Regulation, and Social Signal Interpretation.” In: *17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS2018), Sweden*.

Jones, Hazaël, Nicolas Sabouret, Ionut Damian, Tobias Baur, Elisabeth André, Kaska Porayska-Pomsta, and Paola Rizzo (2014). “Interpreting social cues to generate credible affective reactions of virtual job interviewers.” In: *CoRR* abs/1402.5039.

Mancini, Maurizio, Laurent Ach, Emeline Bantegnie, Tobias Baur, Nadia Berthouze, Debajyoti Datta, Yu Ding, Stéphane Dupont, Harry J Griffin, Florian Lingenfelser, et al. (2013). “Laugh When You’re Winning.” In: *Innovative and Creative Developments in Multimodal Interaction Systems - 9th IFIP WG 5.5 International Summer Workshop on Multimodal Interfaces, eNTERFACE 2013, Lisbon, Portugal, July 15 - August 9, 2013. Proceedings*. Ed. by Yves Rybarczyk, Tiago Cardoso, João Rosas, and Luis M. Camarinha-Matos. Vol. 425. IFIP Advances in Information and Communication Technology. Springer, pp. 50–79.

Mehlmann, Gregor, Markus Häring, Kathrin Janowski, Tobias Baur, Patrick Gebhard, and Elisabeth André (2014). “Exploring a Model of Gaze for Grounding in Multimodal HRI.” In: *Proceedings of the 16th International Conference on Multimodal Interaction, ICMI 2014, Istanbul, Turkey, November 12-16, 2014*. Ed. by Albert Ali Salah, Jeffrey F. Cohn, Björn W. Schuller, Oya Aran, Louis-Philippe Morency, and Philip R. Cohen. ACM, pp. 247–254.

Mehlmann, Gregor, Kathrin Janowski, Tobias Baur, Markus Häring, Elisabeth André, and Patrick Gebhard (2014). “Modeling Gaze Mechanisms for Grounding in HRI.” In: *ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014)*. Ed. by Torsten Schaub, Gerhard Friedrich, and Barry O’Sullivan. Vol. 263. Frontiers in Artificial Intelligence and Applications. IOS Press, pp. 1069–1070.

Niewiadomski, Radoslaw, Jennifer Hofmann, Jérôme Urbain, Tracey Platt, Johannes Wagner, Bilal Piot, Huseyin Cakmak, Sathish Pammi, Tobias Baur, Stephane Dupont, et al. (2013). "Laugh-aware virtual agent and its impact on user amusement." In: *International conference on Autonomous Agents and Multi-Agent Systems, AAMAS '13, Saint Paul, MN, USA, May 6-10, 2013*. Ed. by Maria L. Gini, Onn Shehory, Takayuki Ito, and Catholijn M. Jonker. IFAAMAS, pp. 619–626.

Niewiadomski, Radoslaw, Maurizio Mancini, Tobias Baur, Giovanna Varni, Harry J. Griffin, and Min S. H. Aung (2013). "MMLI: Multimodal Multiperson Corpus of Laughter in Interaction." In: *Human Behavior Understanding - 4th International Workshop, HBU 2013, Barcelona, Spain, October 22, 2013. Proceedings*. Ed. by Albert Ali Salah, Hayley Hung, Oya Aran, and Hatice Gunes. Vol. 8212. Lecture Notes in Computer Science. Springer, pp. 184–195.

Porayska-Pomsta, Kaska, Anderson Keith, Ionut Damian, Tobias Baur, Elisabeth André, Sara Bernardini, and Paola Rizzo (2013). "Modelling Users' Affect in Job Interviews: Technological Demo." In: *User Modeling, Adaptation, and Personalization - 21th International Conference, UMAP 2013, Rome, Italy, June 10-14, 2013, Proceedings*. Ed. by Sandra Carberry, Stephan Weibelzahl, Alessandro Micarelli, and Giovanni Semeraro. Vol. 7899. Lecture Notes in Computer Science. Springer, pp. 353–355.

Porayska-Pomsta, Kaska, Paola Rizzo, Ionut Damian, Tobias Baur, Elisabeth André, Nicolas Sabouret, Hazaël Jones, Keith Anderson, and Evi Chryssafidou (2014). "Who's Afraid of Job Interviews? Definitely a Question for User Modelling." In: *User Modeling, Adaptation, and Personalization - 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings*. Ed. by Vania Dimitrova, Tsvi Kuflik, David Chin, Francesco Ricci, Peter Dolog, and Geert-Jan Houben. Vol. 8538. Lecture Notes in Computer Science. Springer, pp. 411–422.

Ritschel, Hannes, Tobias Baur, and Elisabeth André (2017). "Adapting a Robot's linguistic style based on socially-aware reinforcement learning." In: *26th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2017, Lisbon, Portugal, August 28 - Sept. 1, 2017*. IEEE, pp. 378–384.

Urbain, Jérôme, Radoslaw Niewiadomski, Jennifer Hofmann, Emeline Bantegnie, Tobias Baur, Nadia Berthouze, Hüseyin Cakmak, Richard Thomas Cruz, Stéphane Dupont, Matthieu Geist, et al. (2013). "Laugh machine." In: *Proceedings eNTERFACE* 12, pp. 13–34.

Valstar, Michel F., Tobias Baur, Angelo Cafaro, Alexandru Ghitulescu, Blaise Potard, Johannes Wagner, Elisabeth André, Laurent Durieu, Matthew Aylett, Soumia Dermouche, et al. (2016). "Ask Alice: an artificial retrieval of information agent." In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016*. ACM, pp. 419–420.

Wagner, Johannes, Tobias Baur, Yue Zhang, Michel F. Valstar, Björn W. Schuller, and Elisabeth André (2018). “Applying Cooperative Machine Learning to Speed Up the Annotation of Social Signals in Large Multi-modal Corpora.” In: *CoRR* abs/1802.02565.

Wagner, Johannes, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth André (2013). “The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time.” In: *ACM Multimedia Conference, MM ’13, Barcelona, Spain, October 21-25, 2013*. Ed. by Alejandro Jaimes, Nicu Sebe, Nozha Boujemaa, Daniel Gatica-Perez, David A. Shamma, Marcel Worring, and Roger Zimmermann. ACM, pp. 831–834.

"All that is gold does not glitter,
Not all those who wander are lost.
The old that is strong does not wither,
Deep roots are not reached by the frost."

— J.R.R. Tolkien

ACKNOWLEDGMENTS

I would like to thank all the people that have supported me during the process of writing this thesis. My special thanks goes to Prof. Dr. Elisabeth André for supervising this thesis, especially for her help and advice. I would also like to thank Prof. Dr. Björn Schuller and Prof. Dr. Cristina Conati for consenting to volunteer as reviewers. It is an honour that such experienced and well recognised researchers took their time to consider this dissertation.

Further I would like to thank all my current and former colleagues at the Human Centred Multimedia lab for the friendly and respectful atmosphere. A special thanks goes to my colleagues Johannes Wagner, Florian Lingenfelser and Dominik Schiller for the fruitful discussions and collaboration on countless projects. I'd also like to thank Ionut Damian for the time we shared on the TARDIS project when we both started our PhD, and the lecture we gave together on multi-modal interaction and interactive machine learning.

Of course to a huge amount I like to thank my family and friends for their support. To my parents who constantly supported me. To my good friend Kuzman for the inspirational discussions. Last but not least to my partner in life Melissa, who endured me when I was stressed out and desperate at so many points, for her help with many of the illustrations in this thesis and for everything. I can't thank you enough.

CONTENTS

MOTIVATION AND BACKGROUND

MULTI-MODAL AND MULTI-PERSON DATA COLLECTION AND SOCIAL SIGNAL PROCESSING

COOPERATIVE MACHINE LEARNING FOR SPEEDING UP THE ANNOTATION OF CONTINUOUS MULTI-MODAL DATA

EXPLAINABLE AND CONTEXT-SENSITIVE MODELLING OF COMPLEX SOCIAL SIGNALS

CONTRIBUTIONS

APPENDIX

LIST OF FIGURES

LIST OF TABLES

MOTIVATION AND BACKGROUND

1 INTRODUCTION

"Before we teach computers to love,
maybe we should teach them about personal boundaries"

— Joey Comeau, 2014

1.1 Motivation

When humans communicate with each other, they exchange information not only by written or spoken language. In many cases it is by far more important *how* and *under which circumstances* a message is communicated, rather than *what* was actually said.

To give an example: Imagine a close relative entering the room and asking a question like: *"Did you empty the milk carton?"*. The content of the message is actually not that relevant because obviously now the milk is empty. If the person finding the empty milk carton makes an angry face and is asking that question with a rather aggressive tone of voice, he or she communicates: *"I wanted to have some milk and you drank it up, so now I'm mad at you"*. The message here crucially differs from the literal spoken content. But maybe the person finding the empty milk carton says it with a friendly tone of voice that communicates something like: *"Ok, the milk carton is empty, let's put it on the shopping list"*. Given the same content, the message in both cases is a different one. In addition, the style the message is conveyed reveals a different social attitude towards the interlocutor - in the first case the tendency is rather negative, in the second case still quite positive.

In contrary to human-human communication, computers –as we use them today– are usually ignorant of a user's social attitude towards them. Even modern devices only care about explicit input - either from traditional peripherals, such as mouse or keyboard, as well as speech commands or gesture-based input. At the same time, an early study by Reeves and Nass (1996) revealed that people tend to show a social attitude towards computer systems. That means - even though it might be on an unconscious level - humans seem to respond socially to computer systems in a similar way as they would to human interlocutors.

The need for computer systems to "understand" the user's affective reactions and emotions in some way, has been widely discussed since the mid 1990s. Presumably, pioneer work in this area was Rosalind Picard's book *"Affective Computing"* from 1995 (Picard, 1995). In the following years new sensory devices appeared on the market, starting with high resolution cameras and wearable devices, leading to motion capture systems, daily smart wearables and other state-of-the art devices which pushed the field of computational emotion recognition. To mention an example, early work in this area by André and colleagues (Wagner, Kim, and André, 2005; Vogt, André, and Bee, 2008) had a great impact on the field of research (as well as on this thesis). It took until the year 2007 when Alex Sandy Pentland coined the term of *"social signal processing"* (Pentland, 2007) which established as a new area of research. A decent and well-known overview on the topic of social signal processing is given in the survey by Vinciarelli, Pantic, and Bourlard (2009). Since then, sensory devices and detection algorithms of course improved even further and new problems and challenges arose, such as online processing or the fusion of multiple modalities. To resolve such problems, advanced software solutions like the *Social Signal Interpretation* (SSI) framework (Wagner, Lingenfelser, et al., 2013) were developed.

One might argue that a computer system that analyses implicitly conveyed social cues of humans engaged in tasks, such as browsing the web or creating documents, would be rather disturbing. However, the need to emulate certain aspects of human-like social behaviour becomes more apparent in scenarios with virtual agents or humanoid robots. Often such systems are used in social settings where they replace or assist a coach, a medical practitioner or a caregiver. It is very likely that this process will be even more common in the future. Typical use-cases include negotiation-practice scenarios (Traum et al., 2012), psychotherapy (Kang et al., 2012), job interviews (Baur, Mehlmann, et al., 2015; Hoque et al., 2013) or elderly care (Broekens, Heerink, and Rosendal, 2009). It is obvious that in such scenarios, a basic "understanding" –in a sense of correctly interpreting the users social cues– would be desirable.

In recent years, progress has been made in teaching computers and robots specific tasks with several computational learning algorithms. The Defense Advanced Research Projects Agency (DARPA) identified three waves of artificial intelligence evolutions. In the first phase, mostly human-created knowledge was used to build rule-based systems. That way, problems that were intellectually difficult for humans could be solved using a computer's fast processing power. In the second phase, in which we are up to this date, statistical learning was introduced. The main idea here is that engineers create statistical models for specific problems, based on huge amounts of data. This mostly includes tasks which appear to be easily solvable by humans, but whose rules are difficult to be described in a mathematical way. Such tasks include the automated recognition of speech, or the detection of facial expressions. A trending topic nowadays are artificial neural networks and deep learning. That is mostly because they outpaced other methodologies in specific recognition tasks e.g. in computer vision-based object detection (e.g. Ciresan et al., 2012) and automated speech recognition (e.g. Weninger et al., 2015). Statistical models yet come with a couple of drawbacks. For example, they require huge amounts of data and processing power for training. The data itself and the according labels are often problematic. In data-hungry recognition tasks, such as deep learning, algorithms often rely on poorly labelled data from social media channels and other sources that were not intended for the recognition problem at hand. Another big drawback of statistical models is that they lack ways of representing causal relationship, performing logical inferences and are mostly so called *"black-box"* approaches. That means, while they are extremely good at learning and predicting on huge amounts of data, they only deliver little or no ways of comprehending their decisions. As a human, in many cases it is of vast importance to comprehend and understand decisions a system is making. For instance, when we think about self-driving cars that may have to decide if they kill the driver or a pedestrian, we would expect that an AI must be capable of explaining itself (Goodall, 2016). As we are approaching a state where autonomous systems "perceive, learn decide and act on their own" (Gunning, 2017) without any way to comprehend their decisions, DARPA suggests that a future third wave of AI should be able to contextually adapt and to be able to explain its decisions. This new research trend is often summarised under the term *"eXplainable Artificial Intelligence"* (XAI). XAI aims to find methods that enable humans to understand, trust and manage future AI systems. There are several sub-areas that address the issue of making AI more explainable.

For example, the area of *"interactive machine learning"* (Amershi, Cakmak, et al., 2014) identified the benefits of involving humans in the process of machine learning on several abstraction layers. Foremost, the potential of machine learning should not be limited to experts, but rather should be made available to a broad group of people, by developing interfaces and tools that allow them to interactively contribute to the process while at the same time they get a clear and transparent view on the processes and decisions the machine is making.

The aspect of contextual-adaptation has been only rarely considered up to this date. Yet when we think of analysing human behaviours, as humans we don't perceive others communicative signals in isolation, but rather we consider the contextual surroundings. That is for example, the relation with the person we are talking to, the common knowledge we share or cultural aspects. People are very diverse and sometimes mean different things when talking about the same content. Current statistical models trying to interpret the actual meaning of a communicative signal will make rather naive assumptions. To give an example here: state-of-the-art algorithms that are trained to recognise emotions from facial expressions will always interpret a smile as a sign of happiness and a frown as a sign of sadness. There might be situations where this works perfectly fine, but there are many use cases where this scheme won't fit. Imagine a job interview where the interviewer is asking a harsh question. The interviewee might be embarrassed and smile as he or she is overplaying the situation. In this case current systems would still recognise this person as happy, which most probably is not the case here.

A main contribution of this thesis is to provide a framework for future research to consider aspects, such as transparency, comprehensibility and context-sensitivity in the process of recognising and analysing complex social behaviours, both in human to human and human to agent interactions. The next section describes the research objectives of this thesis in more detail.

1.2 Research Objectives

To address the aspects elaborated in the previous section, this thesis focuses primarily on including the human in the process of machine learning for the recognition of social signals. Nowadays, this is mostly restricted to experts in the field, yet researchers from multiple disciplines could profit by incorporating machine-aided techniques in their daily work-flow. Especially the transparency of computational models allows non-machine learning experts to understand and comprehend the decisions of such a model. This thesis proposes concepts and tools to extend the accessibility of such techniques. This includes gathering new rich and useful data containing samples of various phenomena in multi-person and multi-modal interactions, support to identify such phenomena in the data in convenient ways and tools to automatically interpret complex multi-modal observations. In more detail, the main research objectives addressed in this thesis are:

- **Creation of multi-modal, multi-person corpora, containing social phenomena in various contexts.**

  For recognising social signals in an automated manner we first require relevant data that contains those social phenomena. For example, the success of artificial neural networks in image classification is to a vast amount based on the availability of masses of labelled images as training data. In the research area of social signal processing, gathering adequate data is more problematic, as human social signals are more complex, context-dependent and vary largely between individuals. For achieving adequate classification rates, models are often trained on acted samples, or the few available natural corpora are reused for recognition problems they were not intended to be used for. When such models are applied in another context, or under non-perfect conditions they often don't work as precise any more. Various researchers in the field conclude that non-acted corpora with realistic behaviours are essential for the success of the whole field of social signal processing. Behaviours should always be considered in context, e.g. in interactions between two or more participants. In this thesis, we therefore propose concepts and tools for collecting data of multiple modalities from multiple persons in synchronised non-acted interactions.

To this end, we describe the process of creating a large multimodal and multi-person corpus named NOXI which has been made available to, and is widely used by, the research community. Further, we briefly introduce two additional corpora concerned with other contexts that make use of the proposed techniques.

- **Conception and Implementation of a collaborative open-source annotation tool**

  Once the raw data of a new corpus has been recorded, we want to analyse it in order to identify relevant or interesting behaviours. To improve machine learning algorithms, and other data analysis tasks, experts most often annotate relevant parts in the recordings to apply statistical methods to them. The annotation process, especially for non-acted, natural, and emotional data is a time consuming task and can go beyond a reasonable scope in terms of time, cost and effort.

  As one main contribution of this thesis, we suggest a collaborative annotation and analysis tool named NOVA, that allows researchers of multiple disciplines to annotate continuous data in a more convenient way and to share and organise various types of annotations. When the workload is split between multiple raters, other problems need to be handled, such as the organisation of annotations and data, inter-rater agreement and rater-reliability questions and the management of annotation tasks. The NOVA tool addresses theses issues by connecting multiple annotators with a shared database and multiple functions to manage data and annotations, to find inter-rater correlations and merge annotations from multiple raters.

  NOVA has been made publicly available and is successfully employed in multiple international research projects.

- **Integration of cooperative machine learning to speed up the annotation process and to make machine learning more transparent**

  Given the circumstance that we strongly focus on annotating data for training machine learning models to classify and predict social signals, an interesting question is: Why don't we make use of machine learning algorithms already during the early annotation steps, to 1) support annotators with their task at hand and 2) make the process of training machine learning models more transparent to non-expert users? We directly equipped the NOVA tool with capabilities to train models at various stages of the annotation process. In a session completion step, a model is trained on manual annotations of the first few minutes of a continuous recording, while the remaining part is annotated by a machine annotator that uses this session-specific model. The machine annotator actively highlights the segments it is not entirely confident of. A human rater then only has to correct these parts. This step can be repeated arbitrarily often, so that the models predictions and their improvements become more comprehensible and transparent to the user, while at the same time the manual annotation effort is drastically reduced. Once continuous recordings are fully annotated, a more general model may be trained on multiple recordings and be used to predict completely unlabelled sessions.

We evaluated this process with the problem of detecting spoken utterances from fillers, breath sounds and silence, concluding that more than a third of the required time, when starting with a blank dataset can be saved, compared to manual annotation. We further conclude that when applying a hybrid approach, where a human annotator decides if the model should predict whole sessions with a general model or should created a person-dependent model for new sessions, even more time can be saved and predictions become more accurate.

While similar considerations exist in previous work, for the first time, we included such strategies in a tool that can be used by non-experts for multiple recognition tasks. Feature-Sets and classification models can easily be extended using general interfaces, allowing CML strategies to be applied to almost any social signal recognition problem.

- **Context-sensitive analysis and prediction of complex social behaviours**

  Complex social phenomena, such as emotions or social attitudes, are more complicated to be analysed in an automated matter than single social cues, as we need to consider multiple aspects at the same time. In many computational models nowadays, the presence of single social cues is interpreted in a straightforward manner as the presence or absence of an emotion. For example, most models predict the presence of a smile as an indicator for happiness, while in reality, people also smile in other situations, for example to overplay embarrassment or simply to appear friendly towards another person.

In other research areas that are concerned with analysing human behaviours, such as behavioural psychology, various models and theories exist that aim to explain multiple aspects of complex social behaviours. Researchers have been studying these phenomena in various contexts throughout the past century and therefore, we argue for an interdisciplinary approach by incorporating important findings and theories when building computational models. We give a literature overview on important results and theories at the beginning of this thesis that focus on complex phenomena, such as conversational engagement or emotion regulation strategies.

In most of today's machine learning and fusion approaches, expert knowledge can only be considered in a very limited way, for example with the selection of features and relevant channels. Once these inputs are selected, a classifier uses internal logic and rules to map training samples to classes or values. If we want to validate existing theories, or even verify our own theories about behaviours, a transparent model that allows to simulate interventions for finding correlations between single social cues, context information, interaction dynamics and complex phenomena should be preferred.

In this thesis we investigate the combination of "black box" learning algorithms, to map low level features to abstractions of social cues, and "white box" models to investigate correlations between multiple of these cues and their relation to complex phenomena. Here we not only consider social cues of a single person, but of multiple interaction partners. To this end we employ dynamic Bayesian networks for combining multiple social cues and context information which provides several advantages. The structure of the graph may be modelled using theory-based expert knowledge while probabilities and weights in the network may also be learned from data using machine learning, resulting in a cooperative work-flow of human-intelligence and computational power.

To this end, we introduce a tool chain that allows researchers to train the parameters of a DBN based on (semi-)automatically generated annotations. We further investigate and illustrate the process with two concrete use-cases of complex phenomena. First, we propose a model of conversational engagement that takes observations of multiple interlocutors and multiple types of context information into account. In the second use-case we combine our approach with a emotion-simulation component to generate explanations on recognised emotion regulation strategies for the emotion "shame".

1.3 Structure of the Thesis

This thesis is outlined as follows:

- **Chapter 2** introduces basic theories on human non-verbal communication. If we want to give our computer system "emotional intelligence" we first need to be aware of what emotional intelligence is and how it is evoked in humans. For decades, psychologists have been studying non-verbal communication with all its facets, including social cues and their relation to emotions and social attitudes. This chapter gives an overview on relevant terms and theories that we include in later chapters for the computer-based recognition and interpretation of complex behaviours.

- **Chapter 3** discusses challenges and solutions for designing and recording multi-modal corpora with multiple participants. For building robust recognition models that are applied to analyse a person's behaviours in an interaction, a considerable amount of natural and non-acted training data is necessary. Especially if we want to analyse behaviours with regard to context, we require data recorded in adequate and realistic scenarios. We exemplify the complexity of the process with a database that has been recorded to analyse multiple aspects of expert-novice interactions, such as unexpected situations and the engagement of interlocutors. We further give an overview on two additional databases in the context of laughter-inducing tasks and job interview situations.

- **Chapter 4** briefly introduces common social signal processing and machine learning techniques. Sensor devices are the eyes and ears of an agent (or any other affective computer system) and algorithms that give meaning to what was seen or heard build the foundation for further processing steps towards the detection of complex behaviours. Additionally the Social Signal Interpretation framework is introduced in this chapter. It is a practical approach to perform processing and machine learning steps in real-time applications.

- **Chapter 5** is concerned with the annotation of continuous recordings of multi-modal data. For understanding and analysing behaviours from raw data, in various research disciplines, it is common practice to identify and label such behaviours. Also, supervised machine learning algorithms cannot learn from raw, unlabelled data but require annotated training examples first. Therefore data has to be manually coded which can be a time-consuming task. This chapter introduces NOVA, a novel open-source a tool that allows to annotate and analyse continuous multi-modal recordings using multiple types of annotations. It offers capabilities to collaborate on huge databases and (semi-) automated labelling using cooperative machine learning techniques.

- **Chapter 6** describes our approach to *cooperative machine learning*. The main idea behind our approach is that a human rater labels a part of a continuous data recording and, at some point, hands the task over to a machine annotator. The machine predicts the remaining unlabelled parts of the recording and highlights those segments for which a human should investigate. The human annotators then can correct or confirm a subset of the machine's highlighted predictions and repeat the step arbitrarily often until they decide to trust the model. This way, the amount of time for labelling data can be drastically reduced and at the same time a models predictions become more transparent to the human labeller. This chapter further presents an evaluation of this approach with an exemplary recognition problem and describes a work-flow of its implementation in the NOVA tool.

- **Chapter 7** introduces our approach to combine the results of previous chapters with dynamic Bayesian networks. DBNs are employed as modelling framework for inferring social attitudes and emotions, based on concurrent multi-modal and interpersonal social cues and various kinds of context information in a transparent way. Additionally we exemplify the combination of the proposed tools with a model of conversational engagement in dyadic expert-novice interactions and further give an outlook on how advanced appraisal and regulation emotion models may be combined with such an approach.

- **Chapter 8** concludes with the contributions and most important results of the thesis and describes possible directions and chances concerning possible future work in this research area.

2 NON-VERBAL COMMUNICATION

"Non-verbal communication forms a social language that is in many ways richer and more fundamental than our words."

— Leonard Mlodinow, 2012

A main objective of this thesis is to create a framework for building transparent and context-sensitive computational models to recognise complex social signals in a way that is comprehensible to humans. In order to be able to create such models, in this chapter, we introduce important terms and theories related to non-verbal social cues, emotions and social attitudes.

2.1 The Importance of Non-verbal Communication

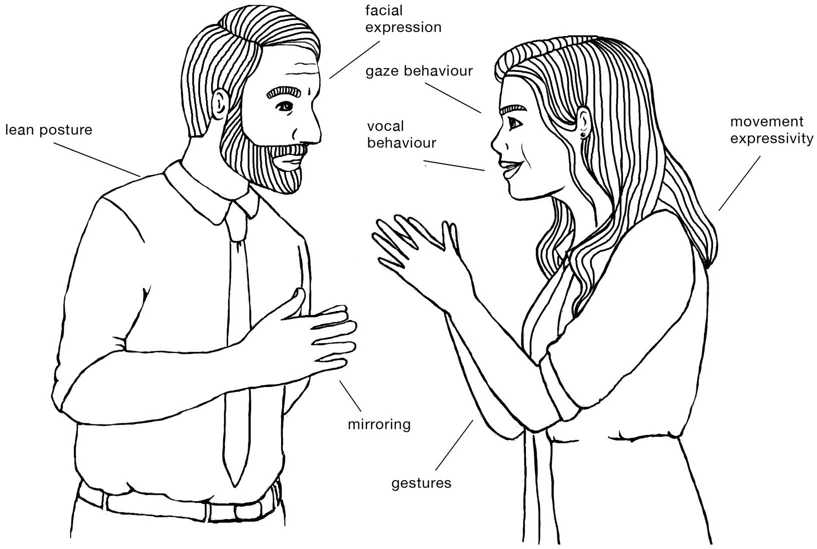


Figure 1.: A variety of social signals is communicated in social interactions

In a conversation between humans, information is not exclusively shared in an explicit manner. On the contrary, a myriad of implicit social signals (see Figure 1) is communicated. Studies suggest that such implicit information may have a higher impact on the outcome of a conversation than the actual content of the message itself. The classical example found in literature is Albert Mehrabian's study (Mehrabian et al., 1971), which claims that the spoken content of an utterance contributes 7% to its success while the impact of the vocal signals (conveyed by the nuances of voice) and the non-verbal signals (including facial expressions) is much higher with 38% and 55% (in the context of job interviews). While Mehrabian himself did not insist on the universality of the numbers that resulted from his study, (he pointed out many limitations of these experiments that often were generalised by other researchers), it boosted the field of behavioural analysis. Mehrabian's initial findings were backed up by a couple of follow-up studies that show non-verbal signals play a huge role in daily conversations and determine a large part of the way we interact with other humans and even animals* .

**studies by Bekoff (2013) suggest that domestic dogs became experts in reading human emotions by analysing our facial expressions*

The vision of a computer "understanding" emotions and non-verbal nuances, in the same way humans would, exists for quite a while now and due to the progress of technology, the research area of social signal processing is now closer than ever to achieve this goal. Before putting things into a more technical perspective in Chapter 4, the next sections aim to give an overview on some of the most important terms and theories about non-verbal behaviour analysis. As we aim to understand what complex behaviours consist of and how we can recognise them, we will first have a look at communicative social signals before we move on to emotions and social attitudes.

The chapter is structured as follows:

- In Section 2.2 we define relevant terms and categories of social signals and introduce findings and theories concerning their appearance in single modalities.
- In Section 2.3 we discuss social cues and phenomena that can be observed in reciprocal exchange only. Such cues will be called interpersonal cues.
- In Section 2.4 we give an overview on common emotion models and their relation to observed social cues. We further discuss concepts such as appraisals and emotion regulation strategies.
- In Section 2.5 we review theories on social attitudes and focus especially on conversational engagement and related concepts.

2.2 Social Signals and Social Cues

Social signals are defined by Poggi and D'Errico (2010) as "communicative and informative signals that directly or indirectly provide information through social interactions, social emotions, social behaviour, and social relationships". As mentioned before, our social communication includes a large variety of such signals, both verbal and non-verbal. Due to their dominant impact we will focus on non-verbal communicative signals here. Kendon, Sebeok, and Umiker-Sebeok (1981) define the term *non-verbal communication* as "most frequently used to refer to all of the ways in which communication is effected between persons when in each others' presence, by means other than words. It refers to the communicational functioning of bodily activities, gestures, facial expressions and orientation, postures and spacing, touch and smell, and of those aspects of utterance that can be considered apart from the referential content of what is said."

While the *verbal channel* is mostly used for the exchange of information, the non-verbal channels original function is to handle interpersonal relationships. Beyond, it may also be used as a replacement of spoken language. The discipline of *behaviour analysis* approaches the interpretation of the actual content of non-verbal messages. This content may also differ (consciously or unconsciously) from the verbal message. If the verbal messages are contrary to the non-verbal ones, an attentive observer will realise that "something is wrong" (e.g., the person is hiding something or is lying). Furthermore, social signals may be used purposely in various situations to reinforce content of spoken messages, or to clarify the role allocation between conversation partners.

An important aspect for the interpretation of social cues is that they must never be viewed in isolation, but rather within the context of the overall complex*. To make inferences about the "hidden" message (which may be an emotion, or an attitude towards the interlocutor) it is of vast significance to have a look at the combination of multiple cues happening simultaneously or in sequences. Figure 2 is meant to illustrate that. The dominant social cue of interest is a gesture one could call "hand supports face". Having a closer look, one can see the pointing finger is directed upwards, and the thumb is supporting the chin. The other arm is crossed defensively in front of the body and the head is tilted.

**complex, in a sense of the entirety of social cues*

The complex of social cues could say something like "I'm thinking this through, but I do not agree with what you are telling me." Overall one could speak of a critical evaluation gesture.

This figure may also be used to illustrate the so called *congruence* between verbal and non-verbal messages. In case the man in this picture claims that he doesn't agree with his interlocutor, both his verbal and non-verbal channels are *congruent*. If he argues the converse, both channels are *contradicted*. Due to their strong impact, when it comes to incongruence between both channels, people are more likely to believe the non-verbal channel (Pease, 1988).

Besides the classification in congruent vs. non-congruent behaviours, Scheflen (1964) distinguishes categories of inclusive vs. non-inclusive and vis-a-vis vs. parallel behaviours. We speak of inclusive behaviours when people are, for example, applying open body postures, while crossed arms used as a barrier are excluding. In a conversation the position of the interlocutors are either vis-a-vis, so that they look at

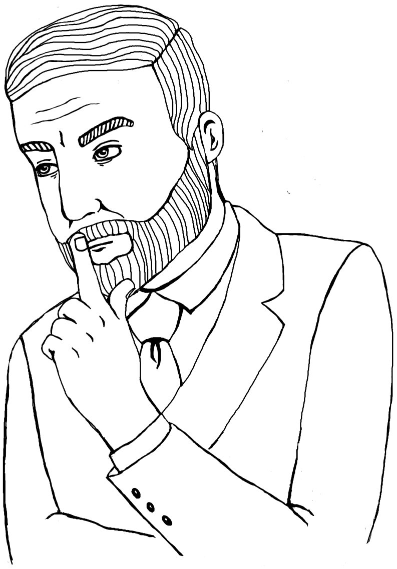


Figure 2.: An example for a complex of critical social cues.

each other, or (partly) parallel. In this case a person is most probably focusing on another object or person, or is in a hurry to get away.

Non-verbal behaviours may have various meanings depending on the context. In the example from Figure 2 we implicitly assumed that the man is a critical listener and inferred, based on the complexity of his behaviours that he doesn't share his counterpart's opinion. Another possibility would be that the man in the picture sits in the waiting room of a medical practice, holding his arm in front of his stomach because he is in pain. He's supporting his head because he has been waiting for quite some time and his critical gaze goes towards the clock because he's getting impatient. By changing the context, the point of view on his behaviours has completely changed. Due to the vast importance of context for social signal analysis, we will discuss general considerations about context information separately in Section 7.2.

Social Cues have been categorised following various schemes. For example, Argyle (2013) identified four primary functions of non-verbal human behaviour:

- Expressing emotion
- Conveying interpersonal attitudes, such as like/dislike, dominance/submission)
- Presenting one's personality to others
- Accompanying speech for the purpose of managing turn taking, feedback, attention, etc.

Pioneers in the area of behaviour analysis Ekman and Friesen (1969a) classified non-verbal behaviours into the following five categories that are still considered as a standard, as they include virtually all types of social cues:

- **Displays of Affect** are behaviours directly connected to a specific emotion or affective state. An example is crying and frowning while being sad. While people are often aware that they show such a behaviour, they normally do not intend to communicate with such.

- **Emblems** are "codes" defined by a society concerning nonverbal behaviours and are usually learned similar to verbal sentences. Examples are head nods as a symbol of agreement or waving as a symbol of greeting. They are regularly shown on all nonverbal channels, although, according to Ekman and Friesen, for example in western cultures emblems are "primarily shown in the face and hands" (Ekman and Friesen, 1981).

- **Illustrators** are slightly less controlled than emblems and go along with verbal content. An example is widely opening the arms while talking about something large, just to illustrate or emphasis the spoken content. More concrete examples will be given in Section 2.2.1.1.

- **Regulators** act to mediate between conversational partners. Regulators are for example back-channels (such as head-nods) or "shifts in posture to bring about greater or lesser attention, or more or less distance" (Ekman and Friesen, 1981). The presence of regulators during a conversation is often a good indicator of engagement and conversational involvement. Due to this specific role, regulators will be discussed more precisely in Section 2.3. As regulators are social cues that only appear in interactions between two or more participants, we call their appearance "interpersonal cues".

- **Adaptors** are defined by Ekman and Friesen (1981) as "movements (that) were first learned as part of adaptive efforts to satisfy self or bodily needs, or to perform bodily actions, or to manage emotions, or to develop or maintain prototypic interpersonal contacts, or to learn instrumental activities."

  Adults apply such learned patterns, especially in social interactions. They do not show the original full behaviours, but rather only fragments of them that are kept out of habit. An example is a child hiding behind an object when it is scared, while an adult would probably use a hand in front of the face to hide in unpleasant situations.

"When the adaptor appears in the adult it is because something in the current environment triggers this habit; something has occurred currently which is relevant to the drive, emotion, relationship or setting originally associated with the learning of the adaptive pattern." (Ekman and Friesen, 1981).

Ekman and Friesen (1969b) further distinguished three subclasses of adaptors, namely self-adaptors, object-adaptors, and alter-adaptors.

Self-adaptors are behaviours that represent learned actions on the own body, e.g., using the hand to wipe over the mouth, covering the eyes, scratching etc.

Alter-adaptors are learned in conjunction with early experiences in interpersonal relationships, like giving and taking, attacking or protecting.

Object-adaptors are movements linked to the manipulation of objects, e.g., tapping a pencil or smoking.

Knapp, Hall, and Horgan (2013) consider adaptors to be always associated with negative feelings. Studies have indicated that "self-touching" (which is an instance of adaptors), is associated with "situational anxiety or stress" (Knapp, Hall, and Horgan, 2013).

One could say there is also an "inofficial" sixth category, which is **meaningless behaviours**, as Ekman and Friesen (1981) stated: "we must admit that there may be actions which are meaningless - random activity or noise, movements which have no regularities in either their encoding or decoding, not even for a single person".

As seen in the examples, social cues fulfil various tasks in social interactions. They happen on purpose or on an unconscious level, sometimes are real and sometimes acted. They can help to reduce ambiguity in social interactions (Sheth et al., 2011) or expose when somebody is lying. Further their granularity may vary from long-lasting obvious cues, such as crossing the arms, to very subtle and short ones, such as batting an eye lash.

In the next subsections we will give a more detailed description of social cues related to the single modalities that are relevant to our non-verbal communication and later chapters of this thesis.

2.2.1 Kinesics

> *"The body never lies."*
>
> — Martha Graham

The term *kinesics* was first introduced by Birdwhistell (1952). In popular literature, "body language" is a term often used as a synonym and includes gestures, postures and general body movements. This section gives an overview on kinesics related to body movements, while facial expressions, gaze behaviour (which are sometimes also included in the definition of "body language") and others are discussed in subsequent sections.

"kinesics" origins from the Greek word "kinesis", which translates as "movement" or "motion"

To differentiate between gestures and postures we refer to the definitions by Poyatos, 1981:

Gestures are "conscious or unconscious body movements made mainly with the head, the face alone, or the limbs, learned or somatogenic*.

Postures are "conscious or unconscious general positions of the body, more static than gestures, learned or somatogenic*, and serving as a primary communicative tool, dependent or independent from verbal language; either simultaneous or alternating with it, and modified by the conditioning background (smiles, eye movements, a gesture of beckoning, a tic, etc.) either simultaneous or alternating with verbal language, modified by social norms and by the rest of the conditioning background, and used less as a communicative tool, although it may reveal affective states and social status (sitting, standing, joining both hands behind one's back while walking, etc.)."

**somatogenic: Originating in the soma or body under the influence of external forces. / Having origin in body cells.*

Additionally to general gestures and postures, the *style* of our body movements gives further insights. The impression we leave to others is not only influenced by the gestures we make, but also on *how* they are performed. Every movement costs energy, so depending on the necessity –or what we regard as necessity–, gestures may be performed for example more or less expressive. Such observations are a good indicator of the motivation (or conversational engagement) of a person during an interaction.

2.2.1.1 Gestures

Kendon, Sebeok, and Umiker-Sebeok (1981) describe gestures to be "usually deemed to be an action by which a thought, feeling, or intention is given conventional and voluntary expression". For classifying different types of communicative gestures, psychologists created various schemes. An overview on the probably most important ones is given in Table 1. To stay in the previously introduced terminology of Ekman and Friesen all the classes in these schemes could be seen as *Illustrators* as their main purpose is to support spoken content. There are of course also gestures that match the other categories. An obvious example would be sign language (emblems), but also crossed arm gestures, which have the original function of physical self protection or self touches (adaptors).

Table 1.: Gesture classification schemes, source: (McNeill, 1992)

McNeill, 1992	**Efron, 1941**	**Freedman and Hoffman, 1967**	**Ekman and Friesen, 1969a**
Iconics	Physiographics Kinetographics	Literal-reproductive	Kinetographs Pictographs Spatial Movements
Metaphorics	Ideographics	Concretization Minor and major qualifying	Ideographs Underliners
Deictics	Deictics	-	Deictics
Beats	Batons	-	Batons Rhythmics

In general, all classification schemes aim to distinguish similar types of gestures, partly using different terms or classifying the terms on different granularity levels (basically because they all descent from Efron's scheme from 1941). As McNeill's scheme is the most current (and best documented) one the different "categories" will be illustrated on his terminology, but references to other classification schemes will be drawn where necessary.

- **Iconics** illustrate images of concrete entities or actions. The gesture itself acts as a referential symbol to a specific event or object. Ekman and Friesen (1969a) subdivide iconics in three sub-categories, namely *kinetographs*, *pictographs* and *spatial movements*. Kinetographs depict a bodily action, for example sleeping, cutting or shooting. Pictographs illustrate their reference. Spatial Movements depict spatial relationships. An example is stretching the arms to emphasise that something is very huge, or visualising a location when giving directions to somebody.

- **Metaphorics** represent the depiction of abstract content "as if it had form and/or occupied space." (McNeill, 2006). This also includes illustrating objects without presenting the object itself, but rather an idea or memory that it represents. In a metaphoric gesture McNeill calls the gesture itself *sign*, the (imaginary) object *base*, and the "concept" the *referent*. There are cases where the line between metaphorics and iconics cannot be drawn clearly.

- **Deictics** also called "pointing" gestures are typically performed with an extended index finger, but also other extensible parts of the body, or held objects can be used. They are not limited to indicate a concrete object, person or location, but also to point to abstract or imaginary objects (Krauss, Chen, and Gotfexnum, 2000). Deictic gestures also have the function to "shift focus" and that "focus can be used to disambiguate gestures" (Wahlster, 1998).

- **Beats** have their name because in this type of gestures the hand appears to be beating time. In Effron's and Ekman's schemes they are called "batons" as an analogy from music. "This rhythmicity has made beats seem purely speech-related. However, they also have discourse functionality, signalling the temporal locus of something the speaker feels to be important with respect to the larger context. One can think of a beat as gestural yellow highlighter."(McNeill, 2006).

McNeill claims in his more recent work that these *"categories"* should rather be considered as *"dimensions"*. This is because many gestures contain rather attributes of these classes, but can not be "unambiguously assigned to a specific category" (McNeill, 2006). He suggests to speak of "iconicity, metaphoricity, deixis, 'temporal highlighting' (for beats), social interactivity, or some other equally unmellifluous (but accurate) terms conveying dimensionality" (McNeill, 2006).

Kendon (1980) decomposes a gesture in three hierarchical elements. A *gesture unit* is defined as the "period of time between successive rests of the limbs." It starts the moment the limb begins to move and ends when it has reached the rest position again. One or several *gesture phrases* appear within one gesture unit and is what we probably would intuitively call a *gesture*. A *gesture phrase* consists of a set of the following *gesture phases*:

- **Preparation (optional):** The hand leaves its rest position and gets into the "gesture space" where it prepares the stroke.
- **Stroke**: The main phase that carries the meaning and that is intuitively understood as the gesture. The stroke can be seen as an "object" being presented. It is "prepared for, withheld if need be until the co-expressive speech is ready, and held again until all linked speech is over" (McNeill, 2006).
- **Retraction (optional):** The hand returns to the rest position, which can differ from the start position. This phase is optional if the speaker moves on directly to the next stroke phase.
- **Pre/post-stroke hold phases (optional):**

  A temporary cessation of motions shortly before or right after the stroke motion. Their purpose is "to ensure that the stroke remains semantically active during the co-expressive speech"(Alibali, Kita, and Young, 2000).

2.2.1.2 Postures

Directly related to gestures, postures are good indicators of a person's emotion or attitude (Dael, Mortillaro, and Scherer, 2012a; Richmond, McCroskey, and Payne, 1991; Montepare et al., 1999) as they are foremost performed unconsciously.

Mehrabian (1969) describes a two-dimensional scheme to characterise postures with regard to their role concerning sympathy and status. The degree of *"immediacy"* subsumes touching, closer positioning, forward leaning, eye contact, and direct body orientation. The second dimension *"relaxation"* includes cues that are asymmetrical rather than symmetrical.

Main categories of this scheme include:

- **Body inclination:** During a social interaction a person may turn to lean toward a conversational partner or away from their direction. Such behaviours mostly happen on an unconscious level. Leaning away may signal negative attitudes, such as the desire to end the conversation or general disagreement or disinterest. Head inclinations have similar analogical meanings. A study by James (James, 1932) suggests that a forward learning posture conveys more positive feelings than a reclining position. Reece and Whitman (1962) concluded that an experimenter's behaviour that included forward leaning, smiles and resting hands was perceived with a warm attitude.

- **Body orientation:** The degree to which a communicator's shoulders and legs are turned in the direction of his interlocutor are another category related to body postures. People usually talk directed towards each other, but usually not in a direct confrontational stance, but rather turned towards each other at an angle. When a person is disengaged in a conversation she or he tends to avoid eye contact and turns the head or the body away. Another interesting aspect is the direction of the feet. For example, turning the body and both feet towards the interlocutor while greeting is considered sincerely. If only the torso moves towards the person, while the feet still show towards another direction, this turn is done most probably because of social manners, such as politeness. It shows that the person is not really interested in a conversation (Molcho, 2001). In the previously mentioned study, James (James, 1932) concluded that a more direct orientation is associated with a more positive attitude.

- **Openness of the arrangement of arms and legs:** The openness of the body posture is another aspect that might give insights into the emotions and attitudes of a person. While children often hide behind objects like furniture, this behaviour changes during adolescence towards using adaptors, such as the arms or hands to obstruct the chest or face. At the same time it serves to protect vital organs, such as the throat or abdomen, but also creates a barrier for delimitation (see Figure 3). A closed posture often "gives the impression of detachment, disinterest and hostility and usually conveys unpleasant feelings" (Rossberg-Gempton and Poole, 1993).

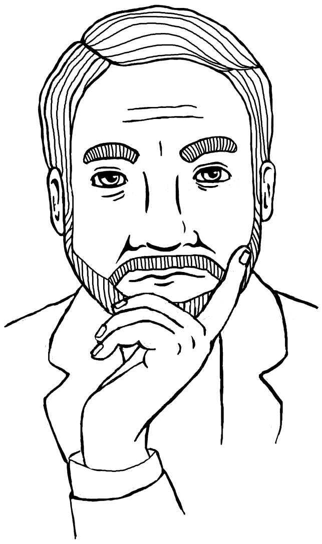


Figure 3.: One arm in front of the torso, the neck and the face: a variation of a closed body posture

The "crossed arms" posture is often applied in variations, e.g., by not crossing both arms but only one as a barrier, or holding ones own hand, such as "parents would do with their children in dangerous situations"(Morris, 1997). Calero (1979) further investigated postures related to crossed hands and concluded that this is mostly a sign of frustration where the person is trying to avoid displaying a negative stance. He further inferred that the elevation of the crossed hands resembles the degree of negativity or even antagonism, where the head position is the most negative degree. Sometimes, people cross their arms or hands when they are cold. This is not a case of contradictions to the behavioural psychology argumentation (Navarro and Karlins, 2008). The feeling of coldness may cause indisposition and discomfort.

In contrast, an open posture is often associated with a positive and friendly attitude. In open postures normally "the head is straight, the feet are spread and the palms of the hands are up and relaxed" (Rossberg-Gempton and Poole, 1993).

- **Relaxation:** According to Mehrabian's scheme, in most cases someone with higher (social) status applies a more relaxed pose than someone with a lower position. For example, he or she normally sits down and talks while the other person, with lower

> status, rather stands until he or she is asked to sit down. People with a lower social status often maintain symmetrical postures by, for example, "placing both hands on their lap or at their sides" (Mehrabian, 1969).

Various efforts have been made to classify and precisely describe body movements and constellations of joints. Bull (1987) introduced the "posture scoring system", which contains exact definitions of postures for the four main categories: head, trunk, arms and legs. A similar approach was taken by Dael, Mortillaro, and Scherer (2012b) with the "body action and posture coding system". It allows the description of postures and other body movements in three abstraction layers: anatomical level (involved body parts), form level (direction and orientation of movements) and a functional level (communicative and self-regulatory functions).

2.2.1.3 Style and Expressiveness

"A blur of blinks, taps, jiggles, pivots and shifts ... the body language of a man wishing urgently to be elsewhere."

— Edward R. Murrow

A third aspect of kinesics that should be considered is the expressiveness of the body movements (for both gestures and postures). As mentioned earlier, body movements may have different meanings in different contexts. There are gestures that are unambiguous, while the possibilities of interpretation vary a lot for others. Especially the performance of a movement may draw conclusions about a person's inner emotional state or social attitude. This follows a simple physical rule: every effort required to perform a gesture costs energy. So basically in situations where we see the necessity to invest such efforts we tend to make more energetic – and also quantitatively more movements. Such situations could be moments that seem important to us, for example job interviews or public speeches, or asking someone out for a date. People often tend to overreact in such situations showing signs of nervousness such as restless hand movements or tapping with a foot. But even on a "normal" level the impression people get from one's body movements often relates to the non-verbal efforts made. This could be simply showing an upright posture (which costs energy), but also the amount of illustrators (see Section 2.2.1.1) used during the conversation.

Wallbott (1998) made efforts in mapping qualitative parameters of movements to emotional states and cultural differences. Hartmann, Mancini, and Pelachaud (2005) were using expressiveness parameters as an approach to map the quality of communicative functions, such as mood, personality and emotion. Based on Wallbots previous studies, Caridakis, Raouzaiou, et al. (2006) concluded to regard the following six expressiveness attributes to make further statements on body movements.

Cardiakis et al. used these parameters in a computer vision approach

- Energy/Power (EN) represents the dynamic properties of a movement (e.g. weak versus strong). It is calculated from the motion vectors' first derivative in all three dimensions.
- Fluidity (FL) differentiates smooth from convulsive movements. This feature aims to capture the continuity between movements. It is calculated as sum of the variance of the motion vectors norms $(\vec{l}, \vec{r})$ of both hands, respectively feet for leg postures.
- Spatial extent (SE) is modelled as the space occupied for gesticulation in front of a person. It calculates as maximum Euclidean distance hands' position (l,r) (respectively feet for leg postures).
- Overall activation (OA) represents the quantity of the movement (passive versus active). It is calculated as the sum of the motion vectors' norm of both hands (respectively feet for leg postures).
- Temporal extent (TP) represents the duration of a gesture (short vs sustained).
- Repetition (RE) includes information about the repetition of the stroke (e.g., for beats).

Concrete formulas for the different expressiveness dimensions can be found in Appendix A.

The way gestures are executed, depends on individual and social factors such as personality, emotional state or culture. Gallaher (1992) speaks more generally of the behavioural "style" of a person. She identified the four dimensions expressiveness (which Hartman et al. used as the overall term), animation, expansiveness and coordination. These dimensions prove to be consistent for an individual, stable over time, as well as stable across raters. In general they also match the previously introduced scheme, but relate more on a person's general way to communicate.

In the next section we will turn to facial expressions, which are considered one of the most relevant modalities when analysing emotions.

2.2.2 Facial Expressions

> *"Smiles are probably the most underrated facial expressions, much more complicated than most people realise. There are dozens of smiles, each differing in appearance and in the message expressed."*
>
> — Paul Ekman

The first recorded scientific studies into facial expressions – more precisely smiles – were performed in the middle of the nineteenth century by Guillaume Duchenne de Boulogne (1849) who used electrodiagnostics to distinguish between a smile of real enjoyment and other kinds of smiling. He found out that smiles are controlled by two sets of muscles, the *zygomatic major* which connect the mouth and the side of the face and the *orbicularis oculi* which pulls the eyes back. The *zygomatic majors* are consciously controlled, which allows to "fake" signs of enjoyment to appear more friendly. In contrast, the *orbicularis oculi* are independent and normally not controlled consciously. Nowadays smiles involving both groups of muscles are named after their discoverer *Duchenne*. Messinger, Fogel, and Dickson (2001) suggest that *Duchenne smiles* are uniquely associated with positive emotions. Besides the contradiction of the *orbicularis oculi*, more unconscious facial expressions exist that reveal true feelings and attitudes, whether or not we want them to be observed by others. Ekman (2009) calls these involuntary expressions micro-expressions, as they usually last for only less than half a second.

In 1978, Ekman and Friesen introduced the "Facial Action Coding System" (FACS), a system to categorise human facial expressions (Ekman, Friesen, and Hager, 1978). It is based on a set of 64 Action Units (AU) which allow human annotators to describe nearly any possible facial configuration. Coming back to the example given at the beginning of this section, FACS allows to distinguish a *Duchenne* from a *non-Duchenne* smile by looking at the raise of the lip corners (*zygomaticus major*, AU 12) and the contraction of the muscles responsible for raising the cheeks (orbicularis oculi, AU 6) (Ekman, Davidson, and Friesen, 1990).

Besides expressing emotions, moods and attitudes, people use facial expressions for providing non-verbal feedback during social interactions and to reflect their interpersonal attitude (Knapp, Hall, and Horgan, 2013). For example, we use the face to signal the opening and closing of a conversation, complement or qualify verbal and/or nonverbal responses, and even replace speech (Knapp, Hall, and Horgan, 2013). During a conversation, smiles are often used as a signal of attentiveness and involvement (Brunner, 1979). Facial expressions are also used as what Ekman and Friesen denote as (facial) emblems (Ekman, 2003). A facial emblem is the expression of an emotion, without actually being in this particular emotion. For example a person may twist the mouth or nose and lower the eyes to share sympathy for a unlucky situation someone is telling about.

While sometimes included in the category of facial expressions, we want to have a look at gaze behaviour (or oculesics) as a modality of its own in the next section.

2.2.3 Oculesics

> *"Eye rolling is one of the nonverbal signs that is pretty much always aggressive."*
>
> — Steve Watts

Besides their primary use (giving us the ability to see), our eyes are also a strong medium for communication, as it is easy for others to follow our gazing (Kobayashi and Kohshima, 2008) and our eyes may give further insights on our emotional state.

Reis and Sprecher (2009) distinguish four dimensions of oculesics:

- **Eye movements and blinking** are signs of cognitive activity to which receivers attribute considerable relational meaning. Eye movements occur both voluntarily or involuntarily. They include "changing eye direction, changing focus, or following objects with the eyes." (Van der Stigchel, Meeter, and Theeuwes, 2006).
- **Pupil dilation** is an important subliminal cue that sends messages of warmth, affection, and interest to receivers. The change in the size of the pupil is called *pupillary response*. This change happens when the focus switches to new objects, but also when the appearance of objects is indicated (Moresi, 2009).
- **Gaze directions** also send relational messages of attentiveness or disinterest and degrees of conversational involvement. Directing someone's gaze on purpose can be used to shift focus towards another person or an object (Senju and Csibra, 2008).
- **Eye contact** is probably the most important oculesic cue. Another term we will use here is *mutual gaze*. Eye contact is an "invitation to communicate in initial interaction and a crucial communication channel in close relationships. Eye contact can also regulate communication, increase immediacy, monitor ongoing interaction, intimidate, promote flirtation, provide turn-taking cues, signal attentiveness, increase warmth, and perhaps most importantly, express involvement and intimacy" (Reis and Sprecher, 2009).

The duration we look another person in the eye is also of importance. Keeping permanent eye contact can be interpreted as unpleasant or even aggressive by others.

2.2.4 Paralinguistics

> *"The most important thing in communication is hearing what isn't said."*
>
> — Peter F. Drucker

Speech is one of humankind's greatest achievements. Vocal communication frees a speaker's hands, can occur in darkness, and does not require looking at the individuals who are signalling (Lieberman et al., 2007). In its complex form, speech is a unique human trait and guaranteed survival throughout history. Speech offers humans a medium to communicate their needs, thoughts, intentions, memories, and knowledge. Paralinguistic, or also called vocalic communication is non-verbal information produced by our voice, the tongue and the lips. It is that part of speech that does not contain the actual verbal content encoded by speech, but the way we use our voice to support the meaning of the message, or to communicate our emotions and attitude about it.

Some vocalic cues, such as *sighs, groans, cries, laughter, inhalations, fillers* and *yawns*, stand alone, but vocalic information accompanies all spoken communication. Depending on *how* it is said, a word such as "okay" can communicate agreement, confusion, boredom and many other messages.

Vocalic cues such as *pitch, cadence, speed, resonance, control, duration, loudness, silence*, and *accent* lend meaning to every spoken word. (Reis and Sprecher, 2009; Knapp, Hall, and Horgan, 2013). In daily situations, recognising emotions from voice is not always as easy as it might seem. Social constraints might keep a speaker from communicating their true intentions and context is again crucial when analysing such cues.

2.3 Interpersonal Cues

> *"It is only once in a while that you see someone whose electricity and presence matches yours at that moment."*
>
> — Charles Bukowski

From a behaviour analysis point-of-view, the last section was mostly concerned with social cues that are to be observed when having a look at a single person in a conversation. In this section we want to have a closer look at what Ekman and Friesen called *regulators*. Considering regulators from a meta-perspective, one could even say that their appearance is depending on the behaviour of both interlocutors in dyadic conversation, respectively multiple interlocutors in group conversations.

Ekman and Friesen (1969a) coin the term *regulators* to draw attention "to actions which maintain and regulate the back-and-forth nature of speaking and listening between two or more interactants. They tell the speaker to continue, repeat, elaborate, hurry up, become more interesting, less salacious, give the other a chance to talk etc. They tell the listener to pay special attention, to wait just a minute more, to talk etc." Although, as already discussed, conversational partners' goals, social norms, and even environmental factors influence the course of interactions, there is typically pressure for stability and predictability in our contacts with others. The most common pattern of exchange in interactions among acquaintances, friends, and loved ones may be described as *reciprocation*. That is, the *involvement* initiated by one person is matched or reciprocated by the partner. This kind of adjustment may be reactive when the second person responds to the first person's initial behaviour with a similar change in involvement. For example, the first person might lean forward and smile (increased involvement) and the partner responds with an increased amount of mutual gaze and a friendly touch (Patterson, 1982).

Next we'll have a look at the most common interpersonal behaviour patterns, especially related to conversational engagement.

2.3.1 Synchronicity

A main aspect of interpersonal behaviour is *synchronicity*. Related concepts found in literature are *mirroring* or *mimicry*. The latter is adapted from flora and fauna where certain species mimic others to gain advantages of the environmental surroundings (think for example of hover flies, which are harmless pollinators, but use the same colour patterns as wasps to protect themselves). In human communication, the term describes the unconscious tendency to imitate the behaviour of our opposite. Such imitation of gestures, postures, facial expressions and others, plays an essential role in "empathy, affiliation, and rapport" (Chartrand and Van Baaren, 2009) which are considered related concepts to engagement. In contrast, a lack of synchronous behaviour may appear forced or unpleasant.

Synchronisation between interlocutors can often be observed, but most of the time people have no conscious perception of it. Cowie, Pelachaud, and Petta (2010) describe that this is not necessary because "one can perceive another person's engagement or disengagement without explicitly knowing which aspects of his or her behaviour triggered such awareness."

Chartrand and Bargh (1999) studied the existence of such unconscious behaviour patterns and named them the *"chameleon effect"*. They also found that the perception of another person's behaviour increases the likelihood of "engaging in that behaviour oneself". A common theory about mirroring is that "individuals understand others by replacing their states, using their own somatosensory resources" (Van Baaren et al., 2009).

This theory came up by observations in experiments with rhesus macaques (Rizzolatti and Craighero, 2004; Binkofski and Buccino, 2006). In this study, neurons got activated in a specific area of the brain (the *"ventral premotor cortex"*) when the monkey grasped an object. The same neurons showed activity when the monkey watched the experimenter take the object. It is argued that the mirror neuron system (MNS) was "evolutionarily selected to subserve action-understanding and promote social learning" (Iacoboni, 2009).

Of course we do not imitate other people at all times. Unconsciously mimicking others is "moderated by both enduring and temporary characteristics of the mimicker and the mimickee" (Van Baaren et al., 2009).

2.3.2 Listener Responses

Next we'll have a look at some of the most relevant behaviours that appear especially for listeners in conversations. Such cues are, for example, related to giving feedback, or efforts to take the turn. Zimmermann (1996) discusses the importance of the listener in a conversation and even suggests that the "quality of a conversation depends largely on what takes place in the person to whom words are directed." Allwood (1993) distincts the following four basic communicative functions on which a listener may give feedback:

- *Contact* describes whether the person is willing "to continue the interaction"
- *Perception* describes whether the person is willing "to perceive the message"
- *Understanding* describes whether the person is willing "to understand the message"
- *Attitudinal reactions* describes whether the person is willing "to react and (adequately) respond to the message, specifically if he/she accepts or rejects it."

2.3.2.1 Backchannel Expressions

Backchannel expressions are "short, typically mono- or disyllabic" (Gardner, 2001) listener responses that appear either verbally or nonverbally. They are used primarily as response in one-way communications, or in a conversation while the other person has the turn. The term "backchannel" implies that there are "two channels of communication operating simultaneously during a conversation" (White, 1989) and "appears very important in providing the monitoring of the quality of communication" (Yngve, 1970). Yngve (1970) defines a backchannel as follows:

"Both the person who has the turn and his partner are simultaneously engaged in both speaking and listening. This is because of the existence of what I call the back-channel, over which the person who has the turn receives short messages such as 'yes' and 'uh-huh' without relinquishing the turn. The partner, of course, is not only listening, but speaking occasionally as he sends the short messages in the back-channel".

While the primary channel is used by the main speaker to communicate, the secondary channel (the backchannel) is used to provide feedback and express comprehension and interest. Their main purpose is to respond in a supporting (i.e. non-disagreeing or non-challenging) manner to the other participants immediately preceding or current vocalization (Iwasaki, 1997). Backchannels indicate that the listener is "following and understanding the speaker" (Heldner, Hjalmarsson, and Edlund, 2013).

Iwasaki (1997) further distinguishes between three types of backchannels:

- **Non-lexical backchannels** are vocalic sounds which have little or no referential meaning and form a closed set. For example, in English, German and other languages, sounds like "uh-huh" and "hmm" serve this role.
- **Phrasal backchannels** are expressions with more substantive meaning than non-lexical backchannels. Examples are sentences like "Are you kidding?" or "Seriously?".
- **Substantive backchannels** go beyond phrasal backchannels as they refer to actual content. Their main purpose is for example asking for clarifications, or giving short comments.

In practice, some vocal backchannels may be replaced or accompanied by visual backchannels such as head nods or facial expressions (like raising the eyebrows, smile) (Boholm and Allwood, 2010; Wlodarczak et al., 2012; Levow and Duncan, 2012; Brunner, 1979), or the interlocutor passing "the opportunity to produce a backchannel because it would occur too close in time to the previous one." (Heldner, Hjalmarsson, and Edlund, 2013).

2.3.2.2 Turn-Taking

In contrast to listener-responses, general turn-talking cues signal the intent to take the lead in a conversation. It involves both, verbal and non-verbal cues. Duncan (1972) formulated a system of turn-taking signals and rules. The interlocutor, wishing to get the speaking turn, displays a 'turn-yielding' signal, e.g. the termination of any hand gesticulation and/or a drop in paralinguistic pitch. Kendon (1967) suggests that the conversational interchanges between a speaker and a listener are regulated partly by gaze. He found that a speaker typically looks away at the opening of a long statement (which effectively signals the listener that he is about to speak) and looks at the listener as the end of the utterance is reached to provide an indication that the "offer of the turn" has been accepted. The auditor, if he wishes to accept the floor, displays a "speaker-state" signal, e.g., a shift away in head direction and a sharp intake of breath. At the same time, the original speaker switches to the auditor mode (Thomas and Bull, 1981).

In general, according to Duncan (1972), shifts of the head and body posture are involved in regulating turns. They are marking semantic, as well as syntactic boundaries of concurrent speech and managing speech disfluencies (Scheflen, 1964; Kendon, 1972; Thomas and Bull, 1981). For example, *postural shifts* often occur towards the beginning and the end of turns in regulating speaking and listening but also "accompany topic changes in speech content when marking semantic boundaries." (Scheflen, 1964). Syntactic functions imply the occurrence between sentences or groups of clauses (Kendon, 1972). Hadar et al. (1984) found that postural shifts happen mostly towards the beginning of speech after long pauses or directly after listening.

While the turn-taking structure is generally universal for humans, conventions vary by culture and community, e.g. how turns are distributed, how transitions are signalled and how much overlapping in speech is acceptable (Sidnell, 2007).

A conversation is regarded fluent and dynamic (and in most cases interesting/engaging) if turn-talking conventions aren't violated. We will have a look at *adjacency pairs* as an example of successful turn-taking, and *interruptions* as an example of failed turn-taking.

- **Adjacency pairs** provide an example of positive conversational turn-taking. In linguistics "an adjacency pair consists of two utterances by two speakers, with minimal overlap or gap between them, so that the first utterance provokes the second" (Rich, Ponsleur, et al., 2010). For example, the speaker communicates the first turn: "what is your name?" This is followed by a delay (according to the study by Rich, Ponsleur, et al. (2010) between 0 and 1.1 seconds, 0.4 seconds on average). The *first turn* requires the interlocutor to answer in the next turn, for fulfilling the adjacency pair. Furthermore, a satisfying *second turn* by the interlocutor has to be given, for example in my case: "my name is Tobias". An irrelevant or unfitting response, such as "I like turtles!" would not satisfy the adjacency pair. In their work, Rich, Ponsleur, et al. (2010) generalise the concept of linguistic adjacency pairs to also include non-verbal communication acts, such as "head nods" as replacement for a spoken "yes".

- **Interruptions** are a violation of the turn-taking flow. Dyadic conversations are basically structured in a way that no interruption occurs (Sacks, Schlegloff, and Jefferson, 1974). As discussed before, the synchronisation between the speaker and listener follows a protocol where the speaker sends signals when they want to offer the turn, respectively the listener that they want to take the turn. In an ideal scenario, the listener waits for the turn to be finished. An infringement of that protocol is an *interruption* as the conversation structure is disturbed. In face-to-face conversations, interruptions can be considered as turn-taking violations (Beattie, 1981) (e.g. claiming the turn by interrupting the current speaker), but they can also serve as important social displays which reflect interpersonal attitudes (e.g., dominance or cooperation) as well as involvement in the interaction (Murata, 1994).

Roger, Bull, and Smith (1988) defined a coding system to distinguish between different classes of interruptions. It is set up in a hierarchic structure. The root node decides whether a first and second speaker exist. If not, they label such an event a *false start*, respectively unintended simultaneous speech. Typically, such events are "followed by some form of repair, for example, a pause followed by an apology and a sequence of mutual floor-offering" (Roger, Bull, and Smith, 1988).

In case a second speaker is identified, they further distinguish if the second speaker's turn was disrupting the first speaker's utterance or not. In case it is not disrupting it leads to "non-interruptive simultaneous speech", which determines whether the first speaker continues the utterance. In case he or she does continue, the event is labelled as an *overlap* (also see (Drummond, 1989)), otherwise as a *listener response* (such as a back-channel) or as an *afterthought*. An afterthought could be seen as an utterance that follows the main utterance with a sentence such as ". . . or something" or ". . . well, anyway"'. The second speaker's utterance isn't regarded as interruptive if she or he begins to "speak before or during the afterthought, because the first speaker is regarded as having effectively completed his or her turn before the afterthought occurs" (Roger, Bull, and Smith, 1988). Now if the second speaker is disrupting the first, they further distinguish interruptions depending on the number of attempts to interrupt, more precisely if it's one attempt (*single interruption*) or if it is the case that there are multiple attempts (*complex interruption*). They further discriminate if the second speaker succeeds in stopping the first from finishing her or his utterance.

The effect of interruption in dyadic conversations between humans (but also between humans and virtual agents), has been extensively investigated (Tannen, 1994; Oviatt et al., 2015; Cafaro, Glas, and Pelachaud, 2016; Heldner and Edlund, 2010). Most of those works examine aspects that influence the perception of the interrupter (e.g. status, and sex (Beattie, 1981; Robinson and Reis, 1989)). Also, the type and strategy of an interruption is frequently researched (Cafaro, Glas, and Pelachaud, 2016; Murata, 1994; ter Maat, Truong, and Heylen, 2010).

In this and the previous sections important aspects of human behaviour analysis have been introduced. Next, we will have a closer look at phenomena that cause such behaviours – or regarded from the opposite perspective – that may be inferred from observed behaviours. Therefore we'll investigate complex phenomena, such as emotions and social attitudes.

2.4 Affect, Emotion, Feelings and Mood

"I continue to be fascinated by the fact that feelings are not just the shady side of reason but that they help us to reach decisions as well."

— Antonio Damasio

2.4.1 What is an Emotion?

Zimmer (1988) associates such phenomena with concepts like feelings, emotions, affect, sensation, drives, passion, instinct, mood, temperament and motivation. He argues that some of these concepts have the same meaning, others coincide, but overall philosophers and psychologists haven't been able to find clear borders that separate these terms precisely. More generally speaking, we assume that all of these concepts are at least connected, and we preeminently use the terms *emotion* or *affect* to describe them, keeping in mind that differences to other (sub) concepts exist. The existence of so many descriptions is due to the fact that philosophers have been concerned with trying to understand and describe emotions since the antiquity. For example, Plato (428-348 B.C) proposed a three-part division of the mind (Adam et al., 1902). Reason, appetite (which either controls or is controlled by reason) and in between spirit, exemplified by anger. Aristoteles (384–322 B.C.) understood emotions (as he called passions) as psychic experience corresponded to appetites or capacities (Bostock, 2000).

Modern theorists like Damasio or Hume differentiate between *feelings*, *emotions* and *moods*.

"*Emotions* are complex, largely automated programs of actions concocted by evolution. The actions are complemented by a cognitive program that includes certain ideas and modes of cognition, but the world of emotions is largely one of actions carried out in our bodies, from facial expressions and postures to change in viscera and internal milieu." (Damasio, 2010). Emotions play a critical role in cognitive processes such as perception, learning, and decision-making and are equally critical in the maintenance of health. (Damasio, 2001)

"*Feelings* of emotion, on the other hand, are composite perceptions of what happens in our body and mind when we are emoting. As far as the body is concerned, feelings are images of actions rather than actions themselves; the world of feelings is one of perceptions executed in brain maps. But there is a qualification to be made here: the perceptions we call feelings of emotion contain a special ingredient that corresponds to the primordial feelings discussed earlier. Those feelings are based on the unique relations between body and brain that privileges interoception* ." (Damasio, 2010).

**interoception: sensitivity to stimuli originating inside of the body*

Moods are "feelings that tend to be less intense than emotions and that often lack a contextual stimulus" (Hume, 2012). It is argued that mood lasts longer in duration than of emotions (Morris and Reilly, 1987).

As implied before, there are many different approaches to understand and define what an emotion is. Next, we'll have a look at categories of emotion models, and illustrate these with examples.

2.4.2 Models of Emotion

In general there are multiple classifications of models and some are not disjoint from each other. Some models are focused on classifying emotions, others on describing the origin of emotions. Here we distinguish between models which view an emotion as a category (e.g, joy, sadness) and models that measure an emotion on a dimensional scale (e.g., in a valence arousal space).

2.4.2.1 Categorical Models

- **Basic models:** The most known categorical emotion models are so called "basic models". They share the idea that an emotion is a discrete entity that appears in all people from all cultures in a similar way. That means, when observing an expression of emotion, one can directly infer a person's emotional state. Emotions are expressed by certain behaviour patterns that are exclusive to each emotion. In the simplest examples a person shows a ("duchenne") smile when she or he is happy, or frowns and tears when she or he is sad. (As mentioned before, this perspective often can lead to wrong assumptions, leaving out contextual information). In general, especially in basic models, facial expressions are seen to be the most reliable cues to infer the actual emotion. One reason for that is that the same brain regions are activated for processing facial expressions, but also for processing emotional information (Wronka and Walentowska, 2011). Paul Ekman is probably the most prominent representative of basic emotion models who came up with the theory that emotions are recognisable between different cultures (Ekman and Friesen, 1971).

- **Psychological and social construction models:** Other forms of categorical models are psychological construction models. The main difference to basic emotion models is that in addition to typical non-verbal response (smiling for being happy), they also account for a wider variability in emotion expression (to stay in the example: crying while being extremely happy, or smiling of embarrassment). Barrett (2011) sees such expressions to be "better understood as symbols of emotion, rather than signals." In this case symbols do have emotional meaning but can not be seen as a direct mapping to a person's emotional state. It is argued that there are individual, but also cultural differences in expressing emotions. This challenges Ekman's previously described experiments, that involved isolated emotional expressions, reflecting a Western notion of emotion (Barrett, 2006; Feldman Barrett and Russell, 1998). In general, physiological construction models suggest that emotion results from basic psychological processes (like positive or negative feelings and physiological activation), combined with other factors such as previous experiences, language and cognitive control.

  Another perspective on categorical models is represented in social construction models. Their main idea is that they consider emotions to be only based on experiences and context, without a biological origin. Extreme representatives of this theory claim that emotions only exist in reciprocal exchange (see Section 2.3). According to this theory, emotions are not internal emotional or mental states, but rather performances that follow the common cultural and social rules. Depending on the local conditions and practices, the same emotion can be expressed in various ways by different cultures (Harré, 1986).

2.4.2.2 Dimensional Models

In dimensional models, emotions are represented in several dimensions. The classical and most known model is the valance-arousal Model (also called circumplex model) (Posner, Russell, and Peterson, 2005). Valence differentiates positive from negative states of emotion, arousal is related to mental and physical activity. Based on these dimensions a variety of emotions can be described (see Figure 4).

Mehrabian and Russell (1974) extend the original model with a third dimension, namely dominance (sometimes referred as potency). It describes an individuals sense that she/he has the power to deal with relevant events.

There have been multiple approaches to additionally add dimensions such as the Unpredictability (Fontaine, Scherer, Roesch, et al., 2007) or interpersonal engagement (Kitayama, Markus, and Kurokawa, 2000). *Engagement* (opposed to detachment) is sometimes also referred to by other terms. For example, Ortony, Clore, and Collins (1990) use

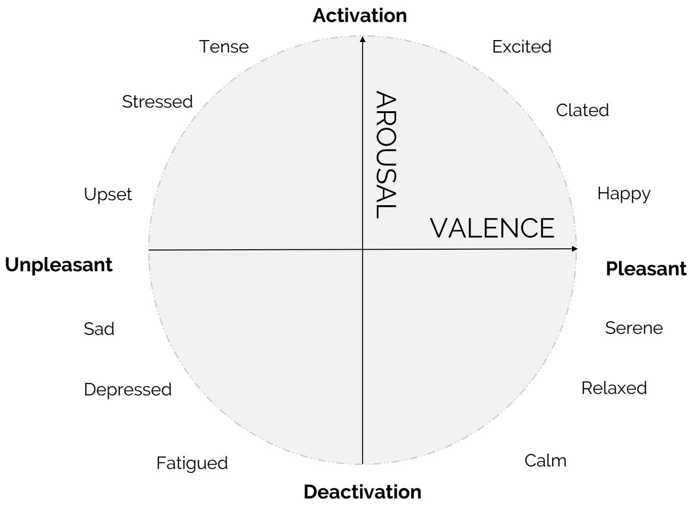


Figure 4.: The Valence-Arousal Circumplex Emotion Model

the term *"caring"* to express such a concept. Multi-dimensional models offer the advantage to express less intense, as well as blended emotions (Plutchik and Kellerman, 2013). Nevertheless, just as the basic models, these models mostly leave out context and interpersonal regulation.

2.4.2.3 Appraisal Models

Appraisal models are similar to basic emotion models in a sense that they also consider an emotion, once triggered, as biologically predetermined. Additionally they assume that emotions are triggered by evaluations of external events or situations (Scherer and Zentner, 2001). To give an example, we think of a person being in a job interview. If the interview works out well, the person might feel happiness or excitement because they appraise this event as good. If, in contrary, the interview doesn't go well and the candidate gets rejected she or he might feel sad and negative. In future job interviews the candidate might then evaluate questions as very negative and stressful, as last time "it all went wrong there". Different people might have "individual variances in emotional reactions to the same event" (Smith and Lazarus, 1990) based on previous experiences and appraisals.

The most famous representative of appraisal models is the OCC model (Ortony, Clore, and Collins, 1990). It differentiates six classes of emotion types namely:

- Prospect-based: appraisal of events related to future events for oneself (hope and fear (confirmed: satisfaction, fears-confirmed; disconfirmed: relief, disappointment))

- Fortunes-of-others: appraisal of events related to other persons (happy-for, resentment, gloating, pitty)

- Attraction: appraisal of things (love, hate).
- Well-being: events related to oneself (joy, distress)
- Attribution: appraisal of own/others acts (pride, shame, admiration, reproach)
- Well-being/attribution: combination of well-being/attribution emotions (gratification, remorse, gratitude, anger)

It is noticeable here that a main emphasis in the explanation of emotions is put on others and how we feel towards others and events that occur to them (as well as to oneself). This is closely intertwined to emotional regulation, based on specific events and past appraisal, our strategy of coping is influenced.

Another example is Scherer's (Scherer, 2005) component-process-model of emotions. In this model, an emotion is seen as process and consists of five components that need to be coordinated for an emotion to be experienced.

- Cognitive appraisal: evaluation of external events.
- Bodily symptoms: physiological part of emotional experience, e.g. raised heart-rate.
- Action tendencies: preparation and direction of actions.
- Expression: emotional expressions signal reactions and courses of action.
- Feelings: experience of the emotional state.

Scherer's model includes emotion regulation in form of action tendencies and coping processes. In the next section these particular concepts are further elaborated.

2.4.3 Emotion Regulation

As we have seen in the last section, there are many approaches for explaining and modelling emotions and their origin. One important aspect in newer approaches (especially in appraisal theory) is, besides physiological processes and previous appraisals, external events influence our emotional state and the expression of emotion. This being said, emotions are not always expressed the same way they are felt. *Emotion regulation* describes "the way people attempt to regulate their emotions, for instance by denying, intensifying, weakening, curtailing, masking, or completely hiding them" (Gross, 2002).

Two interesting sub-concepts of emotion regulation will be investigated in more detail in the next subsections: The first one is *emotional coping strategies*, the second related concept is the previously introduced *action tendencies*.

2.4.3.1 Emotional Coping Strategies

As described before, newer emotion models include cognition as a factor besides basic feelings. Cognitive processes in these theories mostly include judgements, appraisals or thoughts and are a prerequisite for an emotion to occur. According to Fontaine, Scherer, and Soriano (2013) the function of emotions is to prepare the organism for adaptive behaviour in specific situations. Evolutionary, the preparation of behaviours in recurrent situations increases an organisms chance of survival. In the simplest case one can observe the preparations of appetitive (approach) or defensive (withdrawal) reactions in organisms (Frijda, 2010). Besides appetitive and defensive motivations also others exist. Examples are "anxiety, motivating reticence in the face of possible punishment" (Gray and McNaughton, 2003), and cognitive motivation, such as "wonder, interest, curiosity, exploration and fascination, that are all elicited by pleasant, as well as unpleasant events" (Rimé, 2009). In other words one could say emotions function as heuristic mechanism of selecting behaviours.

In experiments in the 1960s Lazarus and Alfert (1964) found that cognition plays a crucial role on emotions. They showed probands silent movies of ritual genital mutilation in aborigine tribes and measured the stress factor of the participants with skin conductivity sensors. When a trivialised comment was added during the movie, the stress factor went down. It declined even more when the comments were given before the video which lowered the stress factor due to the changed expectations towards the movie.

Lazarus and Folkman (1984) distinguish between three degrees of appraisals towards situations

- **Primary appraisal:** According to Lazarus theory, situations can be evaluated as either positive, irrelevant, or stressful. A stressful situation may have one of the following three levels:

  a) challenges in situations that seem to be manageable b) threat when damage or harm is expected c) harm/loss when damage is already done.

- **Secondary appraisal:** In the phase of secondary appraisal it is evaluated how a situation may be managed with available resources. If the available resources are not sufficient, a stress reaction is caused (for example physiological signals such as increased heartbeat frequency or sweating). A so called coping strategy is developed that depends on the situation itself, as well as on an individuals cognitive structures. Coping strategies are for example, aggression, aversion, withdrawal, or denial of the situation. Feedback on success or failure of a strategy may help a person to selectively chose an according strategy.

- **Reappraisal:** In a third step, the success of a coping strategy is evaluated to allow the dynamic adaptation towards the new situation. If the person learns to cope with a thread (primary appraisal), it might turn out that the thread is rather a challenge, and not as threatening.

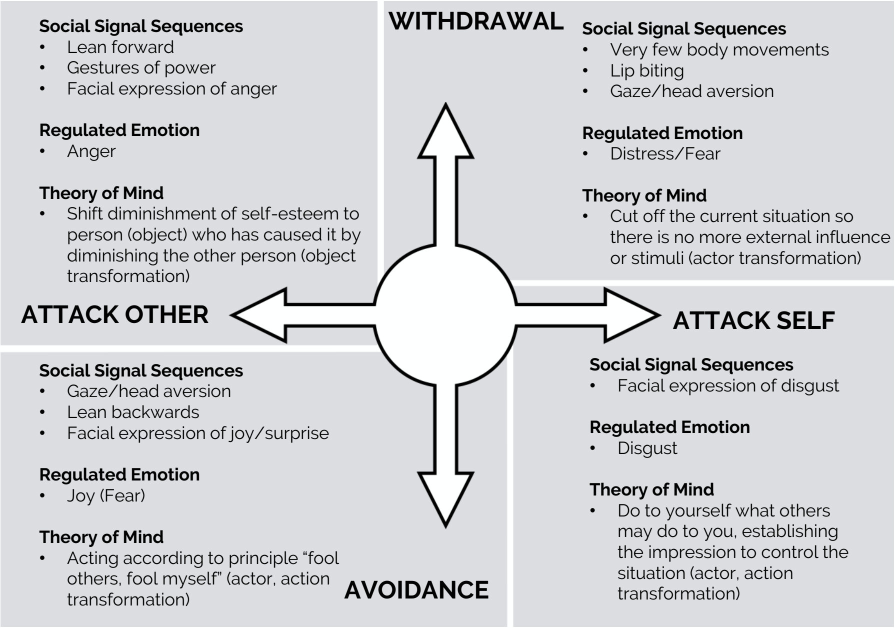
Coping strategies themselves can be divided into two main classes:

- problem-focused coping that has the goal to manage the stressor*
- emotion-focused coping which is concerned with one's affective responses to the stressor

**stress defined as a stimulus, is often referred to as a stressor*

Tomkins (1984) proposed that adult emotions are almost always regulated. The regulation of emotions describes the process of suppressing or changing emotions if they do not fit the current individual situation. The main purpose of the regulation process is to "cover" an unwanted emotion with others in order to (re-)establish the feeling of being secure in a particular situation (Tamir, 2011).

The regulation process changes the situational appraisal information, which elicits different emotions, reflecting a "better" (with regard to the individuals situational appraisal) coping of the situation. The employed regulation strategy changes situational values of an individuals internal situational representation.



ATTACK OTHER

Social Signal Sequences
- Lean forward
- Gestures of power
- Facial expression of anger

Regulated Emotion
- Anger

Theory of Mind
- Shift diminishment of self-esteem to person (object) who has caused it by diminishing the other person (object transformation)

WITHDRAWAL

Social Signal Sequences
- Very few body movements
- Lip biting
- Gaze/head aversion

Regulated Emotion
- Distress/Fear

Theory of Mind
- Cut off the current situation so there is no more external influence or stimuli (actor transformation)

AVOIDANCE

Social Signal Sequences
- Gaze/head aversion
- Lean backwards
- Facial expression of joy/surprise

Regulated Emotion
- Joy (Fear)

Theory of Mind
- Acting according to principle "fool others, fool myself" (actor, action transformation)

ATTACK SELF

Social Signal Sequences
- Facial expression of disgust

Regulated Emotion
- Disgust

Theory of Mind
- Do to yourself what others may do to you, establishing the impression to control the situation (actor, action transformation)

Figure 5.: Possible shame regulation strategies, related sequences of social signals, and explanation examples.

There is evidence that regulation processes can be observed through related social signals (Bänninger-Huber, 1996; Nathanson, 1994; Benecke, 2002; Schwab, 2000; Moser and von Zeppelin, 2005). For example, a *"theory of mind" (ToM)** model for the structural emotion "shame", was introduced by Nathanson. It takes clinical observations, individual background and information about personal motivations, and typical sequences of social signals of emotion regulation in account (Nathanson, 1994). For the regulation of the structural emotion shame, Nathanson describes four regulation strategies with related social signals and regulated emotions: *avoidance*, *attack self*, *attack other*, and *withdrawal* (Figure 5). Regulated emotions are partly expressed (as a communicative emotion) in the sequence of social signals that is related to the individually chosen regulation strategy.

**ToM is the competence to attribute emotions, knowledge, etc. to oneself, but also to others, and to understand that others have desires and perspectives that differ from one's own (Premack and Woodruff, 1978)*

For example, *withdrawal* is accompanied by head adaptors, lip biting, slight body movements, or averting head/gaze. *Avoidance* is accompanied by averting head/gaze or gaze wandering. Social signals indicating a regulation process can be similar as shown in the example. In the case of *attack other*, related social signals are directed gaze, spacious gestures/postures. Both, 1) the social signals of the regulation process (while processing the regulation strategy), and 2) the social signals of the regulated emotion compose an identifiable signal pattern that allows inferring the regulation process and strategy. In the case of *avoidance*, the regulated emotion is joy with the corresponding facial expression: *smile*.

2.4.3.2 Action Tendencies

A related concept to Lazarus coping strategies is the so called action tendencies. In the definition of Frijda (1986), action tendencies are "states of readiness to execute a given kind of action, (which) is defined by its end result aimed at or achieved". Emotions are, in this point of view, tendencies to engage in behaviour, influenced by the needs of the person in the specific situation. The difference between action tendencies and intentions is that they are "not goal-directed, but rather stimulus-driven" (Frijda, 1986).

Table 2.: Frijda's classification of relational action tendencies, source: (Frijda, 1986)

Emotion	**Function**	**Action tendency**	**End state**
Desire	Consume	Approach	Access
Joy	Readiness	Free activation	-
Anger	Control	Agnostic	Obstruction removed
Fear	Protection	Avoidance	Own inaccessibility
Interest	Orientation	Attending	Identification
Disgust	Protection	Rejecting	Object removed
Anxiety	Caution	Inhibition	Absence of response
Contentment	Recuperation	Inactivity	.

The "end state" differs between positive and negative emotions. For negative emotions, the "end state" that is aimed for, mitigates the experience of the negative emotion (for example, a person starts to feel save, once she or he starts to believe the object of fear is out of reach). On the contrary, positive emotions cause a person to get in a mode of relational action readiness. For example, the person could be ready to start a new interaction. Table 2 gives an overview on emotions, their functions, the according action tendencies and end states, where possible, according to Frijda (1986).

2.4.4 Emotional Intelligence

As a final aspect related to emotions we want to dedicate this section to *emotional intelligence* (EI). Gunderman (2011) defines *emotional intelligence* as "the ability to understand and respond to emotions in daily life". While some researchers argue that EI is learned, others state it is innate. In the definition of Mayer, Salovey, and Caruso (2004) EI describes "the capacity to reason about emotions, and of emotions, to enhance thinking. It includes the abilities to accurately perceive emotions, to access and generate emotions so as to assist thought, to understand emotions and emotional knowledge, and to reflectively regulate emotions so as to promote emotional and intellectual growth." Several models to comprehend emotional intelligence have been designed. Goleman (1998) outlines five main EI constructs, namely self-awareness, self-regulation, social skill, empathy, and motivation. Basically, these aspects describe the ability of being aware of and able to handle one's own, as well as other peoples' emotions. These concepts are directly related to "social attitudes", which describe our mindset towards others in social interactions. We will introduce social attitudes in more detail in the next section.

2.5 Social Attitudes

> *"People may hear your words, but they feel your attitude."*
>
> — John C. Maxwell

social attitudes are often also referred to as interpersonal attitudes

A concept, related - yet different - to emotions are *social attitudes*. As discussed earlier, social attitudes are also used in some dimensional emotion models as additional dimension (dominance, engagement, etc). Eagly and Chaiken (1998) define the term *social attitude* as a "psychological tendency that is expressed by evaluating a particular entity with some degree of favour or disfavour" . According to Rosenberg and Hovland (1960), a social attitude has three main components: affective, behavioural and cognitive.

- **Affective component** involves a person's feelings and emotions about an attitude object*. As an example, we take a person who has been invited to a job interview.

  Let us assume this person is very uncomfortable with the interview situation in general. In this case the interview situation represents the object the attitude is directed towards. Whenever this person is exposed to an interview or thinks about one she or he feels anxious and nervous. Those feelings form the affective component of a social attitude.

- **Behavioural component** refers to the way the attitude we have, influences how act or behave. Let us consider again the person in the job interview. Since the candidate is scared of the situation, he or she might show a behaviour that includes tense and nervous gestures, such as crossing arms and avoiding eye contact, especially when being asked difficult questions.

- **Cognitive component** involves a person's beliefs and knowledge about an attitude object. Now that we have seen how our job interview candidate behaves, the question arises of what he or she thinks about the interview. Probably, he or she thinks about being unemployed for a long time and the pressure of getting that job. Beyond the physical and emotional reactions to the situation, there is also the cognitive component of his or her attitude.

**attitude objects are entities we make judgements about or have feelings towards*

One can further distinguish between explicit and implicit attitudes (Greenwald and Banaji, 1995). Explicit attitudes appear at a conscious level. In that case, people are aware of them and usually know how they determine their behaviours and beliefs. For example, our candidate might have a negative attitude towards the interviewer, but tries to hide any negative feelings in order to get the job. On the opposite, implicit attitudes are at the unconscious level. In this case, our candidate would not be aware of his or her negative attitude towards the interviewer even though it might strongly influence his or her behaviour.

Often a person's social attitude is consciously or unconsciously reflected by their behaviour. In the context of our job interview scenario, the interviewer might conclude from the candidate's slouched body posture (behavioural component) that the candidate is bored (affective component) and finds the job unattractive (cognitive component). Overall, the candidate's behaviour portrays a negative social attitude towards the situation *job interview*.

2.5.1 Dominance, Friendliness, and Closeness

The conscious or unconscious evaluation of others' behaviour is related to an assessment of dominance, friendliness and closeness, one person has regarding another (Argyle, 2013). There are several models of interpersonal attitude. Argyle's model is based on two dimensions, derived from the two main aspects of interpersonal behaviour (Foa, 1961). The first dimension affiliation ranges from unfriendly to friendly. The second dimension status is related to power on a scale from submissive to dominant. A dominant person has the disposition to control others (Cashdan, 1998); a highly affiliated person is interested in a high friendliness and closeness level between him and the interaction partner (Kasap et al., 2009). Several aspects influence how the interpersonal attitude of a human interaction partner is perceived (Moskowitz, 1993; Foa, 1961).

2.5.2 Conversational Engagement

In earlier sections we used the term *engagement* implicitly (e.g. as a dimension in emotion models), but by looking at the literature, it turns out that there is no unique definition available. On the contrary there are many definitions of engagement and for related concepts such as interest, or involvement.

For example, Sidner et al. (2004) define engagement as *"the process by which two (or more) participants establish, maintain and end their perceived connection"*. According to their definition, this process not only includes the initial contact and beginning of a collaboration, but also functions to evaluate whether to stay involved or when a conversation should be ended. Bohus and Horvitz (2009) adopted Sidner et al.'s notion of seeing engagement as a process. They extended it with capabilities they deemed necessary for dealing with multi-party interactions. Therefore, they additionally characterised engagement as *"the process subsuming the joint, coordinated activities by which participants initiate, maintain, join, abandon, suspend, resume, or terminate an interaction"*. Poggi (2007) defines engagement as: *"the value that a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction."*. Here engagement is considered a value that is measurable, in contrast to a process as in the previous definitions. Another definition, specific to empathic engagement by Hall et al. (2005) is: *"empathic engagement is the fostering of emotional involvement intending to create a coherent cognitive and emotional experience which results in empathic relations between a user and a synthetic character."* Reck et al. (2011) consider engagement in the context of infant-caregiver engagement phases, where they associated positive engagement to positive emotional and interested expressions, such as facial expressions of joy, or gazing towards the other person. Negative engagement is associated to negative emotional coping strategies like withdrawal, hostility and protest (e.g. by showing facial expressions).

The definitions of engagement vary in some aspects, and sometimes are related to a specific context, but they all share a common ground. In their research, Glas and Pelachaud (2015) made some interesting implications about what various definitions of engagement share and where they differ. They concluded that in all the definitions, while sometimes implicitly and in other cases explicitly mentioned, "the level of *connection* and *cooperation* is positively related to the presence or level of engagement". They further conclude that due to the many definitions, there is also no consensus on the measurement of engagement. Some definitions measure engagement as a process to get to a certain state, others see engagement as a state (resulting from connection and cooperation) that can be estimated at any moment in time. In the latter, measurements of engagement can further be distinguished between those that assign a binary value to engagement (presence or absence), and those that imply a continuous notion of engagement (e.g. value or degree). Due to the strong dependency on context and the specific use case, they argue in favour of multiple coexisting definitions. Previous studies focus on a particular aspect of engagement without being able to cover nearly the entire range of interpretations. In Chapter 7 we present our own approach by using a probabilistic model for the automated detection of engagement based on signal processing techniques and various types of context information.

2.6 Conclusions

This chapter is meant to give an overview on the complexity of human social signals. Even tough some of the phenomena might sometimes be simplified and obviously can not cover each and every aspect of human non-verbal behaviour, the outline of the chapter is that the automated recognition alone, but even more the interpretation of such social cues, is an ambitious task. Even for humans it is sometimes not possible to identify all nuances and understand our counterpart's behaviours out of context. Especially when aiming to infer "high-level" interpretations such as emotions or social attitudes, we argue that aspects like earlier appraisals, interaction dynamics and external context need to be addressed.

In the next chapters we will introduce state-of-the-art approaches to "teach" computers (including agents and robots) how to recognise specific behaviours using machine learning. Therefore, we'll first have a look at multi-modal and multi-person corpora in different contexts. Social signals in such datasets are annotated are processed, so that machine learning models can be trained afterwards. After the next chapter we'll briefly introduce common social signal processing and machine learning techniques, before we present our approach to speed up the annotation process of continuous multi-modal data in a transparent way. In Chapter 7 we introduce an approach to model complex emotions and social attitudes.

MULTI-MODAL AND MULTI-PERSON DATA COLLECTION AND SOCIAL SIGNAL PROCESSING

3 MULTI-MODAL MULTI-PERSON DATA COLLECTION

"Data is a precious thing and will last longer than the systems themselves."

— Tim Berners-Lee - inventor of the World Wide Web.

For analysing humans in social interactions, we first need to gather an appropriate amount of prototypical data. This concerns equally both, the behavioural analysis research area, as well as automated social signal processing. Prototypical data here is a bit of a problematic term. Humans are very diverse, and depending on personality, culture, relationship, environmental surroundings and other factors the expressiveness of emotions and attitudes varies. For creating general automatic recognition systems, it is essential to rely on large databases, that include many variations of a problem at hand to fit as good as possible in new situations.

3.1 Challenges in the Creation of Corpora of Social Interactions

"It is a capital mistake to theorise before one has data."

— Sherlock Holmes. (Arthur Conan Doyle)

The quantity of training data is crucial for generalising social signal recognisers as much and as broad as possible, but what's even more important is the quality. Here, not only the actual technical quality of the raw streams is meant (in a sense of having a high resolution and being free of noise etc.) but the naturalness and reproducibility of behaviours. It used to be, and still is, common practise to analyse behaviours on data that has been recorded by professional (or not so-professional) actors. This of course has its justification for some valid reasons. It is easier to instruct a person to show certain signs of emotions, or perform specific actions, including multiple variations. Further it often is cheaper and less time consuming to rely on a few instructed people than gathering a large pool or participants. Yet, this allows quickly gathering a large number of examples for training a model. This procedure comes with two major problems. First, such recordings often happen in "perfect" conditions, lacking artefacts, disturbances and movements. As much as this sounds like a highly desirable scenario, when we want to apply a model, that has been trained on such data, in realistic environments where these conditions are not given, it will most likely fail to give correct predictions. The second problem is so called "overfitting". That means if we train our models on a couple of actors, it will most probably give decent results when it is applied on this group of people. In a simple example, imagine the actors are only 80 year old females, it is likely that the recognisers for many behaviours do not work as well on males, or on other younger females because they might for example move in a slight different way. This applies to almost any aspect of non-verbal behaviours in-between different humans. Not only as in this example, gender and age, but also cultural background, personality and situational context influence the way non-verbal behaviours are performed and are therefore context-dependent. To create more general models, examples from diverse groups of people are required in the training process. Researchers conclude that for the area of *social signal processing* to be successful, there is a growing need of annotated data of human interactions (Pantic, Pentland, et al., 2007; Vinciarelli, Pantic, Heylen, et al., 2012; Eerekoviae, 2014). Since, as discussed in Chapter 2, humans transmit non-verbal messages through a number of channels (voice, face, gestures, etc.) and due to the complex interplay between the channels (think, for instance, of a duchenne versus a non-duchenne smile), progress in SSP is directly linked to the availability of large and well described multi-modal databases rich of human behaviour under varying context and different environmental settings (Douglas-Cowie, Campbell, et al., 2003). Not least because state-of-the art algorithms, such as deep neural networks (DNNs) require - and directly improve with - large amounts of annotated training data (Sun, Shrivastava, et al., 2017).

Newer approaches like generative adversarial networks (GAN) aim to synthesise new training data based on a smaller training set (Goodfellow et al., 2014). Here, two models are simultaneously trained: a generative model G which captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than from G. The training procedure for G is to maximise the probability of D making a mistake. While this seems to be a promising approach for gathering more training data, it also comes with a couple of limitations. For example, a GAN can only learn to output data that contains the same variance and standard deviations of the data that it learned in the original training set. This means, newly generated data will contain more permutations of the same sample distributions it was originally learned on. Data-hungry algorithms such as neural networks will train better on the same distribution but this may also lead to over-training. Given the diversity and richness of human expressions and the complexity of human emotions, as well as their usage within various contexts and time sequences, such approaches seem to solve the original problem of gathering diverse training data only to a limited amount. This rather raises the need for new databases that "move from the simple collection of data samples to more realistic, non-acted databases" (Cowie, Douglas-Cowie, and Cox, 2005). Common challenges with creating such datasets lie in the high degree of naturalness demanded of the recording scenarios, how well one recording scenario generalises to others, the number of human raters needed to reach a consensus on labels, and of course sheer volume of data. In the next section we'll give an overview on some of the more prominent corpora containing data of social interactions in various contexts.

3.2 Existing Multi-modal Corpora

Unfortunately, the core of databases that have been collected in the past contain either acted behaviour recorded by few professional actors (e. g. the database of kinetic facial expressions (DaFEx) (Battocchi, Pianesi, and Goren-Bar, 2005) or the Berlin database of emotional speech (Emo-DB) (Burkhardt et al., 2005)) or isolated snapshots (e. g. Belfast naturalistic database (Douglas-Cowie, Cowie, and Schröder, 2000) or the "Vera am Mittag" (VAM) talk-show corpus (Grimm, Kroschel, and Narayanan, 2008)). The proper training of an online recogniser, however, requires long and continuous recordings collected under preferable natural conditions (Douglas-Cowie, Cowie, et al., 2007). An example is the SEMAINE corpus (McKeown, Valstar, et al., 2010), which is composed of 100 sessions (each about 5 minutes) of emotionally coloured, yet free conversations. SEMAINE has been successfully applied to many computer vision problems, such as facial muscle action (FAC) detection (Jiang, Valstar, and Pantic, 2011), head nod and shake detection (Gunes and Pantic, 2010), non-verbal communication estimation (Eyben, Wöllmer, Valstar, et al., 2011), segmentation (Nicolaou, Gunes, and Pantic, 2010) and emotions (Schröder et al., 2012). However, like other comparable corpora (e. g. SAL (Douglas-Cowie, Cowie, Cox, et al., 2008) or IEMOCAP (Busso et al., 2008)) it features audio-visual content only. There are numerous examples of such datasets, with a wide range of applications, such as speech recognition (Cooke et al., 2006), behaviour analysis (Zeng et al., 2008), segmentation, emotion recognition (Caridakis, Castellano, et al., 2007) and depression detection (Dibeklioglu, Hammal, and Cohn, 2018). Other relevant multi-modal datasets featuring dyadic interactions include: the Cardiff Conversation Dataset (CCDB) (Aubrey et al., 2013), an audio-visual database focusing on non-scripted interactions that do not predetermine the participants' roles (speaker/listener); the MAHNOB-Mimicry dataset, designed to analyse mimicry in dyadic scenarios where subjects act with a significant amount of resemblance and/or synchrony (Sun, Lichtenauer, et al., 2011) and the SEWA project, which include video-chat recordings, audio transcript and hand-gesture annotations of human interactions. Another example is the RECOLA (Ringeval et al., 2013) corpus that contains audio-visual recordings (but here also including physiological sensors) of collaborative, emotional video conference recordings in French. Furthermore, an interesting collection of multimodal datasets can be found at TalkBank (MacWhinney, 2007), a web-accessible database of audio-visual recordings of both human and animal communication. One of the major benefits of TalkBank is that it also includes the transcripts of the conversations recorded. Also a couple of multi-party datasets have been created. Popular examples of such datasets are the Belfast storytelling dataset (McKeown, Curran, et al., 2015), which collects spontaneous social interactions with laughter, and the AMI meeting corpus (Valente, Kim, and Motlı cek, 2012), which collects mult—modal data from recordings of meetings.

Given the richness of observable social expressions, but also the flexibility in sensory devices, there is still a huge lack of data. For instance, in none of the named corpora, motion capture information was considered as part of a social interaction. In this chapter we will introduce three novel muli-modal and multi-person databases containing samples of non-acted social signals which we have designed, recorded, annotated and provided to the research community.

3.3 The NOXI (Novice-Expert Interaction) Corpus

> *"Becoming an expert in almost anything requires literally years of work. People will do this only if they have some initial success, enjoy the work, and are supported by the social climate. Expertise is not solely a cognitive affair."*
>
> — Earl Hunt

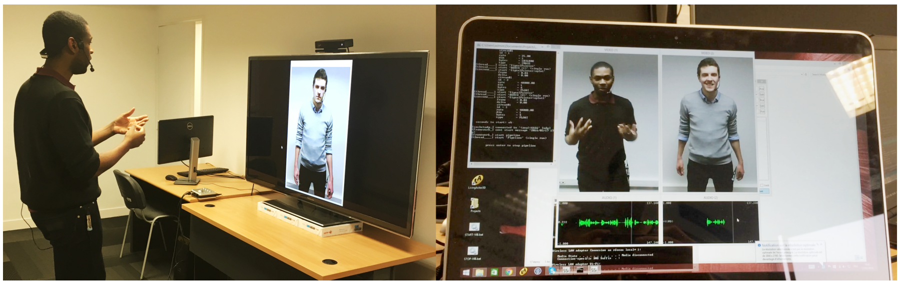


Figure 6.: Snapshots of user interaction (left) and observer screen (right) during a recording session for the NOXI database.

This section is based on the publication: Cafaro, Wagner, Baur, Dermouche, Torres, Pelachaud, André, and Valstar, 2017

This section presents **NOXI**: the **NO**vice e**X**pert **I**nteraction database. NOXI is a corpus designed for studying and understanding human social behaviour during an information retrieval task targeting multiple languages, multiple topics, and the occurrence of unexpected situations. It contains screen-mediated face-to-face interactions recorded at three locations (France, Germany and UK), spoken in eight languages (English, French, German, Spanish, Indonesian, Arabic, Dutch and Italian) discussing a wide range of topics. Our first aim was to collect data from natural dyadic knowledge exchanges in the context of the Aria Valuspa project (Valstar, Baur, et al., 2016) about retrieval of information supported by assistants in the form of virtual anthropomorphic agents with linguistic and social skills.

NOXI has been designed from the beginning with the aim of being used by a wider audience of researchers in a variety of applications other than information retrieval (see Section 3.3.1). The dataset offers over 25 hours (x2) of recordings of dyadic interactions in natural settings, featuring synchronised audio, video, and motion capture data (using a Kinect 2.0). We aimed to obtain data of spontaneous behaviour in a natural setting on a variety of discussion topics. Therefore, one of the main design goals was to match recorded participants based on their common interests. This means that we first gathered potential experts willing to share their knowledge about one or more topics they were knowledgeable and passionate about, and secondly we recruited novices willing to discuss or learn more about the available set of topics offered by experts.

Eliciting unexpected situations was another emphasis in creating NOXI, and therefore it includes controlled interruptions made by a confederate to one of the participants during the recordings as well as spontaneous interruptions made by one of the interactants in a subset of recordings. Efforts have been made to add manual and automatic annotations to the database. Both discrete and continuous annotations are created including low level behaviour (e.g. head movements, gestures, etc., see Section 2.2) and high level user states such as valence and arousal (see Section 2.4.2.2 or engagement (see Section 2.5.2), as well as, speech transcriptions on word and sentence level. A web interface is publicly available, thus allowing the research community to search the database using a variety of criteria (e.g. topic, language, etc...) and download the data. In this section, NOXIs design, availability and annotations are presented.

3.3.1 Data Collection

The idea behind NOXI was to obtain a dataset of natural interactions between human dyads in an expert-novice knowledge sharing context. In a recording session one participant assumes the role of an expert and the other participant the role of a novice. When recruiting participants, potential experts offered to discuss about one or more topics they were passionate and knowledgeable about, whereas novices applied to a recording session based on their willingness to discuss, learn more and retrieve information about a topic of interest among those offered by experts. A matching of interests was found when a novice chose an expert's topic, then the dyad was recorded. This served the purpose of obtaining spontaneous dialogues on a variety of different topics for which the participants were passionate/-knowledgeable about.

3.3.2 Design Principles

We prepared the recording protocol with the following design principles.

- **Setting:** We opted for a screen-mediated recording for a twofold purpose: first it allowed us to record a face-to-face conversation without the need of multiple cameras recording from different angles as in classical face-to-face settings of other corpora (e.g. (McKeown, Curran, et al., 2015)). Secondly, this setup is closer to a scenario where a virtual agent is displayed on a screen. Participants were recorded while standing to meet the second design principle listed next.

- **Data:** We aimed at capturing full body movements (e. g. postural changes, torso leaning) in addition to facial expressions, gestures and speech. Furthermore, we wanted to enrich the dataset with the above mentioned data and possibly new formats not being captured in existing databases (e. g. Kinect depth maps).

- **Interaction:** We wanted to record spontaneous interactions but at the same time we were interested in measuring both participant's engagement in the interaction and occurrences of unexpected events (e. g. interruptions). Participants were allowed to continue their conversation until it reached a natural end. This results in quite long interactions for such a database: the minimum duration for a recording session was set to seven minutes, the maximum to almost 40 minutes.

- **Participants:** The recruitment was carried out in two stages. First we gathered potential experts that offered their availability on several topics they were knowledgeable and interested about. We then recruited novices by proposing them the schedule of available experts and their topics, and we let participants select multiple topics they were interested in. This way, we obtained a natural matching of dyads between novices and experts driven by their own interests. We primarily recruited in our research facilities, but also from our immediate social surroundings, as well as on social networks like Facebook. Therefore we obtained dyads of both colleagues and friends, as well as entirely unacquainted persons, thus providing us the opportunity to record a number of zero-acquaintance situations (Ambady, Hallahan, and Rosenthal, 1995).

- **Unexpected Events:** One of the aims was to obtain occurrences of unexpected events. We primed both participants before recording a session by encouraging them to interrupt each other, provide opinions, suggest slight topic changes and induce a mild debate whenever possible. The instructions for both, the expert and novice can be found in Appendix B. Moreover, we artificially injected an unexpected event during the recordings. More specifically, we introduced two possible functional interruptions (e.g. pretending that the microphone was not in the good position) that would result in an unexpected event, for the expert, from an external source (i.e. not being within the interaction or caused by the interactants). For this purpose, we informed the novice about the possibility to (1) call him on his/her mobile phone (i.e. CALL-IN) or (2) physically enter the recording room (i.e. WALK-IN).

3.3.3 Recording Protocol

The recording protocol had several steps. We first received participants in two different rooms. One room was for the novice (Room I) and one for the expert (Room II). Room III was used to monitor the session and synchronise the data collection. We primed the novice about the functional interruption as described above. In zero-acquaintance situations we did not introduce participants to each other, therefore their very first interaction happened when they both saw each other on the displays. We read instructions to both participants, set up their microphone and showed the position where to stand (also indicated by a marker on the floor) and prior to begin the recordings we obtained their informed consent.

Participants were informed about the sole possible usage of the recorded data for scientific research and non-commercial applications. Moreover, they had three (non exclusive) choices concerning the usage of their data: (1) data available within Aria-Valuspa consortium only, (2) data available for dissemination purposes in academic conferences, publications and/or as part of teaching material, and (3) data available for academic and non-commercial applications to third-party users through the web interface described later in this section.

The session was monitored in Room III and when both participants agreed to end, the experimenter(s) gave participants questionnaires (described in Section 3.3.3.2). Finally, participants were debriefed and compensated.

3.3.3.1 Recording System

We used the Microsoft's Kinect 2 as recording devices. Kinect supports the capture of video streams in full HD quality and provides optical motion capturing to track the body and face position of the

Table 3.: List of the recorded signals. Audio signals were sampled at 48 KHz, video signals at 25 fps.

Sensor	Channel	Resolution	Depth
Kinect	Audio	mono	16 bit signed
	Video	1920 x 1080	24 bit unsigned
	Depth	512 x 424	8 bit unsigned
	Skeleton	25 joints	32 bit float
	Confidence	25 values	32 bit float
	Face	1347 points	32 bit float
	Head	Pitch, roll, yaw	32 bit float
	Action Units	17 values	32 bit float
Headset	Audio	mono	16 bit signed

user. The inbuilt microphone was used to capture the ambient sound in the room. Additionally, to obtain low-noise recordings of the voice we equipped users with a dynamic head-set microphone (Shure WH 20 XLR connected through a TASCAM US-322). The setup, was distributed over three rooms. The rooms for the novice and the expert were equipped with a Kinect device put on top of a 55″ flat screen. Kinect and headset are plugged to a PC (i7, 16 GB RAM). In each room a local hard drive (2 TB) was used to store the captured signal streams. A third PC was put in the observer room to monitor the interaction. The three PCs were connected in a wired local network.

The signals separately recorded for each user are listed in Table 3. Skeleton data had 14 values per joint, whereas for the face we had 3 values per point. The raw captured streams would require 154 MB of drive space per second for a single user. To ease the storage load two compromises were made: (i) the size of the HD video stream was reduced by applying the lossless Ut Video Codec by Takeshi Umezawa[1]. The algorithm builds on the Huffman code, but allows a better compression. Since it runs on multiple threads and uses SSE2 assembly, it is fast enough to compress HD videos in real-time. (ii) the size of the depth images was reduced by decoding each pixel with only one instead of two bytes (i. e. depth values are expressed in the range of [0...255] instead of [0...60000]). The remaining streams (2 x audio, skeleton, tracking confidence, face, head, and face animation units) were stored uncompressed. With the aid of these measures we were able to reduce the bandwidth from ~9.3 GB to ~1.4 GB per minute per user.

To stream audio-visual information from the novice room to the expert room and vice versa, as well as from the novice and expert

[1] http://umezawa.dyndns.info/archive/utvideo/

room to the observer, we needed a very fast and efficient streaming protocol. Due to its popularity in streaming applications, we decided to use the h264 codec provided by the ffmpeg project[2]. For the sake of speed we also decided not to use a streaming server, but to stream directly to the receiver(s). Because of the way the experiment was designed we assumed that participants would stay more or less in the marked spot throughout the recording. Given the horizontal orientation of the video image, we decided to crop the streamed images from full HD to 480 × 720 pixels, thus discarding unused parts (i.e. left and right). Figure 6 shows an expert during interaction with a novice.

To keep recorded signals in sync we rely on a two-step synchronisation. Once all sensors are properly connected and provide a stable data stream, we use a network broadcast to simultaneously start the recordings in the novice and expert room. In the following we then ensure that the captured streams keep a stable sample rate to avoid time drifts between the individual signals. The latter is achieved by regularly checking the number of received samples against the number of expected samples. In the cast that a discrepancy is observed either missing samples are added or additional samples are removed.

The described system was implemented with the Social Signal Interpretation (SSI) framework (see Section 4.2)

3.3.3.2 Collected Data

The experiment was conducted in three different countries – France, Germany and UK. The primary reason for recording in three demographically different locations was the aim of collecting large numbers of interactions of the three languages. In addition to English, French, and German, we also collected a smaller number of recordings of five other languages (Spanish, Indonesian, Italian, Arabic, Dutch). A summary of the recorded sessions is given in Table 4. For the three main languages English, French, and German, we had 40, 25, and 19 interactions. In total, 87 people were recorded during 84 dyadic interactions (some people appeared in more than one session). The total duration of all sessions was 25 hours and 18 minutes (resulting in over 50 hours of data overall).

We also collected demographic information of our participants at the end of each session. These data consisted of their gender, cultural identity, age and education level. The cultural identity was obtained by asking participants to select the country that most represented their cultural identity from a list of all countries in the world. Participants' age is in the range of 21-50 years old.

In addition to demographic information, participants provided a self-assessment of their personality based on the Big 5 model (a.k.a.

[2] https://ffmpeg.org/

Table 4.: Overview of NOXI recordings - Sessions. From left to right: place, number of recording sessions, number of participants (female/male), average and standard deviation of recording duration (mm:ss), total duration (hh:mm).

Place	Sessions	Participants	Avg Dur	Std Dur	Total Dur
DE	19	29 (05/24)	17:56	05:56	05:38
FR	25	32 (10/22)	20:15	06:51	08:26
UK	40	26 (11/15)	16:50	06:41	11:13
Total	84	87 (26/61)	18:06	06:28	25:18

OCEAN) (McCrae and Costa, 1997) by using Saucier's Mini-Markers set of adjectives (Saucier, 1994).

Finally, we collected session specific information that included the social relationship level between participants, the level of expertise on the discussed topic and the proficiency level of the language spoken for that session.

We were very pleasantly surprised by the large diversity of topics covered by our experts. A total of 58 topics were discussed. English sessions had a large variety of topics including travels (5 sessions), technology (4), health (3), cooking (3), sports (2), politics (2) and many others. French sessions were mainly about video-games (4), travels (3), music (3) and photography (2). Finally, German sessions included expert computer science fields (5), various sports (6), car driving, magic tricks and other trivia.

3.3.4 Annotations

Since the recording of NOXI, considerable efforts have been spent to collect descriptions of the database. Due to the large amount of data we wanted to allow multiple annotators from several sites to contribute and share their labelling efforts in an easy and clear way. In addition and to further speed up the process, we decided to make use of semi-automated solutions to accomplish the desired descriptions. In Chapter 5 we introduce the NOVA tool, which we have implemented for this purpose.

More than 30 annotators from three countries (UK, France, Germany) are involved in the annotation of the NOXI database. They use the NOVA tool as a platform to create and share their annotations. Depending on the requirements of the project partners, different annotations were either created by the single groups, or annotation tasks have been coordinated between partners. To reduce human efforts where possible, we also apply automated or semi-automated methods where applicable. In NOXI we created annotations for multiple modalities and with various levels of automation:

Since participants were equipped with close-talk microphones only little background noise can be observed in the voice recordings. Hence, voice activity detection (VAD) was implemented by first normalising the waveforms and afterwards applying a threshold to the intensity. Comparing the sequence of speech segments of both interlocutors makes possible studying their turn taking and interruption strategies. However, completely ignoring the semantic context of speech can lead to wrong annotations. For instance, overlapping segments (i.e. where both interlocutors talk simultaneously) may not necessarily signal an attempt to take the floor but can be a sign of backchanneling, too. Hence, VAD annotations were refined by marking filler events such as hesitation (e.g. "uhm") and backchanneling events (e.g. "ok"). We consider such fillers indicators of a person's turn-taking strategy, but also as indicators of a person's engagement in the conversation. Once a sufficient number of annotations had been manually processed, cooperative machine learning (more details in Chapter 6) was used to predict the annotations for the remaining sessions.

Skeleton tracking from the Kinect 2 was used to assess gestures and movement quality. Gesture recognition was accomplished with the full body interaction framework FUBI (Kistler et al., 2012). It supports the recognition of static postures and dynamic gestures by comparing each skeleton frame against a series of recognition automata. To define the automata, FUBI offers a user-friendly XML based language. From the body we extracted gestures such as arm crossing or leaning front/back; from the head gestures such as nods and shakes. To capture the dynamic properties of the movements we calculated several expressiveness measures such as energy, fluidity, or spatial extent (see Section 2.2.1.3).

The emotional state of the user is currently described along the two affective dimensions valence and arousal. Each dimension is represented by a continuous score between 0 and 1, which we automatically derive from the user's voice and face. To train the speech models we relied on the "Geneva Multimodal Emotion Portrayals" (GEMEP) corpus (Bänziger, Mortillaro, and Scherer, 2012). It contains 1.2 k instances of emotional speech from ten professional actors, which we used to train support vector machines classifiers using the popular LibSVM library (Chang and Lin, 2011). While we argue that acted corpora often lead to wrong assumptions for many phenomena in natural settings, for the initial models we rely on the GEMEP corpus as it is widely used for the task at hand. As feature set we took the ComParE 2013 set (Schuller, Steidl, et al., 2013), which has been extensively demonstrated to be suitable for a wide range of paralinguistic tasks. A total of 6373 features were extracted on a per-chunk level using the OpenSMILE toolkit (Eyben, Weninger, et al., 2013). Whenever

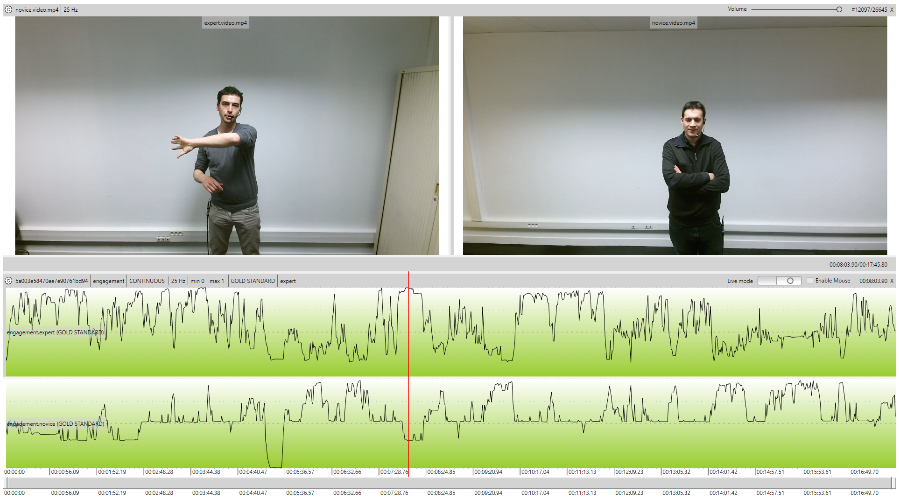


Figure 7.: An example of the Gold standard annotation of engagement. The expert's (on the left) engagement is shown on the upper tier, the novice's (on the right) engagement is shown on the lower tier.

necessary, the annotations have been corrected manually, to reflect the non-acted setting at hand.

Manual transcriptions on word and sentence level are available for different languages. These transcriptions help fine-tune the acoustic and language models, as well as, estimate the performance of the automatic speech recognition (ASR) system implemented in Aria-Valuspa.

Finally, we manually labelled the engagement of participants on a continuous score from 0..1. The annotations are based on multi-modal observations and are performed considering the videos of both interlocutors. This way, raters are not limited to isolated samples but consider context information already during the annotation process. We calculate the inter-rater agreement between multiple annotators and create a gold standard annotation based on the most matching annotations (also see Section 5.3.5.3). An example of an engagement annotation of both, the expert and novice can be seen in Figure 7.

3.3.5 Availability

NOXI is made freely available to the research community and for non-commercial uses. In a web interface, the data is organised in sessions that correspond to an expert-novice full recording with audio-visual data and annotations. The total size of the database is approximately 4 TB, however the database is searchable through the web interface. A user can select multiple criteria (e.g. language or topic of the session, participants' gender, etc...) and can choose the sessions to download from a list or results.

NOXI is available for download at: `https://noxi.aria-agent.eu/`

Furthermore, after selecting the sessions from the search results, a user can choose the files that s/he wishes to include in the download for the expert and the novice. Search results can be saved and shared with other NOXI users of the web interface. This is implemented through the notion of collections, which are predefined sets of sessions grouped by one or more criteria. Users can create a collection starting from the results of a research in the database.

3.4 Further Multi-modal and Multi-person Corpora

Similar to the NOXI corpus, we recorded various other multi-modal and multi-person corpora for a range of other contexts. While we won't go as much into detail for these corpora, we give a brief introduction on two corpora to highlight the phenomena that appear in changing contexts. The first one is concerned with gathering multimodal laughter data, the second corpus aims to collect data of people showing various shame-regulation strategies.

3.4.1 Multi-Modal Multi-Person Corpus of Laughter in Interactions (MMLI)

This subsection is based on the publication: Niewiadomski, Mancini, Baur, Varni, Griffin, and Aung, 2013

Within the EU-founded project ILHAIRE, which has been concerned with the science of laughter, we recorded the *multi-modal and multi-person corpus of laughter in interactions* (MMLI) (Niewiadomski, Mancini, Baur, et al., 2013). The aim of the MMLI corpus was to collect multimodal data of laughter with the focus on full body movements and different laughter types. It contains both, induced and interactive laughters from human triads. In total we collected 500 laugh episodes of 16 participants. The data consists of 3D body position information, facial tracking, multiple audio and video channels as well as physiological data. Laughter is one of the most commonly appearing human communicative signals (Chapman, 1983). It is often associated with spontaneous reactions to humorous stimuli. However, laughter may also have other roles e.g. expressing social attitudes such as irony, or regulating conversation e.g. when used for backchanneling (Provine, 1996). Despite its high incidence, knowledge about the multi-modal expressive pattern of laughter is rather limited. It further is a rather complex behaviour that includes the majority of expressive modalities. Most of today's research focuses on acoustic and facial cues of laughter. However, these are often accompanied by body movements and changes in posture (Ruch and Ekman, 2001) including, among others, head backwards movements and trunk/shoulders vibrations caused by forced exhalations.

To collect multi-modal data we built a complex set-up that allowed us to collect the information from different sources. First of all, three high precision inertial motion capture systems were used to collect high quality data of body movements. These systems were complemented by Microsoft Kinect sensors, high frame rate cameras and a respiration sensor. All the data is synchronised through the SSI framework (see Section 4.2). This allows to analyse not only synchronisation between different modalities in a laughter episode but also intra-subject synchronisation.

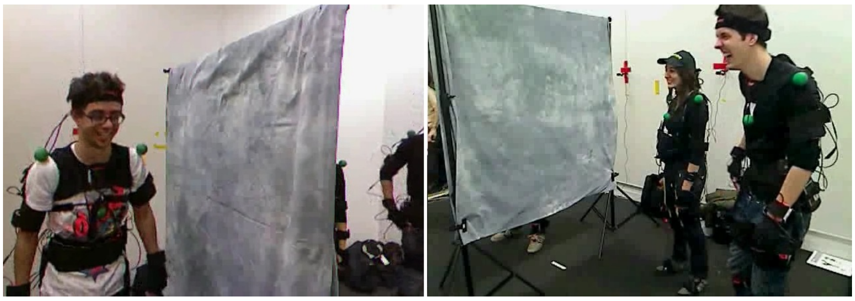


Figure 8.: The views from two synchronised wall cameras in a laughter inducing task.

To capture laughter in different contexts and various laughter types we invited groups of friends and asked them to perform six enjoyable tasks. Beside classical laughter inducing tasks such as watching funny clips we proposed participants to play several simplistic social games, i.e. games regulated by one simple general rule in which participants are free to improvise. We supposed that a lack of detailed rules could encourage easy-going spontaneous behaviours that may include reactions such as commenting, joking, irony, or even embarrassment or schadenfreude. Thus, we expected that the resulting data could consist not only of enjoyment laughter, but also of some other laughter categories. In total, the project partners annotated 439 laughter events, corresponding to 31 minutes of laughter, that is, 12% of total recording time (4 hours and 16 minutes). More details on the annotation process can be found in (Niewiadomski, Mancini, Varni, et al., 2016). This rate is not particularly high because a raw data contains recording of whole sessions. It is important to notice that the rates obtained in other laughter data collections are not much different: 5-8% in meeting recordings (FreeTalk, (Scherer, Schwenker, et al., 2009)), 18% in an laughter inducing study (AVLC, (Urbain et al., 2010)).

The MMLI corpus is freely available for research purposes on the ILHAIRE database website: `http://qub.ac.uk/ilhairelaughter/homepage`.

The MMLI database is the first corpus of this richness in different laughter contexts, containing various data sources (motion capture, depth, audio, video, physiological), a large spectrum of captured modalities and that is synchronised across multiple participants. The proposed scenarios were successful in eliciting the laughter in our participants.

3.4.2 EmpaT: Shame-eliciting Job Interview Corpus

This subsection is based on the publication: Gebhard, Schneeberger, Baur, and André, 2018

In the BMBF (German federal ministry of education and research) founded project EmpaT we created a semi-natural corpus of shame eliciting situations during a job interview situation. In a pre-study two job coaching experts identified six possible shame eliciting situations considering Nathanson's work (see Section 2.4.3). 26 participants (age 18 - 29, *M* = 21.71, *SD* = 2.91) were asked to put themselves into a position of a job applicant experiencing these six different situations. The task of the participants was to describe in their own words how they would react. The answers were analysed by two psychologists and assigned to Nathanson's four shame regulation strategies. Finally, we identified five situations that elicit the structural emotion shame, e.g. *"Before we begin, let me ask a short question: where did you find your outfit? It really doesn't suit you!"*, or *"Other candidates did a better job answering the question!"*

To generate our corpus, we created a 15-minute job interview with the five shame eliciting situations from the pre-study. This job interview was conducted from a female interviewer with 20 participants (10 female, age 19 - 30, *M* = 24.60, *SD* = 4.08). After welcoming the participants, they were asked to practise an application for a student assistant job in their favourite faculty. The participants were sent to the office of the interviewer to be interviewed. Afterwards, the participants answered demographic questions and were compensated. The interviews were recorded with a Kinect2 depth camera and a head-mounted microphone.

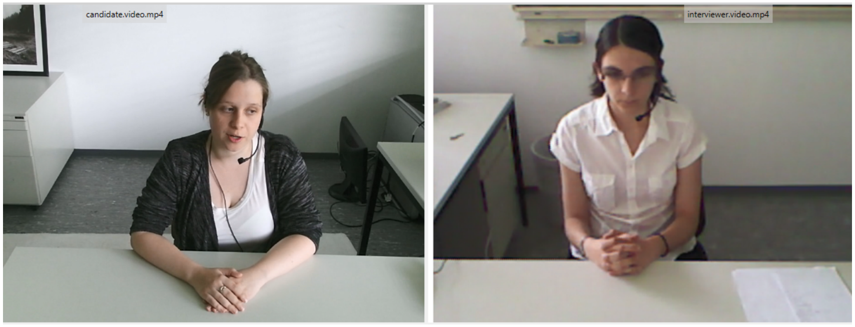


Figure 9.: A job candidate (left) is asked shame-eliciting questions by the interviewer (right)

In total, 100 (20 participants in five situations) shame eliciting situations are included in the corpus for the analysis. We annotated the obtained data in terms of social cues and emotion regulation strategy. Each situation was classified independently by three students, that were not related to the experiment neither knew about the aim of the study. They were trained beforehand to classify Nathanson's four shame regulation strategies. Overall, 300 labels were assigned as follows: 83 Withdrawal, 105 Attack Self, 98 Avoidance and 14 Attack Other. For assessing the reliability of agreement Fleiss' kappa (see Section 5.3.5.2) was calculated for three raters, four labels and 100 data points. With 0.7301 it is considered as substantial agreement.

3.5 Conclusions

For improving the automated recognition of social behaviours, still a lot more databases, containing various aspects of human behaviours are required. The increased interest in multi-modal corpora is reflected by various scientific conferences and workshops such as the international workshop series on multi-modal, tools and resources[3] or the SSPnet portal[4] that came up in the last years. The interest in research ranges from collection efforts, coding, validation and analysis methods, to tools and applications of multi-modal corpora.

We introduced three novel multi-modal and multi-person corpora containing various social cues in natural interactions and different contexts. It is of vast importance that signals are synchronised during recording, so that cues from multiple channels, but also in between users can be aligned and used for later interpretation and analysis. The introduced databases are designed with a specific goal in mind, yet the interactions between users are natural which is a requirement for machine learning models to work in real-life scenarios. The NOXI corpus will be used in later chapters to illustrate multiple aspects of automated recognition modules and serves as evaluation corpus for different algorithms. In the next chapter we we briefly introduce common methodologies for the automated recognition and analysis of social signals and the SSI framework that was used for the recording of the three introduced corpora. Chapter 5 will focus on the challenges for the annotation of such large continuous multi-modal corpora.

[3] http://multimodal-corpora.org/

[4] http://sspnet.eu/

4 SOCIAL SIGNAL PROCESSING

"Computers are useless. They can only give you answers."
— Pablo Picasso.

4.1 How Computers Recognise and Learn Social Signals

As discussed in the motivation chapter of this thesis, the one huge challenge in the area of *social signal processing* is the automated recognition an interpretation of social cues. When communicating with computers, nearly all of our communication relies on explicit commands only (clicks, touches, directed keywords). This is in contrast to the intuition of human communication, which is mostly based on implicit interaction. Aiming for more intuitive interaction is therefore an important goal of research on next-generation human-computer interfaces (Pantic, Nijholt, et al., 2008). However, intuitive interaction requires the computer (computer stands in our case also as a synonym for virtual avatars and robots) to perceive implicit user behaviour. Therefore, we have to equip machines with tools that allow the recognition and interpretation of social signals from various modalities. Developing such tools, able to detect and react to user behaviour in real-time, involves a number of challenges:

- **Synchronisation of signals in multiple modalities:** As humans we are naturally equipped with a variety of senses. We have eyes to see, ears to hear, skin to feel, etc. Our brain is capable to process and align such stimuli in parallel. Similar to the set of sensing organs in biological organisms, a variety of sensor devices have to be used to record various signals that carry social cues. Microphones replace the human ear and capture the human voice and other sounds. Video and depth cameras replace human vision and allow spotting humans to analyse their behaviours. Motion sensors worn by the user measure body posture and motion at very high precision, while physiological sensors monitor physiological signals, such as heart rate or respiration. Signals received by such various types of sensors differ in quantisation and sample rate. To combine information gathered from such varying sources, raw data streams must be synchronised properly.

- **Uncertainties:** Compared to computers, which follow exact defined rules and mechanisms, human communication is rather chaotic. Behaviours are often ambiguous and the task of interpreting them not always a straight-forward one. Therefore, automatic recognisers should make use of probabilistic models for recognising and interpreting observed behaviours. In general, recognition of social signals consists of : (a) *pre-processing* to separate relevant information on a signal from irrelevant (b) *data segmentation*, to detect on- and offsets of actions, which hold relevant information about the user's intention and goals. (c) *feature extraction*, which relates to the reduction of a raw sensor stream to a set of concrete features to keep and consider only relevant information of a signal for further steps. (d) *classification*, for mapping a observed feature vector onto a set of discrete classes or continuous scores. A large set of samples is required to fulfil these tasks. Especially in real-time non laboratory settings, high amount of realistic data is required.

- **Fusing multi-modal data:**

  To analyse the complex of human behaviours, information gathered from observing multiple modalities need to be combined to analyse and interpret the full complex of behaviours. This may happen at various abstraction levels. For example, channels may be fused already at data level (e.g. a monochrome high resolution camera is filled with colour information from a second camera). The next abstraction is at feature level when features of multiple channels are combined in a single feature vector, e.g. when considering the audio and video channel to recognise laughter and smiles as enjoyment value. A third abstraction is at decision level, that means the decisions of multiple classifiers that happen simultaneously are combined. The highest abstraction is on event level. It is especially used when social cues happen with a slight offset, or when social cues are detected with various classifiers on an event level. For example someone speaks and then turns away the head. To chose which abstraction level should be used widely depends on the kind of information and the application of interest.

- **Real-time processing:**

  In interactions with conversational systems, we expect reactions to happen as fast and natural as possible. From the recognition point of view, therefore information gathered from sensor devices needs to be processed in real-time (also called online). That means for example, sample values are at the same time constantly read by sensor devices while feature extraction and recognition are applied simultaneously to these samples. The window length on which these steps happen depend on the type of signal. In early processing steps, mostly overlapping small windows of a few milliseconds are applied, while in later processing steps typically segments of a few seconds are common. Real-time processing comes along with further challenges compared to offline training and classification.

4.1.1 From Analogue to Digital Signals

We are confronted with signals all the time, and we as humans have "sensors" to receive and "actors" to send such signals, e.g. when we communicate, we are able to perceive and decode vibrations in the air as sounds with our ears, and we are able to encode information back to vibrations with our lips, tongue and vocal cords. This of course does not only concern audio waves but also visual images and other modalities. Generally speaking, any information about a physical quantity that changes over time could be considered as a signal (Sinha, 2009). By measuring the state of this physical quantity in regular intervals we receive a curve, as for example the one in Figure 10.

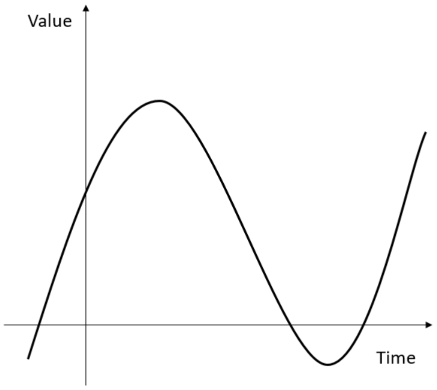


Figure 10.: Continuous physical signal.

4.1.1.1 Sampling

This curve is rather theoretic, because in fact, both humans and technological sensor devices are not able to measure a signal with an infinite degree of detail. Instead, a signal is observed and measured in constant, regular time intervals. Each of these measurements represents one *sample* of the observed signal. The frequency in which samples are measured is called the *sample rate* and is referred to as *Hertz* (Hz) or in *frames per second* (FPS). On the value axis the current measured sample is quantified to the nearest discrete value of the *target resolution*, because just as on the time axis, the value axis has theoretically an infinite degree of detail.

The resolution for example might be 1 byte which consist of 8 bit, resulting in $2^8 = 256$ different levels we could map our signal to. Using a higher resolution, e.g. short (16 bit = $2^{16} = 65536$) or int/float

(32 bit = 2^{32} = over 4 billion values), allows to represent the signal with a higher precision. Figure 11 illustrates the analogue signal from the previous example, quantified in discrete time and value units.

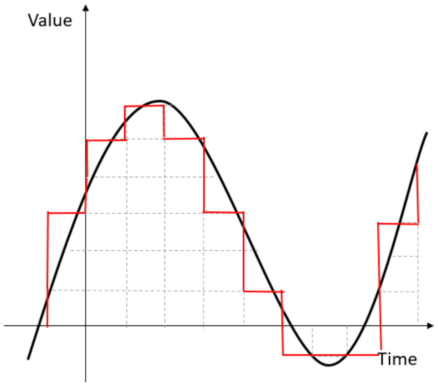


Figure 11.: A digital signal (red) represented with a sample rate and target resolution

As seen in this figure, the overall signal (in red) is captured, but it differs from the original signal to a wide extend. By increasing either the target resolution or the sample rate - or preferably both - the digital representation of the signal approaches the original curve. Which resolution is to chose depends on the signal. If the signal converter is able to represent signal levels below the background noise (measured in signal-to-noise ratio (SNR)) additional bits will no longer contribute useful information, so for the sake of space and resources, lower resolutions would then be sufficient and should be preferred.

4.1.1.2 Streaming

Once we measured single samples in our original signal, we use these samples to represent the signal in a digital form. The representation of the signal is called a *stream*. Each sample in our stream might have one value, as in the previous example, but also multiple values that are represented by *dimensions*. For example if we measure the user's body movements with a Microsoft Kinect sensor, the sensor processes information about 25 body joints, each containing further information, like their x, y, z position, rotation and tracking confidence (14 values in total) leading to a dimension of 25*14 = 350 per sample. Running at a sample rate of 30Hz, the sensor delivers 350 * 30 = 10500 values each second. And finally, each value has the target resolution of a float value, respectively 32 bit.

4.1.2 Signal Processing

Now that we converted our real-world signal into a digital stream, we are able to further process it. Depending on the sensor, the raw data stream might already be preprocessed (like in the previous example, the Microsoft Kinect sensor, already provides a pre-calculated skeleton), other sensor streams require more processing steps. That is due to the fact that sensors like a microphone or camera do of course not only capture human behaviours, but also any other physical quantity. Algorithms that extract relevant information from these general observations are needed.

4.1.2.1 Pre-Processing

Typically, a first step when processing a raw stream is to enhance and filter the signal. This process is called *pre-processing*. In controlled laboratory conditions (e.g. where we have a white background in our camera recording) this step may be less important, but in naturalistic settings, with a lot of background noise it is essentially important.

There a multiple rather "standard" procedures to pre-process streams. An example is using filters to cut of values above or below a certain threshold to remove artefacts. Finding such thresholds is not always straight-forward, because especially in real-time applications with changing conditions they might be variable.

Noise reduction (Schuller, Wöllmer, et al., 2009) as another example is applied to recover an original signal that got corrupted, or has background noises that would make it hard to further process the signal. Bandpass filtering (Proakis and Manolakis, 1992) allows to focus on frequency bands in which one would expected usable information.

Often, before applying such and similar transformations, a signal needs to be transformed to another value range or domain. This happens e.g. by normalisation to a common scale or by transforming the signal into the frequency domain. Also, sometimes the sample rate of a signal is reduced to save further processing time.

4.1.2.2 Segmentation and Activity Detection

Depending on the task, a simple activity detection might be already enough to identify behaviours. An example is *voice activity detection*. Imagine we transformed the audio signal to calculate statistical measures like the energy. When the energy value goes above a certain threshold (that depends on the microphone) we can already be sure that the audio channel is not silent anymore (we can not be sure the signal is really voice either, as noise or breath can trigger the energy, but this will be addressed in later chapters). If we now want to calculate audio features we can concentrate on the parts that are not *silence*. Another example is in facial feature detection, algorithms try to match certain prominent points of the face to a pre-trained model. Background information is not relevant as only the area around the face is of interest. Therefore algorithms often first try to find a face in the image before having a more detailed look at single facial landmarks.

4.1.2.3 Feature Extraction

A final step is the reduction of the signal to relevant features. A *feature* in machine learning is considered as a characteristic measurable property (Bishop and Nasrabadi, 2007) for a chunk in the signal. For example, if we consider again our Kinect stream with a *framesize* of one second (meaning we process our stream in chunks of one second), we would reduce the 10500 values (see Section 4.1.1.2) to a single one. The value could for example represent the average movement energy in this particular second. Typically, one does not calculate a single feature, but multiple ones that describe various discriminated characteristics of the signal. As an example, in Section 2.2.1.3 we introduced formulas to calculate the expressiveness of movements. The output of such calculations leave us with single features. Besides a value for the energy, we could also retrieve the overall activation, fluidity, the spatial extent of hand movements and others. The single feature outputs are represented by numeric values and are combined in a so called *feature vector*. Depending on the given input signal and task at hand, different features might be more or less appropriate. Finding relevant features is an essential step for the success of any machine learning algorithm . Often a feature can be directly interpreted while the single sample values would not be meaningful themselves. At the same time the data size is drastically reduced.

Typically, the best feature combinations for a specific problem are brute-forced, e.g. by using floating search (Pudil, Novovicová, and Kittler, 1994)

A recent development is that the process of designing features becomes (theoretically) obsolete (see e.g. Trigeorgis et al., 2016). In end-to-end learning approaches, network structures such as *convolutional neural networks* (CNN) are applied. The speciality of *convolutional neural networks* is to find "features" by considering locally connected neurons (that e.g represent pixels of an image). CNNs achieve impressing precision rates for image recognition tasks. E.g. on the classical MNIST recognition problem (recognising hand-written letters) a CNN achieved an error rate of 0.23% in 2016 which is considered the lowest error rate among all tested algorithms. They are mostly used in image and speech recognition tasks. A CNN consists of one or multiple convolutional layers, followed by a pooling layer. A pooling layer is responsible for discarding unnecessary information. Such a construct can be repeated multiple times. If this process is repeated often enough we talk of deep convolutional networks.

End-to-end learning (Trigeorgis et al., 2016) and End-to-feature learning are promising approaches, but require a lot of data and processing power nowadays, which is a problem especially in SSP as there is still a lack of adequate real-world data (as discussed in Chapter 3). For the next sections we will presuppose that we are talking about "hand-engineered" features, but keep in mind that in general they are interchangeable with features that have automatically been learned from raw data.

4.1.3 Machine Learning

Now that we converted our signal to concrete feature vectors, the final step is to identify behaviour patterns based on these features. Machine learning (ML) describes the artificial generation of knowledge, based on expertise. During ML an algorithm learns patterns and rules from given examples by itself, and is then ideally able to generalise these on new samples that is has not seen before.

4.1.3.1 Categories of Machine Learning

To break down the term machine learning, we first consider the categories that machine learning can be divided into:

- Supervised learning: an algorithm learns a function of given input and output pairs provided by a human "teacher". The aim is to learn general patterns and rules based on these examples and often to automatically classify inputs to output classes/values.
- Unsupervised learning: the machine tries to find patterns in the input data, while the targeted output is not known before the learning phase and the machine gets no feedback on success or failure. One general use case is to automatically group samples into clusters.
- Semi-supervised learning: only for a part of the inputs the according outputs are known. The other part of the input is unknown, and algorithms try match the unknown labels based on several assumptions.
- Active learning: an algorithm aims to overcome the bottleneck of labelling data by asking queries of unlabelled instances to be labelled by an oracle (e.g., a human annotator). (Settles, 2010)
- Reinforcement Learning: the algorithm learns via a carrot and stick approach when to change its behaviour to improve its usefulness. That means the direction of the algorithm is changed based on rewards and punishments.

Another way to distinguish ML algorithms is between online and offline learning. During offline learning all data is present, and therefore the learning process may be repeated as often as desired. Online-learning accesses the data only one time to learn from it and then the data is lost. When distinguishing between outputs, main categories are *classification*, that matches input values to discrete classes and *regression*, that matches input values to continuous output scores. This is especially useful when we want to map our prediction not on a set of discrete classes but on a continuous level.

The mentioned categories are not all necessary distinct, and depending on the task some ML techniques are not suitable for being used. For detecting patterns in human social signals we will concentrate on *supervised learning* here. In Chapter 6 we will focus on semi-supervised and active learning for the task of annotating huge databases.

4.1.3.2 Training

For our machine learning algorithm to learn patterns, we first need to provide it with training data. In the previous step we split our stream into chunks with a fixed frame-size and calculated features on each chunk. Additionally, we need to add an *annotation* that contains the label of the output class (or the score in regression problems) for each sample. The annotation process itself will be elaborated in more detail in Chapter 5. A *classifier* now tries to find a mapping function from the feature vector to the output class. Depending on the classifier, the required amount of training data varies. To be able to generalise a model the input training samples should be as balanced as possible in quantity (based on their output classes / scores). Once the training of a model is completed we can evaluate it by applying it on other annotated datasets, by splitting our data into a training and evaluation set, or by applying a cross-validation approach where the model is trained multiple times on different parts of the data and evaluated on the respective other parts.

4.1.3.3 Classifiers

To map input samples to a single class or score, we just generally introduced classifiers. In this section we have a look at two of the most commonly used classifiers, especially in the context of identifying social signals and social cues.

- **Support vector machines (SVMs)** are to this date, the latest supervised learning technique (Cortes and Vapnik, 1995). They aim to construct a so called hyper-plane which separates classes by their maximum distance. Maximising the margin reduces an upper bound on the expected generalisation error. This also means the more distinct features are, the easier it gets to find a maximum margin.

  Figure 12 exemplifies this process for a classification problem. Labelled data samples are fed in a SVM during the training process. In this case samples for "hand moving" and "hand still" exist. They are represented by a hyperspace, meaning for each dimension of our input feature vector, a sample is placed on the

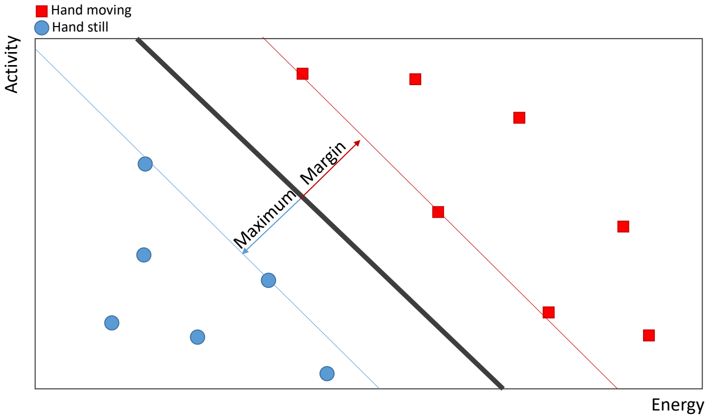


Figure 12.: The hyperplane in the 2D space. The red squares represent samples of the class *"hand moving"*, the blue circles are samples for *"hand still"*. According to their feature-vector (activity and energy) they are placed in the hyperspace. The SVM tries to find the hyperplane with the maximum margin between both classes. In a most scenarios, one does often not have 2 but multiple dimensions.

according axis of the feature value. For illustration purposes, in this example we have a feature vector holding two features, *energy* and *activity*. The SVM calculates the maximum margin between these groups of samples. Once trained, new samples with similar attributes are placed in the same side of the hyperplane, indicating that this samples belongs to a specific class. The probability of the classification depends on the distance of the new sample to the hyperplane and is transformed with a parametric method of logistic regression.

SVMs can either be linear or use a radial (or Gaussian) kernel. A linear kernel usually is much fast for the prediction while a non-linear kernel often is better concerning the predictive performance. Yet, when the number of features is large, data does not need to be mapped to a higher dimensional space, so a linear kernel performs not significantly worse than an non-linear kernel (Keerthi and Lin, 2003).

- **Artificial neural networks**

  An artificial neural network (ANN) is a computational model that is inspired by biological neural networks in the human brain.The concept of ANNs is not exactly new, as it got popular in the mid 1980s (Le Cun et al., 1988; Rumelhart, Hinton, and Williams, 1985) when the back-propagation method, to simultaneously construct the coefficients of the neurons, was introduced. After the appearance of support vector machines in the 1990s, ANNS almost completely disappeared. This was mostly due to the fact that for most tasks, labelled databases were way to small, and computational processing power was too low. In the late 2000s and early 2010s ANNs experienced their renaissance due to the boost in advances in computer hardware developments. E.g. graphic card producer NVIDIA introduced the CUDA framework which allows the GPU to perform general purpose processing. This enables strong parallelisation in computation and extreme speed boosts.

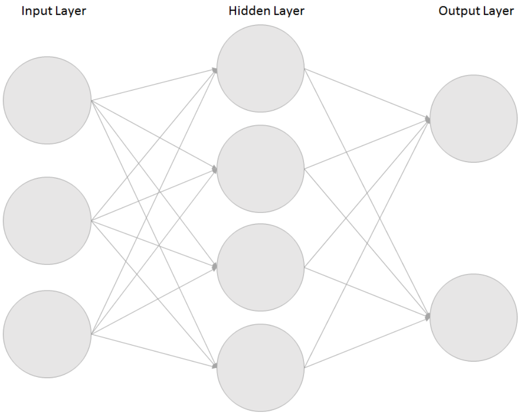


Figure 13.: A neural network with an input layer (each neuron represents one feature of the feature-vector), one (or many) hidden layers, and an output layer where each neuron represents the target classes

The biggest breakthrough was a paper by Hinton, Osindero, and Teh (2006) introducing what today is known as deep learning. In deep learning multiple (> 3) hidden layers exist. Each layer of neurons is trained and reconstructed with the previous layer to generate more abstract features, finally leading to the final output layer.

The hidden layers are basically activation functions that map values from the previous layer to a new value (e.g. sigmoid, rectified linear unit (relus)) . To train an artificial neural network a learning algorithm called *back-propagation* is applied. The goal is to learn weights on the edges between neurons that influence the calculations in the next layers. Back-propagation consists of the following six steps (Kotsiantis, 2007):

1. Present a training sample to the ANN. For example if we hand over a feature-vector, each input neuron represents one feature.
2. For the sample compare the desired output (the annotation) with the network's output and calculate the error in each output neuron.
3. For each neuron in the network, calculate the desired output, and a scaling factor, how much lower or higher the out-

put must be adjusted to match this output. This is called the local error.

4. Next, adjust the weights of each neuron to make the local error as small as possible.
5. For the local error assign "blame" to the neurons at the previous layer, giving greater responsibility to neurons connected by stronger weights.
6. Repeat the steps above on the neurons at the previous layer, using each one's "blame" as its error.

These steps are repeated for each training sample and weights between neurons get adjusted.

A traditional neural network has no capabilities to consider previous observations in the learning and prediction phases. This seems to be a major shortcoming, as for example, humans don't throw away previous knowledge when they interpret observations. As an example, for interpreting a word in a sentence we always consider previous words as well. This issue is addressed by *recurrent neural networks* (RNNs). They contain loops within their structure, allowing them to keep information. This is especially useful if we want to consider previous temporal context for the next prediction. Imagine we have a language model and want to predict a word based on a sentence like: "the colour of a plant is ..", the obvious prediction will be "green". Considering the last words is enough context to make such an prediction. In other cases, it might not be sufficient to consider the immediate previous context, but rather we want to consider information from a sentence or even a paragraph ago. With a growing gap between information, RNNs become unable to connect such information. To overcome these limitations, Hochreiter and Schmidhuber (1997) introduced *"long short-term memory"* (LSTM) cells. An RNN that contains layers of such LSTM units is often referred to as LSTM network. An LSTM cell consists of three major components: an input gate, an output gate and a forget gate. Using these gates, a LSTM cell has the capabilities to regulate the extent of information it captures, keeps/-forgets or forwards to the next module. This allows them to keep short-term memories over a long period of time.

Various other classifiers exist, but for the given task of recognising particularly social cues we limit ourselves to the proposed ones here. Which of these classifiers is to be chosen depends on the given task, the data and the available resources. While ANNs are currently considered state-of-the art for many tasks, SVMs produce comparable or sometimes even superior results for many tasks, while requiring less training data and time. The main conclusion here is that, once we created features from our raw data stream we can use these classifiers interchangeably, even tough for specific problems one might work better than the others.

4.2 The Social Signal Interpretation Framework

This section is based on the publication: Wagner, Lingenfelser, Baur, Damian, Kistler, and André, 2013

Now that we conceptually described the process of how social signals are recognised by a machine, the next question is, how can we effectively achieve this in a practical approach? One of the most advanced open-source frameworks is the *Social Signal Interpretation* framework (SSI). SSI was originally started by Dr. Johannes Wagner, and has been extended in previous years with new sensor and processing plugins, machine learning frameworks and mobile interfaces.

SSI offers tools for the development of real-time social signal recognition systems based on multiple synchronised sensors. It tackles the challenges introduced at the beginning of this chapter in multiple ways:

SSI is available for download at `https://github.com/hcmlab/ssi`

- The architecture of SSI allows to manage various signals coherently, independently of the signal being a hand motion, a biological measurement signal, a video image or a waveform.
- It offers necessary real-time processing capabilities such as synchronisation of streams, threading and buffering.
- SSI includes the full machine learning pipeline. All tasks, such as sensor capturing, signal processing, low- and high level feature extraction and classification might be performed with the framework in-real time. Also, training models based on features and annotations is supported.
- Multiple fusion strategies are available to combine information from multiple modalities and at various levels.
- SSI supports many different types of sensors, filter and feature algorithms to handle signals.

SSI addresses both developers by providing a C++ API, for implementing own components, but also for deployment, recognition pipelines may be defined in an XML language. SSI also offers a Python interface, allowing to integrate existing algorithms and processing frameworks.

4.2.1 Existing Tools

SSI combines tools for social signal processing and machine learning, with the both the recording and real-time processing of social signals in mind. Other existing frameworks focus mostly on either machine learning tools or are specialised on specific social signals.

Weka[1] is a popular framework that includes various machine learning algorithms that focus on data mining tasks.

Matlab[2] is a commercial tool that supports a scripting language and provides toolboxes for applying multiple algorithms to process signals.

Other tools that allow processing specific social signals are most often designed for a specific modality, like Praat[3] for audio processing.

Tools and frameworks that allow the processing of multi-modal live sensor input are only rare as of today. Some commercial frameworks exists that offer pre-trained models to process emotions from audio-visual data, e.g. affectiva[4] and from biological sensor devices, such as iMotions [5]. Some frameworks for multi-modal live sensor processing in an academic context are Pure Data[6], EyesWeb (Camurri et al., 2007), OpenInterface (Serrano et al., 2008) and the openEAR toolkit which is build on openSMILE (Eyben, Wöllmer, and Schuller, 2010). In such frameworks, developers may use sets of sensor, processing and output components of variable size.

However, such frameworks do not focus on enabling users to build their own machine learning models, including the task of collecting multi-modal corpora, training models and apply these in a real-time scenario.

4.2.2 Architecture

SSI is written in C/C++ and optimised for multi-core processing. It has originally been developed on Microsoft Windows, but now also versions for Linux, Mac OS X and Android are available (Flutura et al., 2016; Damian, Dietz, et al., 2016). Since the integration of sensors largely depends on the availability of drivers and APIs, the availability of sensors and processing algorithms varies depending on the platform. Many external libraries have been included directly to the framework to avoid additional dependencies (e.g. OpenCV, libcurl)

4.2.2.1 Data Structures

In SSI two basic data structures exist, namely the earlier introduced *streams* and *events*, that represent the beginning or end of an activity.

As described earlier, signals captured by a sensor are converted into a continuous stream that contains single samples. A stream further holds its sample rate, dimension and datatype.

A stream then may be manipulated using filter and feature algorithms. In such a conversion, the sample-rate, as well as the dimen-

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] http://www.mathworks.com/products/matlab/
[3] http://www.praat.org
[4] https://www.affectiva.com/
[5] https://imotions.com/
[6] http://puredata.info/

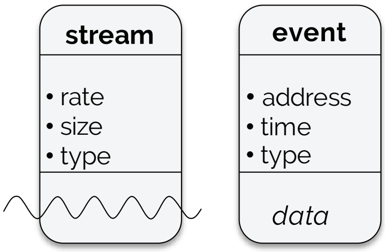


Figure 14.: In SSI signals are organized in streams. Events offer a possibility to hold information about relevant parts in the stream.

sion of a stream may change. The number of retrieved samples is constantly monitored by the framework. In case a mismatch with the expected sample rate is detected, SSI adjusts the stream accordingly by either removing samples or adding samples via interpolation.

Events in SSI, represent relevant parts in streams. For example and event might be the appearance (start and end) of a voice activity in an audio stream. Based on such events feature extraction and classification might be triggered. Events might also be logged and used as automated annotations, if a recognisers works reliably well. The duration of an event is variable and often contains additional data, e. g. the calculations of a set of features.

4.2.2.2 Recognition

SSI is organised to work with so called *pipelines*. A pipeline consists of several autonomic components that are connected, similar to assembly line. In the beginning of a pipeline, most of the time there is a sensor that captures the real-word signals and outputs them as streams. Then, for example, a transformer may filter the raw streams or to compute compact sets of features (as described earlier). On these filtered streams, we may apply activity recognisers to find the positions in the signal that are relevant for the recognition. Once a such a part is found, an event is created and sent to a shared event-board. Components can subscribe to the event-board and are informed, once an event arrives at the board. Based on events streams can be segmented into chunks which can be fed into a classifier during runtime. SSI supports dynamic and statistical classification for streams with variable length, and respectively on streams that are described by a statistical feature set.

Finally, the classifier generates a new event containing either the recognised probabilities for each class or the current value of a regression prediction, and sends these via the event board to other components, e. g. fusion algorithms. Figure 15 illustrates the pipelining process for the detection of single social cues.

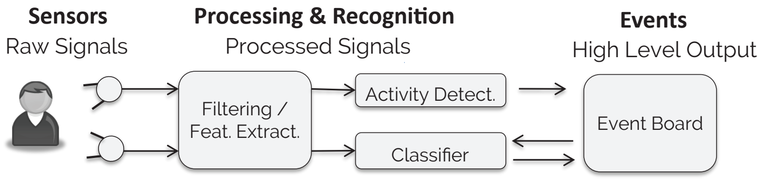


Figure 15.: A recognition pipeline.

The raw- and processed streams, as well as events, can be synchronously stored. Due to SSI's online synchronisation mechanism, all stored data can synchronously be replayed and processed. This opens new possibilities for creating large multi-modal synchronised databases. While synchronised audio and video databases are common, databases including other streams like depth information or physiological sensors are rare (as discussed in the last chapter).

4.2.2.3 Fusion

Figure 16 illustrates three different approaches to fuse information in the SSI framework. The feature vectors extracted by multiple feature extraction components may be plugged together in a common feature vector before a classification model is applied (feature fusion). Similarly, the output of multiple classifiers may be combined (decision fusion, late fusion), e.g. as input for a third classifier. Fusion may also happen at event level. This way, information can be combined at different time scales. We will use event-based fusion in Chapter 7 when information from several components is combined to create a probabilistic model of complex behaviours.

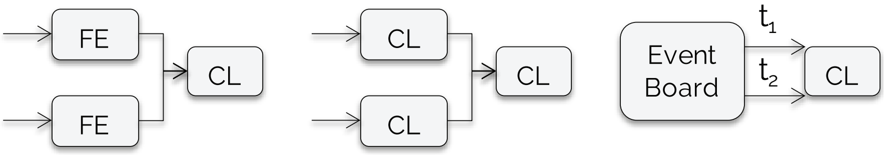


Figure 16.: Different approaches for fusing information in SSI. (FE: feature extraction, CL: classifier)

4.3 Conclusions

In this chapter we investigated how physical social signals can be converted into digital streams and how they are further processed until they are described in concrete features, holding the relevant information of each stream. We further introduced state-of-the-art machine learning techniques that allow the automated recognition of social cues.

To actually perform these steps in the field, we briefly introduced the *Social Signal Interpretation* framework which provides tools to perform these tasks in real-time and thread all modalities in a coherent way. SSI enables researchers to process the full pipeline, from synchronously recording multi-modal raw data, to extract relevant features, training a classifier and predicting output labels for classes of behaviours. When processing data in real-time and in-the-wild scenarios, conditions for recognition models might change, artefacts and unseen sample might appear that a social signal recognition component needs to handle. Also in (close-to) real-time processing, it is not promising to rely on models that are trained on large data segments, as predictions need to be performed as quick as possible to keep systems reactive and interactions natural. The SSI framework specifically aims to handle the parallel real-time processing, feature extracting and prediction of social cues. It's synchronisation capabilities, enable us to create novel, realistic and multi-modal databases, containing various non-acted social signals.

While the machine learning methods briefly introduced in this chapter may be seen as "standard procedures" as of today, in Chapter 6 we address the aspect of putting the human in the loop for training statistical models. To enable non-expert users to make use of such techniques we first need a tool that applies the suggested algorithms in practice. Therefore we introduce our implementation of an annotation tool that directly incorporates cooperative machine learning strategies in the next chapters.

COOPERATIVE MACHINE LEARNING FOR SPEEDING UP THE ANNOTATION OF CONTINUOUS MULTI-MODAL DATA

5 COLLABORATIVE ANNOTATION OF MULTI-MODAL DATA

"The goal is to turn data into information, and information into insight."

— Carly Fiorina

In previous chapters we introduced techniques to synchronously record and process multi-modal human behaviour data. Research areas like behavioural psychology, anthropology, medicine and others are concerned with analysing such data of interactions by identifying relevant information within a corpus. Analogously, if we want to use machine learning to train models for detecting certain behaviours in an automated manner with the help of e.g. artificial neural networks or support vector machines, we first have to provide them with labelled training data. During training, supervised learning techniques rely on pre-annotated samples. Once given enough examples, they are eventually enabled to match new unlabelled samples to a class or value. Also semi- and unsupervised learning algorithms at least start with a set of labelled data. The *annotation* process is also called *coding*, *transcription* or simply *labelling*. A person performing such annotations is called *annotator* or *rater*.

This chapter is based on the publications: Baur, Mehlmann, Damian, Lingenfelser, Wagner, Lugrin, André, and Gebhard, 2015 Wagner, Baur, Zhang, Valstar, Schuller, and André, 2018

The chapter is structured as follows:

- In Section 5.1 we discuss current challenges in the process of annotating continuous large multi-modal databases.
- In Section 5.2 we give an overview on related work in the area of annotation tools that focus on social signals.
- In Section 5.3 we introduce the NOVA tool and briefly discuss its core features.

5.1 Challenges in the Annotation Process

Annotation of multi-modal natural data comes along with a couple of challenges that concern all research disciplines alike and are addressed by the NOVA tool presented later in this chapter.

- **Handling various types of annotations:** Depending on the task, different types of annotations might be more or less adequate. Some problems can be matched to discrete entities or classes, making discrete segments containing information about the current state a perfect solution. The vocabulary of such segments can be a fixed set, for example when multiple raters should use the same wording and granularity of annotations. Such a fixed set is called an *annotation scheme* which helps reducing ambiguities between raters. In other tasks a limited set is not useful, e.g. when speech is to be transcribed. Given an almost endless set of words in a language, annotators should have the free choice to place in what they actually believe they heard. For tasks that do not fit as clear on discrete segments, such as emotions, social attitudes and similar phenomena, a dimensional scheme might be more appropriate (Metallinou and Narayanan, 2013). This approach for example maps directly on dimensional emotion models, as discussed in Section 2.4.2.2. On a predefined scale, raters decide on their impression of e.g. the current emotional state. When working on detecting social cues from video, e.g. facial expressions, a whole different kind of annotation is needed. Geometric positions are marked on a video for the classifier to learn geometric attributes in an image (or video). Various, free and commercial tools exists to address single types of annotation tasks. In this chapter we will introduce our tool that combines all of these annotation types to incorporate multi-modal phenomena.

impression in this case, may also be influenced by guidelines and agreements

- **Finding inter-rater agreement:** In many tasks, annotations are done for predefined segments of data with a fixed duration (McKeown and Sneddon, 2014), e.g. a short audio snippet, or an image. In natural, continuous human interactions, the task is often way more complex as an additional aspect is for the rater to decide when a segment starts (onset) and when it stops (offset). For example, a hand movement is a dynamic signal, and raters might naturally have different opinions on when a gesture starts and stops. To find a common *ground-truth*, annotations from two or (preferably) more persons should be combined, or at least measured against each other. This is partly done by defining a common annotation scheme, and create guidelines and agreements that all raters should be aware of. When multiple raters should decide on the emotional state of a person, opinions might nevertheless lie far from each other. Statistical approaches aim to find a consensus by comparing the correlation between the annotations and should be applied when building the ground-truth annotation.

- **Collaboration and management of large multi-modal corpora:** The management and exchange of annotations is another challenging task. The exchange of annotations with local files and the coordination of labelling tasks on large databases often is a difficult project. We suggest a centralised database where raters share and discuss their decisions in an easy-to-access way. Collaboration does not have to be limited between human raters. In Chapter 6 we introduce our approach to enhance the annotation process with the support of a machine annotator.

- **Saving time and effort:** Annotation of behaviours in multiple modalities on large databases is a time-consuming task. To reduce the required resources we suggest a combination of three aspects. (a) First, the annotation interface should be easy and intuitive to use, including short-cuts where possible. (b) creating semi-automated segments based on existing social signal processing techniques, as well as other input that can be directly derived from a machine. Existing recognisers already identify a huge amount of behaviours that can be logged during runtime, and be used similar to manual labels. The rater then only needs to approve or correct these segments. Agent states and other context information can be logged automatically in real-time systems as well. (c) As mentioned before, sharing the annotation task with the machine drastically reduces the time that needs to be invested.

To address these challenges we implemented the NOVA ((**No**n)**v**erbal **A**nnotator) tool. Before presenting NOVA in more detail, we first give an overview on the most commonly used annotation tools.

5.2 Related Work

NOVA's general interface has been inspired by existing annotation tools. For instance, EUDICO linguistic annotator (ELAN) (Wittenburg et al., 2006), annotation of video and language (ANVIL) (Kipp, 2013), and EXMARALDA (extensible markup language for discourse annotation) (Schmidt, 2004). These tools offer layer-based tiers to insert time-anchored labelled segments, that is *discrete* annotations. *Continuous* annotations, on the other hand allow an observer to track the content of an audiovisual stimulus over time based on a continuous scale. A tool that allows labellers to trace emotional content in real-time on two dimensions (activation and evaluation) is FEELTRACE (Cowie, Douglas-Cowie, et al., 2000). Its descendant GTRACE (general trace) (Cowie, McKeown, and Douglas-Cowie, 2012) allows the user to define their own dimensions and scales. Other tools to accomplish continuous descriptions are CARMA (continuous affect rating and media annotation) (Girard, 2014) and DARMA (dual axis rating and media annotation) (Girard and Wright, 2016).

An interesting approach for gathering crowd-sourced annotations is iHEARu-PLAY (Hantke et al., 2015), that allows labelling audio material on various scales in form of a browser-game. Whereas most tools are restricted to describe audiovisual data by a single user, REPOVIZZ (Mayor et al., 2013) is an integrated online system to collaboratively annotate streams of heterogeneous data (audio, video, motion capture, physiological signals, etc;). Datasets are stored in an online database, allowing users to interact with the data remotely through a web browser. One approach for incorporating active learning from an annotation tool is ATLAS (Meudt, Bigalke, and Schwenker, 2012). It allows to visualize multiple data streams and supports active learning from the user interface, yet is limited to recognition tasks that can be solved with MFCC features and support vector machines only. Further, the precision of their generated predictions has not been evaluated in their publication.

Though the mentioned tools are of great help to create annotations at a high level of detail, they suffer from several drawbacks. Firstly, they have been developed with a strong focus on audiovisual material, other signals like depth information, e.g. skeleton and face tracking, or physiological data streams are not or only sparsely supported. Secondly, almost none of the tools allows different types of annotations. Since different coding types have certain pros and cons the choice depends on the observed phenomenon and should be selectable on demand. Finally, almost all of the tools offer none or only little automation. However, since labelling of several hours of interaction is an extremely time consuming task, methods to automate the coding process are highly desirable. NOVA overcomes the limitation of other tools to only playback audio and video streams, and supports the display of an arbitrary number of video and time-series tracks. Additionally, it has been advanced with features to create collaborative annotations and to apply cooperative machine learning strategies out of the box for multiple recognition problems (see Section 6.2). To support a truly collaborative work-flow between several annotators and the machine NOVA provides a database back-end to store, exchange, and combine annotation work.

5.3 The (Non)verbal Annotator: NOVA

NOVA is open-source and available for download at `https://github.com/hcmlab/nova`

NOVA has been developed following the requirements of tools for annotating human social interactions. It offers annotations on multiple tiers and is able to visualize multi-modal data, such as audio, video, depth-camera skeletons or facial points.

Figure 17 gives an overview on a typical instance of the graphical user interface. As seen, multiple media files can be opened in parallel. In this instance, not only audio and video files are loaded, but also the extracted face of the left user, as well as the skeleton of the right user. Additionally, the waveforms of the audio are visualised below the videos. Several discrete annotation tracks follow, containing information about who has the turn in the conversation and if there was an interruption (for more details, see Section 7.4.5.1). Such automatically created annotations seamlessly integrate in the NOVA tool with manual annotations. Examples are the next two tracks that show speech and fillers for one participant each. Finally, continuous tiers for valance and arousal are opened. NOVA supports a variety of annotation types, as will be elaborated in Section 5.3.2.

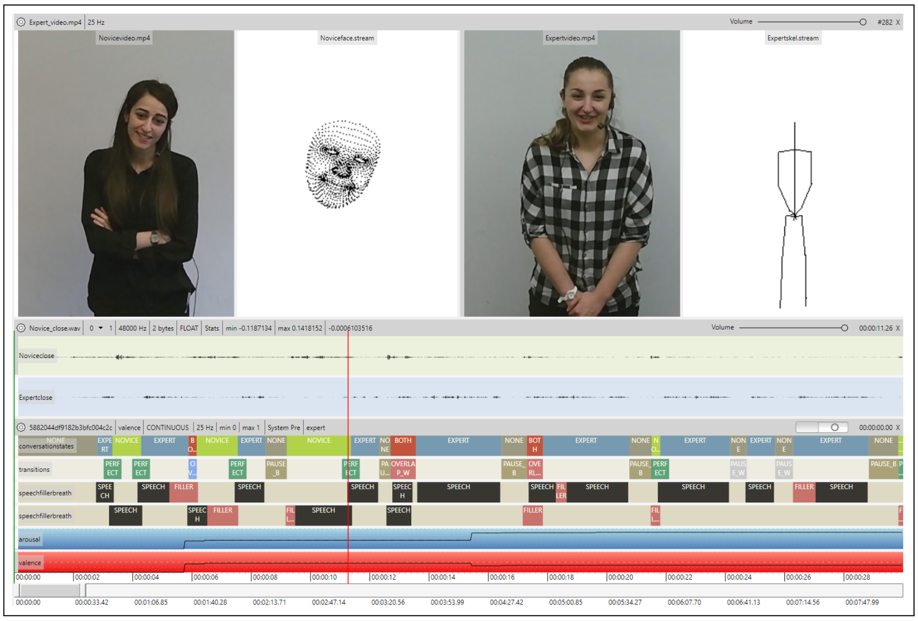


Figure 17.: The interface of NOVA: in the upper part we see videos of two users interacting during a NoXi recording, as well as, results of the face and skeleton tracking. Audio streams are displayed as waveforms in the centre of the figure. Below, several discrete and continuous annotation tracks are shown. Visualised content can be played back in real-time and new annotations can be added on-the-fly.

5.3.1 Media and Streams

The media panel (Figure 17, top part) plays back the recordings of the interaction. All media streams are synchronised during playback, and users can navigate through the video by clicking on any point in the time track. Also navigating forwards and backwards to the next/last frame is possible for precise annotation. NOVA relies for playback on the system's installed codecs, which allows playing almost any media format form the interface. Additionally, data can be visualised, such as the skeleton recorded with the Microsoft Kinect Sensor (or similar depth camera or motion capture sensors.), and face points, recorded with various face trackers (see Figure 17).

5.3.2 Annotation Types

As mentioned earlier in this chapter, the coding process of multimodal data depends on the phenomenon we want to describe. For example, we would prefer a discrete annotation scheme to label behaviour that can be classified into a set of categories (e. g. head nods and head shakes), so that all annotators use the same "vocabulary", whereas variable dimensions like activation and evaluation are better handled on continuous tiers. For tasks like language transcriptions, which consist of hundreds of individual words, we want to assign labels with free text. Finally, we might also want to annotate geometric points in visual material, for example if we want to learn about movements of the face.

To meet the different needs, NOVA supports four kinds of annotations:

1. **Discrete annotations** consist of a list of labelled time segments. Each segment has a start and end point and holds a label name. Segments can vary in length, may overlap and possibly have a gap to adjacent segments. Label names are not arbitrary but chosen from a set of predefined categories (*classes*). For instance, to label head movements in a video stream we choose the class labels "NOD" and "SHAKE" for nods and head shakes, and "OTHER" to code other movements. The label "GARBAGE" is always available and should be used where the annotated stream is corrupted. See Figure 18 for an example of a discrete tier.

2. **Free annotations** are similar to discrete annotations, but allow annotators to assign free label names. This is obviously useful if an annotation task can not easily be reduced to a few classes (for example in case of speech transcriptions). Of course there is the risk that the same phenomenon may be labelled differently (either because a synonym is used or due to misspelling). See Figure 18 for an example of a free tier.

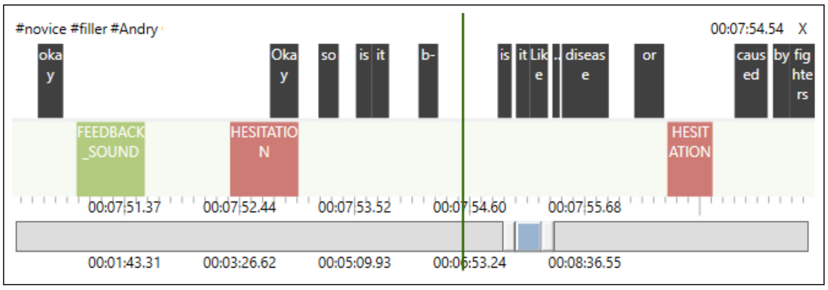


Figure 18.: Example of a discrete (bottom) and free (top) annotation tier. The start- and endpoint of a label can be directly changed with the mouse (even during playback). The name of a label can be changed through a dialogue by using pre-defined 'hot keys'.

3. **Continuous annotations** are continuous in time and space. Instead of names, numerical values (*scores*) are assigned at a constant interval defined by a selectable sample rate. For instance,

a sample rate of 2 Hz means that two scores are assigned per second. Scores have to be within a pre-defined interval.

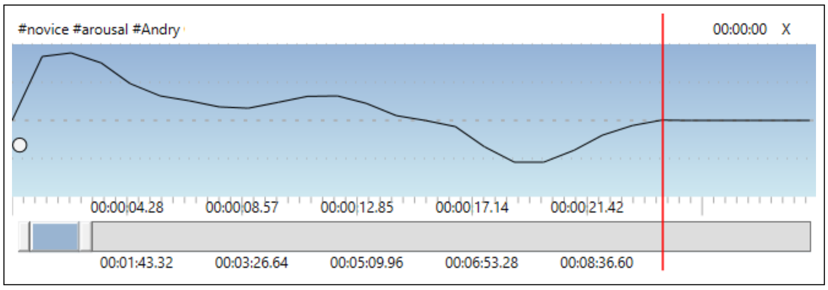
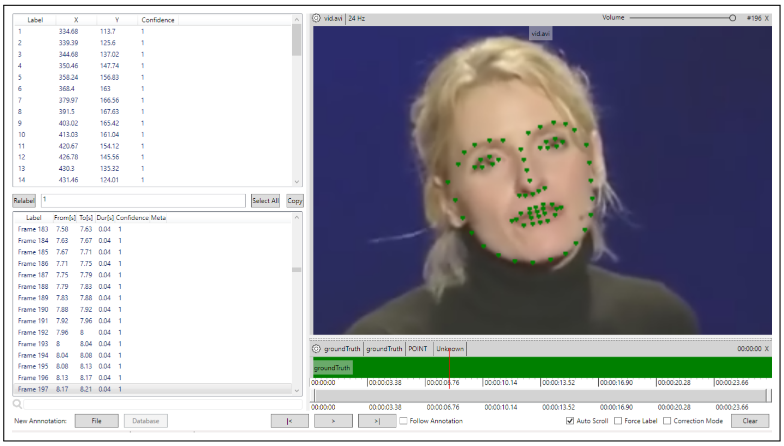


Figure 19.: Example of a continuous annotation tier. A value within a predefined range is assigned at a constant interval. The white dot on the left shows the current score height controlled by the y position of the mouse cursor. In live mode the value is automatically assigned to the current playback position (indicated by the red marker).

A live mode is available that allows annotators to interactively change the score values by moving the mouse or using the up and down keys to the desired level. See Figure 19 for an example of a continuous tier. This is especially useful for regression tasks in machine learning, or for describing emotions and attitudes.

4. **Geometric annotations** are meant for annotation tasks where neither discrete nor continuous annotations are useful. Imagine we want to train a model to recognise facial landmarks, for example to calculate the FACS (see Section 2.2.2) automatically. For this and similar use-cases NOVA also supports geometric annotations, as seen in Figure 20.



Label	X	Y	Confidence
1	334.68	113.7	1
2	339.39	125.6	1
3	344.68	137.02	1
4	350.46	147.74	1
5	358.24	156.83	1
6	368.4	163	1
7	379.97	166.56	1
8	391.5	167.63	1
9	403.02	165.42	1
10	413.03	161.04	1
11	420.67	154.12	1
12	426.78	145.56	1
13	430.3	135.32	1
14	431.46	124.01	1

Relabel 1 Select All Copy

Label	From[s]	To[s]	Dur[s]	Confidence	Meta
Frame 183	7.58	7.63	0.04	1	
Frame 184	7.63	7.67	0.04	1	
Frame 185	7.67	7.71	0.04	1	
Frame 186	7.71	7.75	0.04	1	
Frame 187	7.75	7.79	0.04	1	
Frame 188	7.79	7.83	0.04	1	
Frame 189	7.83	7.88	0.04	1	
Frame 190	7.88	7.92	0.04	1	
Frame 191	7.92	7.96	0.04	1	
Frame 192	7.96	8	0.04	1	
Frame 193	8	8.04	0.04	1	
Frame 194	8.04	8.08	0.04	1	
Frame 195	8.08	8.13	0.04	1	
Frame 196	8.13	8.17	0.04	1	
Frame 197	8.17	8.21	0.04	1	

Figure 20.: Example of a geometric annotation tier. For each frame an annotator can move multiple pre-defined points on the video. To speed up the process each frame can be copied to the next one, so only adjustments need to be done.

5.3.3 Annotation Schemas

Each annotation type comes along with its own annotation scheme. For example, for discrete annotations a scheme contains information such as the annotation's name, the background colour of the tier and the labels allowed on the tier, respectively their colours. Once such a scheme is loaded, the annotator can only chose between these predefined labels. As described before, for FREE annotations, labels are not predefined and can be chosen freely during the coding process. Continuous and geometric schemes contain information such as the sample-rate (see Section 4.1.1.1), the minimum and maximum ranges, and for geometric annotations the number of points per frame. Using annotation schemes allows multiple annotators to create comparable annotations, and helps avoiding errors and misunderstandings.

5.3.4 Database Backend

The coding of a corpus can be a lengthy process involving several annotators from different sites. Widely used annotation tools are limited to store annotations to files on a local disk drive. For a better support of a collaborative annotation process, we have implemented a database back-end, which allows users to load and save annotations from and to a MongoDB running on a central server. This gives involved annotators the possibility to immediately commit changes and follow the annotation progress of the others. MongoDB[1] is an open-source and cross-platform NoSQL database. We have chosen it in favour of a relational database due to its simplicity and fast read-/write operations.

We opt for a design that not only allows to read and write annotations, but manages all relevant meta data of a corpus, too. Generally, each corpus is represented by a single database including several collections (the analogous to tables in relational databases). The collections are (see also Figure 21):

- **Meta**: meta information about a database, including the data server location, and a description
- **Sessions**: stores general information for each recording session, such as location, language and date.
- **Annotators**: stores names and meta information of the involved annotators (human or machine!).
- **Roles**: stores the different roles subjects can take on during a recording session (e. g. listener vs. speaker).

[1] https://www.mongodb.com/

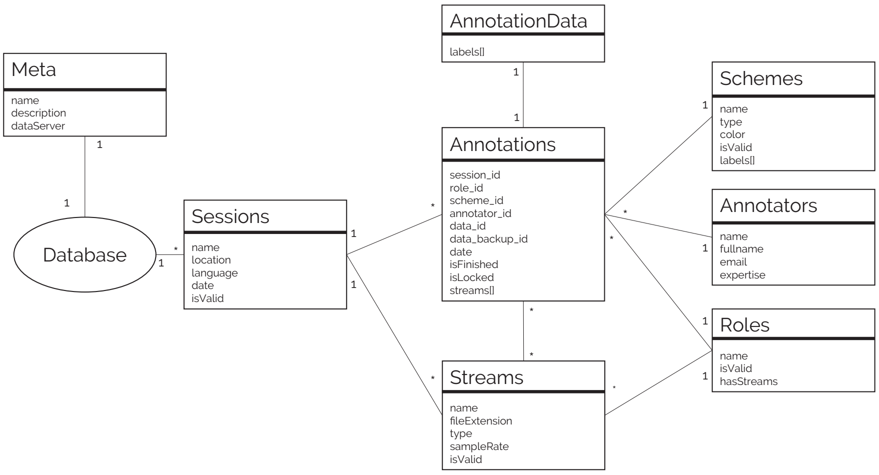
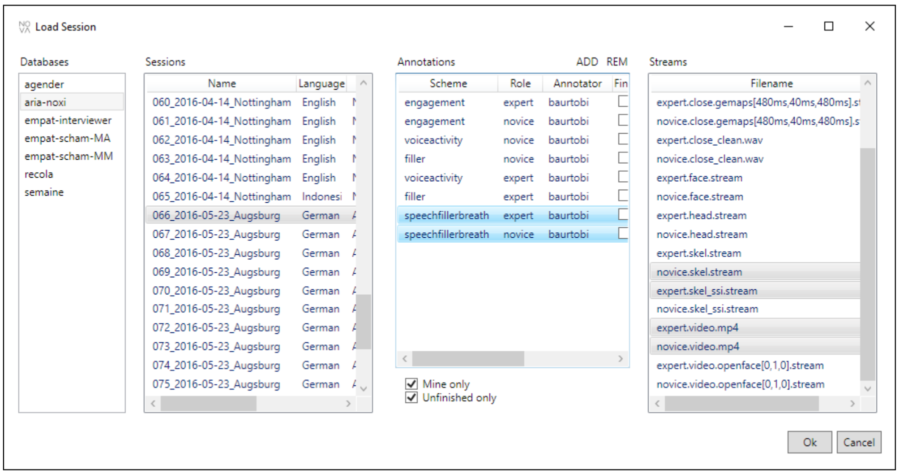


Figure 21.: Overview of NOVA's database structure. Annotations and meta information on subjects, sessions, etc. are stored in different collections. NOVA includes necessary tools to maintain and populate a database.

- **Streams**: stores the recorded stream files. Each file is assigned to a media type, a session, a subject and a role. An url is included that points to the location where the file can be downloaded.
- **Schemes**: stores the available annotation schemes.
- **Annotations**: stores the headers of created annotations. An annotation is linked to an annotator, an annotation scheme, a role and a session. Optionally, a list of stream files is referenced to store which information should be displayed during the annotation process.
- **AnnotationData**: contains the actual annotation data (segments or scores) for an annotation. Additionally a backup is stored for each annotation, allowing the user to go back to the previous version.

As soon as several users collaborate on a common database it becomes crucial to implement adequate security policies. For instance, we want to prevent a situation in which a user accidentally overwrites the annotation of another user. Therefore, standard users can only edit and delete their own annotations. They can, however, load annotations of other users. In that case the annotation is copied and stored under their username. Only users, privileged with admin rights may edit and delete annotations of other users. They can also assign newly created annotations to specific users. This way, an admin can divide up forthcoming annotation tasks among the pool of annotators.

Beside human annotators, a database may also be visited by one or more "machine users". Just like a human operator they can create and access annotations. Hence, the database also functions as a



Load Session

Databases

- agender
- aria-noxi
- empat-interviewer
- empat-scham-MA
- empat-scham-MM
- recola
- semaine

Sessions

Name	Language
060_2016-04-14_Nottingham	English
061_2016-04-14_Nottingham	English
062_2016-04-14_Nottingham	English
063_2016-04-14_Nottingham	English
064_2016-04-14_Nottingham	English
065_2016-04-14_Nottingham	Indonesi
066_2016-05-23_Augsburg	German
067_2016-05-23_Augsburg	German
068_2016-05-23_Augsburg	German
069_2016-05-23_Augsburg	German
070_2016-05-23_Augsburg	German
071_2016-05-23_Augsburg	German
072_2016-05-23_Augsburg	German
073_2016-05-23_Augsburg	German
074_2016-05-23_Augsburg	German
075_2016-05-23_Augsburg	German

Annotations ADD REM

Scheme	Role	Annotator	Fin
engagement	expert	baurtobi	□
engagement	novice	baurtobi	□
voiceactivity	novice	baurtobi	□
filler	novice	baurtobi	□
voiceactivity	expert	baurtobi	□
filler	expert	baurtobi	□
speechfillerbreath	expert	baurtobi	□
speechfillerbreath	novice	baurtobi	□

☑ Mine only
☑ Unfinished only

Streams

Filename
expert.close.gemaps[480ms,40ms,480ms].st
novice.close.gemaps[480ms,40ms,480ms].s
expert.close_clean.wav
novice.close_clean.wav
expert.face.stream
novice.face.stream
expert.head.stream
novice.head.stream
expert.skel.stream
novice.skel.stream
expert.skel_ssi.stream
novice.skel_ssi.stream
expert.video.mp4
novice.video.mp4
expert.video.openface[0,1,0].stream
novice.video.openface[0,1,0].stream

Figure 22.: NOVA's database interface: when an annotator chooses a session, available annotations and stream files are displayed. For each annotation the scheme, role and annotator are shown, as well as, two flags that allow it to mark sessions as being finished or lock them to prevent that they are accidentally overwritten. If necessary, additional credentials can be entered to access the server where the stream files are stored.

mediator between human and machine (more on that in Section 6.2). To control the annotation progress we have introduced a 'isFinished' flag that signals if an annotation requires further fitting or is finished. A second flag 'isLocked' marks whether an annotation is editable or not.

NOVA provides instruments to create and populate a database on a MongoDB server from scratch. At any time new annotators, schemes and additional sessions can be added, without specific knowledge about database structure. Figure 21 illustrates the structure of the database. A screenshot of the dialogue to access annotations from the database is depicted in Figure 22.

5.3.5 Statistics and Inter-Rater Agreement

"If we have data, let's look at data. If all we have are opinions, let's go with mine."

— Jim Barksdale, former Netscape CEO.

NOVA not only allows the storage and management of annotations in its database back-end, but also to create statistics and to merge annotations from multiple annotators. Statistics are created for both annotations and streams, which may help in the direct interpretation of the observes behaviours, but also on the annotations between raters.

5.3.5.1 Statistics on streams

Statistics for streams are visualised for example by pie chart diagrams that represent proportions between single classes, as seen in Figure 23. These charts refer, depending on the user's specification, either to a chosen timespan, or alternatively to the complete recording. This for example helps identifying visually parts where user's showed relative high or low amplitude in signals.

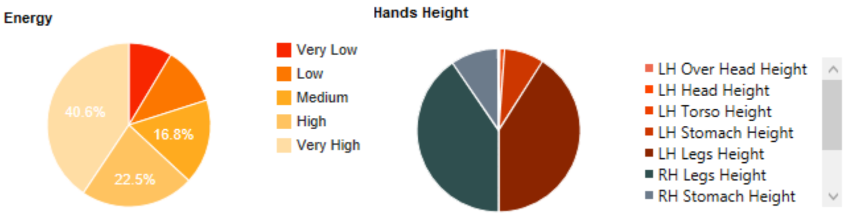


Figure 23.: Examples for pie chart diagrams for energy and the height of the user's hands that visualise user behaviour for a selected section.

5.3.5.2 Statistics on Annotations

For annotations, widely-used measurements to find inter-rater agreement have been implemented directly in the user interface. Especially when we want to merge annotations from multiple raters, identifying those annotations with a high agreement is essentially important to gain information about the reliability of our ground truth. For example, if multiple raters achieve a high correlation value, our merged annotation will be reliable, whereas a low correlation value means that the annotations only correlate by chance. For measurements of continuous and discrete annotations different procedures are applied. To identify the agreement between various raters on a discrete set of classes the κ (kappa) correlation is a common measurement of inter-rater agreement. In NOVA we integrated Cohen's κ that allows to find correlations between two raters, and its extension Fleiss's κ which is suitable for multiple raters.

- Cohen's κ (kappa) (Cohen, 1960)

  $$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

  p_0 represents the relative observed agreement among raters, and p_c represents the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category.

- Fleiss's κ (Fleiss and Cohen, 1973)

  $$\kappa = \frac{\bar{P} - \bar{P_e}}{1 - \bar{P_e}}$$

An extension to Cohens κ is the so called Fleiss κ. The factor $1 - \bar{P_e}$ represents the degree of agreement that is attainable above chance, and, $\bar{P} - \bar{P_e}$ represents the degree of agreement that was actually achieved above chance.

In both measurements, a κ value = 1. represents a perfect agreement while if there is no agreement among the raters then $\kappa \leqslant 0$. (Landis and Koch, 1977) The values in-between represent slight - to almost perfect agreement. This scheme for interpretation, however, is not universally accepted as the κ value also depends on the number of raters and classes. One should keep in mind that having fewer categories will result in higher values.

For comparing continuous annotations, other types of measurements are required as we now have to deal with a regression task. NOVA features the calculation of the root mean square error (RMSE), as well as the Pearson correlation coefficient (r) and Cronbach's α to compare two or more continuous annotations.

- Root Mean Square Error (RMSE) (Willmott, 1981)

  $$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

  To calculate the RMSE each value in a continuous annotation is compared with the value on the same x position in another annotation. This step can for example be performed after an annotation merge (see Section 5.3.5.3) was performed, to see which annotations fit or outlines the averaged curve - in other words which annotators drift away from the averaged groundtruth.

  To calculate the RMSE first we need to determine the difference between the actual value and the predicted values (in our case, the annotation of interest and the averaged annotations).

  In the formula above they are represented by $\hat{y}_i - y_i$, where y_i is the observed value for the ith observation and $\hat{y}_i$ is the predicted value.

  The results can be either positive or negative, as the predicted value under or over estimates the actual value. By squaring the results, averaging the squares and finally taking the square root, we receive the RMSE as a measure of the spread of the y values about the predicted y value.

- Pearson's correlation coefficient (Pearson, 1895)

  $$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$

Pearson's correlation coefficient (PCC) r is a unit-less index that represents the strength of the association between two variables (+ = positive association, - = negative, 0 = no association). r results in a range of $-1 \leqslant r \leqslant 1$ and measures the linear relationship between two subsets X and Y.

PCC tests for significant association by testing whether the population correlation is zero

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}i$$

which is similar to the t-test used to test whether the population r is zero.

It uses probability calculations for the t distribution to get the p-value (2-tailed if interested in association in either direction), 1-tailed test for a positive correlation between X and Y. In other words, one can say it tests H_0 : when $X \uparrow$ does $Y \uparrow$ in the population and the other way round. A value close to +1 represents a high correlation between the subsets, and therefore in our case a high correlation between both annotations. According to Cohen (1988) an $|r| = .10$ is considered a weak correlation, $|r| = .30$ a medium or moderate correlation and $|r| = .50$ a strong correlation. In other interpretations, e.g. in psychological questionnaires values up to .30 are considered a slight correlation where .50 is considered as moderate and .70 - .80 a very high correlation.

- Cronbach's α (Cronbach, 1951; Santos, 1999) named after Lee Cronbach (who rather speaks of the coefficient alpha) is a commonly used correlation measurement in statistics. The formula goes:

  $$\alpha_{st} = \frac{N \cdot \bar{r}}{1 + (N-1) \cdot \bar{r}}$$

  where N represents the number of components (items or subscales) and $\bar{r}$ represents the average correlation between the items.

  As an alternative, Cronbach's α calculates as

  $$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N} \sigma_{Y_i}^2}{\sigma_X^2}\right) \qquad \text{where} \qquad X = \sum_{i=1}^{N} Y_i$$

  here again N represents the number of components (items or subscales) , σ_X^2 represents the variance of the observed overall scores and $\sigma_{Y_i}^2$ represents the variance in component i.

  A Cronbach α value above 0.9 is considered as an excellent agreement, while a value below 0.5 is considered as not acceptable.

5.3.5.3 Merging Annotations

By calculating the correlation values, we can see how well our annotators agree on a given problem. If we find a common agreement between them (or a subset of the annotators) NOVA further allows to directly merge annotations from the interface to create *gold standard* annotations. Depending on the type of annotation, different approaches are offered.

For discrete annotations, a user may select any number of annotations for the same problem (same session, scheme and role), but from multiple annotators. The annotations are split in small chunks (e.g. 40 ms which represents 25 FPS) and compared across all annotators. For each chunk, a majority vote is performed, mapping each chunk to the label that was given in most cases. By option, annotations from specific raters can be weighted, e.g. by the expertise of a rater. Once this step is performed, the small chunks are put back together to coherent larger chunks, representing merged labels. Figure 24 illustrates this process with an example for a scheme that contains VOICE, FILLER and BREATH labels.

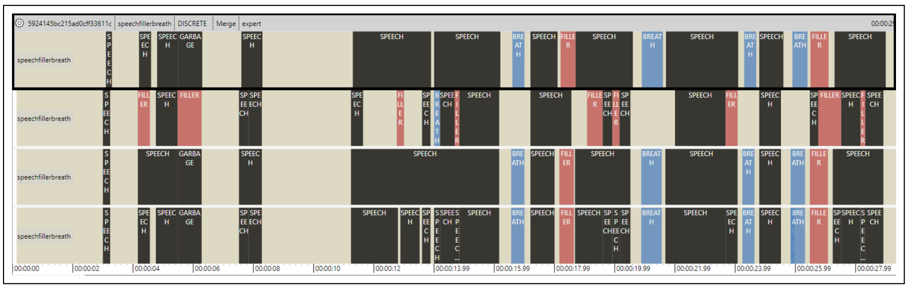


Figure 24.: Discrete Annotation merge in NOVA: the upper tier shows the merged annotations from three raters for a voice/filler/breath scheme

The top tier shows the merged annotations, the three tiers below show the single annotations from three different raters. By using this process on many annotators a common ground truth can be created. Labels are further equipped with a certainty value, in case only a part of raters share the same opinion. For example if 9 out of 10 raters agree, the confidence for the segment will be 0.9.

In the case of continuous annotations, NOVA offers basic signal processing algorithms for the merge process. Multiple continuous annotations can be merged by calculating the mean, for each sample across all annotations. With the inter-rater agreement measurements introduced in the last section, annotations with low agreement might be removed before the merging process, and again, rater's can be weighted based on their reliability or expertise.

5.3.6 Plugin Mechanism

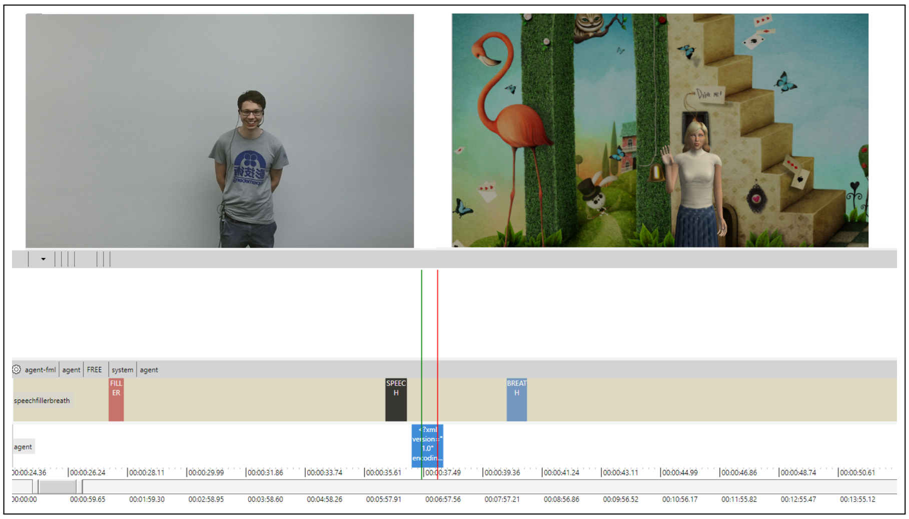


Figure 25.: Using a NOVA plugin, the tool is able to send FML/BML code to the GRETA player engine to replay agent behaviour while replaying the interaction.

NOVA supports a plug-in system, that allows external developers or students to add functionality without changing the core code. Figure 25 shows a plugin that allows replaying agent behaviours in the external GRETA (Pelachaud, 2015) virtual agent engine. During the interaction *functional markup language* (FML) (Schröder, 2010) tags are saved as the virtual agent interacts with the user. Loading these logs, alongside a plugin, NOVA sends FML tags via ActiveMQ or UDP/TCP protocol to the engine, and agent behaviours are replayed alongside the video of the user. This for example enables replaying the interaction with an agent for later analysis.

Another plug-in allows to call SSI XML pipelines directly in the background when an annotation file contains meta tags. The combination of a *OSCSender* and a *pipeline* for example allows sending SSI events directly from the NOVA tool, by sending annotations to an SSI XML pipeline during playback. That way, event based algorithms can be run in a simulated way, without the need to actually record live data from sensors, but rather from pre-recorded events, or even manual annotations that help tuning other components.

5.4 Conclusions

In this section we introduced NOVA, a novel annotation tool that allows a collaborative work-flow on multiple types of annotations and the analysis of human behaviours. NOVA was designed to annotate multiple aspects of human non-verbal behaviours, from discrete events to continuous changes in emotion and social attitude.

Further, NOVA is offers capabilities to manage large multi-modal corpora. This includes the management of corpora and their annotations with a collaborative shared database to allow annotators from multiple sides to cooperate. For handling multiple annotations for the same problem and session, it features functionality to merge annotations to create a common gold standard annotation. Therefore it offers tools to identify inter-rater agreement by calculating correlations between raters. NOVA further has a plugin mechanism that allows third party developers to add new functionality, such as sending annotations via multiple network protocols during playback, or to feed event-based SSI XML pipelines with annotations. Due to its capabilities to visualise automatically generated annotations of behaviours and streams it also may be used as analysis or coaching interface (see Section 7.5.1). This way, NOVA may be used as explanation interface for the task of investigating correlations between recognised social signals. NOVA is open-source and can be found at `http://github.com/hcmlab/nova`.

As a main feature, NOVA supports interfaces to interactive machine learning algorithms that aim to support human annotators by outsourcing parts of the annotation process to a machine. This process will be elaborated in more detail in the next chapter.

6 COOPERATIVE MACHINE LEARNING

A lack of transparency results in distrust and a deep sense of insecurity "

— Dalai Lama

6.1 Motivation

This chapter is based on the publication: Wagner, Baur, Zhang, Valstar, Schuller, and André, 2018

With the NOVA tool introduced in the last section, we presented a solution that focuses on the fast annotation and analysis of social signals in a very efficient and convenient way. Nevertheless, when we think about the many hours of data that is contained in modern corpora (see Section 3.1), the manual annotation process still requires and extensive amount of time and resources - something that no manual annotation tool in the world could ever change. The annotator needs to identify the start and end of a social signal and rate it accordingly. Depending on the granularity of annotation this applies for very short episodes, and most often on multiple modalities. This chapter is concerned with one main question: how can we use machine learning techniques to speed up the annotation process and reduce the amount of time and resources for this process in a transparent and easy-to-access way?

One approach is the *active learning* (AL) (Zhu, 2005) algorithm that interactively query the user to manually label certain data points. The core idea of AL is to extract the most informative instances from a pool of unlabelled data based on a specific query strategy (Settles, 2010) (for more details, see Section 6.2.2). These selected instances are then passed to human annotators and finally – after labelling – a model is derived from this subset. This, of course, reduces the labelling effort. In addition, it has two more positive side effects. First, it speeds up the training since fewer instances have to be processed. Second, it helps improving the maximum accuracy, as it reaches a more coherent learning model (focusing on the most relevant cases). The work by Zhang, Coutinho, Deng, et al. (2015) takes the idea of AL a step forward and combines it with *semi-supervised learning* (SSL) techniques to efficiently share the labelling work between human and machine: a pre-existing classifier is used to derive confidence values for unlabelled data, thus human annotators are involved only for instances predicted with insufficient confidence. Such a strategy allows the performance of an existing classifier to be improved while minimising the costly work of human labelling. To further save labelling efforts one can apply *dynamic active learning* (DAL) by choosing the most reliable raters first (Zhang, Coutinho, Zhang, et al., 2015a). To

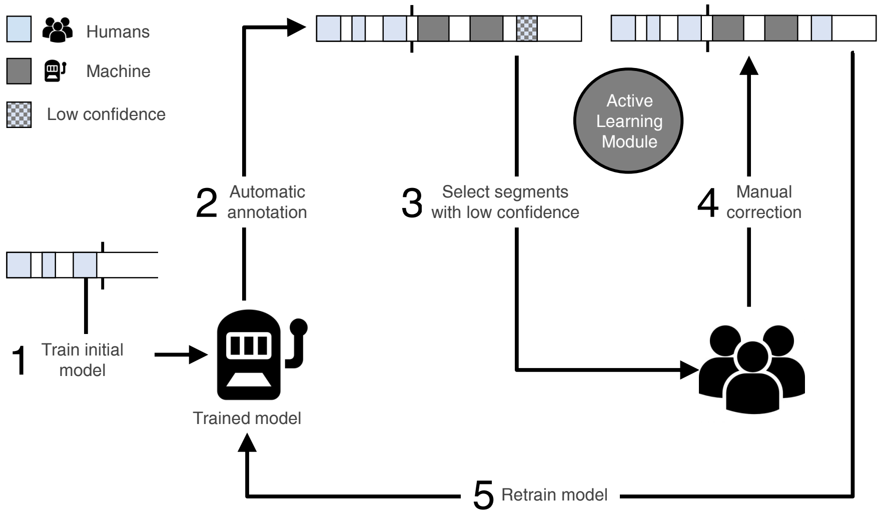


Figure 26.: The scheme depicts the general idea behind cooperative machine learning (CML): (1) An initial model is trained on partially labelled data. (2) The initial model is used to automatically predict unseen data. (3) Labels with a low confidence are selected and (4) manually revised. (5) The initial model is retrained with the predicted / revised data.

put it into other words, the DAL algorithm also considers how many and which annotators should be queried on a per instance level.

Here, we subsume learning approaches that efficiently combine human intelligence with the machine's ability of rapid computation under the term *cooperative machine learning* (CML) (Dong and Sun, 2003; Zhang, Coutinho, Deng, et al., 2015). In Figure 26 we illustrate our own approach towards cooperative machine learning, which creates a loop between a machine learned model and human annotators for continuous recordings: an initial model is trained (1) and used to predict unseen data (2). An active learning module then decides which parts of the prediction are subject to manual revision by human annotators (3+4). Afterwards the initial model is retrained using the new labelled data (5). Now the procedure is repeated until all data is annotated. By actively incorporating human expert knowledge into the learning process it becomes possible to interactively guide and improve the automatic prediction. Hence, the approach bears the potential to considerably cut down manual efforts. For instance, the system may quickly learn to label some simple behaviours, which already facilitates the work load for human annotators at an early stage. Then over time, it could learn to cope with more complex social signals as well, until at some point it is able to finish the task in a completely automatic manner. Such an iterative approach may even help bridging the gap between quantitative and qualitative coding, which still defines a great challenge in many fields in social science (Chen et al., 2016).

In this chapter we aim at examining to what extent the proposed CML approach helps speeding up the annotation of social signals. In addition, we present the integration of this approach in the NOVA tool, introduced in the last chapter, which allows researches to apply the described techniques to their own databases.

We see the main contributions presented in this chapter as follows:

- In Section 6.2 we propose a novel two-step CML strategy: as long as only few labelled instances are available the system is applied to local fractions of the database. Later, as more labelled instances become available, larger parts can be predicted.
- In Section 6.3 we evaluate the proposed strategy on an audio-based annotation task by simulating the incremental injection of additional information during training. Results show that the proposed strategy significantly reduces manual coding efforts.
- In Section 6.4 we present a walk-through to demonstrate the collaborative annotating capability of the NOVA tool introduced in the last chapter.
- In Section 6.5 experiences from users working with the tool are reported and discussed.

For the sake of clarity related work will be given separately for each section.

6.2 CML Approach

Interactive machine learning (Fails and Olsen, 2003; Amershi, Cakmak, et al., 2014) aims to involve users actively in the creation of models for recognition tasks. Most approaches integrate automated data analysis and interactive visualisation tools in order to enable users to inspect data, process features and tune models. In this section, we focus on approaches that facilitate the acquisition of annotated data sets and introduce a novel methodology for applying *cooperative machine learning* (CML) to speed up annotation of social signals in large multi-modal databases.

6.2.1 Related Work

A common approach to reduce human labelling effort is the selection of instances for manual annotation based on active learning techniques. The basic idea is to forward only instances with low prediction certainty or high expected error reduction to human annotators (Settles, 2012).

An art of its own right is how to estimate which are these most informative ones. A whole range of options to choose from exist, such as calculation of 'meaningful' confidence measures, detecting novelty (e. g. by training auto-encoders and seeing for the deviation of input and output when new data runs through the auto-encoder), estimating the degree of model change the data instance would cause (e.g. seeing whether knowing the label of a data point would make a change to the model at all), or trying to track 'scarce' instances, e.g. trying to find those data instances that are rare in terms of the expected label.

Further more sophisticated approaches aggregate the results of machine learning and crowdsourcing processes to increase the efficiency of the labelling process. Kamar, Hacker, and Horvitz (2012) made use of learned probabilistic models to fuse results from computational agents and human labellers. They showed how to allocate tasks to coders in order to optimise crowdsourcing processes based on expected utility. Zhang, Coutinho, Schuller, et al. (2015) developed an agreement-based annotation technique that dynamically determines how many human annotators are required to label a selected instance. The technique considers individual rater reliability and inter-rater agreement to decide on a combination of raters to be allocated to an instance. Active learning has shown great potential in a large variety of areas including document mining (Tong and Koller, 2001), multimedia retrieval (Wang and Hua, 2011), activity recognition (Stikic, Laerhoven, and Schiele, 2008) and emotion recognition (Zhang, Coutinho, Zhang, et al., 2015b).

Most studies in this area focus on the gain obtained by the application of specific active learning techniques. However, little emphasis is given to the question of how to assist users in the application of these techniques for the creation of their own corpora. While the benefits of integrating active learning with annotation tasks has been demonstrated in a variety of experiments, annotation tools that provide users with access to active learning techniques are rare. Recent developments for audio, image and video annotation that make use of active learning include CAMOMILE (Poignant et al., 2016) and iHEARu-PLAY (Hantke et al., 2015). However, systematic studies focusing on the potential benefits of the active learning approach within the annotation environment from a user's point of view have been performed only rarely (Cheng and Bernstein, 2015; Kim and Pardo, 2017).

While techniques that enable systems to learn from human raters have become widespread, little attention has been paid to usability challenges of the remaining tasks left to end-users (Amershi, Cakmak, et al., 2014). Rosenthal and Dey (2010) investigated which kind of information should be provided to users in order to reduce annotation errors in a setting for active learning. They found out that contextual information and predictions of the learning algorithms were in particular useful for the annotation of activity data. In contrast, uncertainty information had no effect on the accuracy of the labels, but just indicated to the labellers that classification was hard. Amershi, Fogarty, et al. (2009) investigated how to empower users to select samples for training by appropriate visualisation techniques. They found that a representative overview of best and worst matching examples is of higher value than a set of high-certainty images and conjecture that high-certainty images do not provide much information to the learning processing due to their similarity to already labelled images. In another paper by Amershi, Chickering, et al. (2015) the authors suggest an interactive visualisation technique to assess model performance by sorting samples according to their prediction scores. In their tool the user can directly inspect samples to retrieve additional information and annotate them for better performance tracking. This way, the tool allows users to monitor the performance of individual samples while the model is iteratively retrained.

The approaches above supported users in the annotation and selection of samples for training. As an alternative, graphical user interfaces have been developed that enable users to create their own annotated examples for training models. Typically, the labels are given by instructions or stimuli to be provided to the users to evoke particular behaviours. An example includes SSI/ModelUI (Wagner, André, Kugler, et al., 2010). It presents users with a graphical user interface that allows them to test different machine learning algorithms on labelled data. Labels are acquired by stimuli which may include textual instructions, but also images or videos. However, users have to determine themselves which kind of stimuli and data are most useful to create and tune models.

Summing up, it may be said that many studies experientially investigate the potential of novel techniques to minimise human labour.

In addition, few studies were run to actually label novel data, rather than test whether such method could save effort. Also note that the prevailing choice is merely active learning rather than the combination with semi-supervised learning, e.g. cooperative machine learning.

Relatively little attention has been paid, however, to the question of how to make these techniques available to human labellers. There is a high demand for annotation tools that integrate cooperative machine learning in order to reduce human effort — in particular in the area of social signal processing where human raters typically disagree on the labels (Lotfian and Busso, 2017).

In such a setting, dynamic active or cooperative strategies appear particularly promising, e.g. not only learning the target task, but also as much as possible about the raters and their reliability depending on the labels and the content being labelled. Likewise, it can be learned ‘whom to trust when’ to further reduce annotation effort by only requiring labels from the ‘right persons at the right time’.

6.2.2 Active and Semi-Supervised Learning Query Strategies

Before presenting our approach for cooperative machine learning, we'd first like to give an overview on possible strategies on how to decide which segments should be presented to the human annotator for correction. As cooperative machine learning basically consists of active and semi-supervised learning techniques we introduce existing strategies for both sub-areas and for combined approaches. This subsection is based on the work of Settles (2010) and Han et al. (2016).

Active learning (AL) algorithms select unlabelled instances with, ideally, a high potential to improve a model's performance. Settles (2010) generalises three different query strategies:

- The most commonly used strategy is for the active learning component to calculate the **certainty of the predictions** (COP) on previously unlabelled data. For this step, a pre-trained model is employed and the classifications with low confidence are then queried the human oracle (the annotator).

- **Query-by-committee** (QBC) is a strategy where multiple classifiers are trained for the same problem, and the parts where the classifiers disagree the most are provided to the human oracle. In regression tasks such a strategy can also be applied, e.g. by considering the variance among the predictions of the classifiers (Burbidge, Rowland, and King, 2007).

- **Expected error reduction** (EER) methods estimate how likely it is that a model's generalisation error may be reduced (Roy and McCallum, 2001). Such methods often lead to improvements over COP and QBC strategies. Often, the EER method is also the most computationally expensive one.

While active learning methods can greatly reduce human label effort, still a lot of human annotation input is required. Semi-supervised learning (SSL) techniques luckily also aim to handle unlabelled data for training and improving models. Two categories of SSL are currently focused on: self-training and co-training (Settles, 2010).

- **Self-training** allows to annotate unlabelled data automatically by employing a pre-existing model that was trained on a small labelled set first. The main idea is that the model's predictions that have a high degree of confidence are then included in the training set. The classifier is then re-trained with the automatically labelled set. This process is repeated iteratively, so that more and more labels are included in the training set. Such techniques are especially useful to improve the robustness of data-hungry classifiers such as ANNs as no intervention from human annotators is required. Yet, no new information is added in the process which reduces the effect on some classifiers.

- **Co-training** is a "multi-view learning" technique. The idea is that different models are trained concurrently on the same recognition task, yet on different feature sets ("views") (Settles, 2010). In co-training, two models are trained, each with a distinct feature set, yet on the same annotated data. The predictions on the previously unlabelled set of each classifier with high confidence are then included in the training set, so that the models train each other. According to (Han et al., 2016), three assumptions need to be given for the algorithm: "(a) sufficiency: each "view" is sufficient for classification on its own, (b) compatibility: the target functions in both "views" predict the same labels for co-occurring features with high probability, and (c) conditional independence: the "views" are conditionally independent given the class label (Otherwise it would not be guaranteed that new information is added)".

Active learning strategies alone have successfully been applied reduce the time-consuming and expensive human labelling work while at the same time leading to great performance improvements (Settles, 2010). However there are situations where it is not possible or at least not practical to obtain large amounts of human annotations. Here semi-supervised learning delivers the strategies to handle unlabelled data, yet without a human oracle intervening. In order to combine the advantages of both approaches and to overcome their disadvantages, a combination of both techniques seems like a considerable solution.

For example it is imaginable to use self-training for labels with high confidence, while predicted labels with low or medium confidence are passed to a human annotator as suggested in (Zhang, Coutinho, Deng, et al., 2015). In their work they investigated three different possibilities to combine active and semi-supervised learning techniques: (a) single-view cooperative learning (svCL), which combines active learning and Self-Training, (b) mixed-view cooperative learning (xvCL), a combination of active learning and Co-Training, and (c) multi-view cooperative learning (mvCL), which explores the use of coAL and co-Training. They conclude that all of these cooperative learning strategies improve the accuracy and stability of the SVM classifier, with the combination of active learning and Co-Training resulting in the best accuracy and robustness for their given problem.

6.2.3 Two-fold Strategy

The cooperative machine learning strategy we propose here is a two-fold one. It is divided into a *session completion* (SC) step during which information of a fraction of a single session is used to complete the remaining part of the session, and a *session transfer* (ST) step during which information from a set of labelled sessions is used to predict a set of unlabelled sessions. We define a session as a single continuous and self-contained recording. The sessions of a database can be captured on different dates and sites involving different subjects.

The division is motivated by the lack of labelled data in the beginning of an annotation process, which usually does not allow building models that are robust enough to generalise well to the unseen parts. This is especially true if the recording conditions and the involved subjects vary between the individual sessions. Nevertheless, already small fractions of labelled data can be sufficient to build models that are able to make reliable predictions on data that resembles the instances that have been seen so far. An example is data recorded from the same subject under comparable conditions – something we can generally expect from snapshots of the same session. Even if these models are too "weak" to make reliable predictions for the whole dataset, they can help to speed up the early annotation process. In the following, we refer to a classifier trained on samples of a single session as a *session-dependent* classifier. Once enough sessions have been completed, a *session-independent* model can be trained and used to accomplish remaining sessions.

To ensure the quality of the recognition, manual verification of the outcome of the classification might be necessary. This procedure can be accelerated by rating the predictions, e.g. by adding confidence values to the predicted instances. Instead of reviewing everything annotators can concentrate on parts with *low confidence*, e.g. labels that have been predicted with a high uncertainty[1]. The proposed strategy can be summarised as follows:

1. **Session completion**: manually assign labels to a fraction of a session and train a session-dependent classifier. Apply it to complete the remaining fraction. Based on the confidence values generated by the model, query a human for manual revision.

2. **Session transfer**: take all (with aid of step 1) fully labelled sessions and train a session-independent classifier. Apply it to predict annotations for remaining sessions. Again, based on the generated confidence values decide which parts require manual adjustment.

[1] In a multi-class classification task uncertainty can e.g be derived from the distance a predicted sample has to the decision boundaries of the other classes. We will subsume this strategy here, but one, or a combination of multiple query strategies proposed in Section 6.2.2 could also be applied.

So far we have distinguished between session-dependent and session-independent classification. Depending on the corpus to which the strategy is applied, this may not necessarily be the best practise. For instance, if a dataset is composed of recordings that are too short to apply the first step we can adapt the strategy and initially complete recordings belonging to the same subject. Once we have labelled data from a sufficient number of individual subjects, we continue by training a subject-independent model and apply it to the remaining recordings. Likewise, we can use the described strategy across several databases, too. In that case we would concentrate on individual databases first and afterwards obtain a database-independent model that we use to label the remaining databases.

6.2.4 Implementation

To efficiently apply the described strategy, we would like to know the *sweet spot* for applying the *session completion* and the *session transfer* step. On the one hand, if we apply it too early the model becomes unstable and predictions will be poor. On the other hand, if we annotate more data than necessary we give away precious time. To avoid any of the described situations, we are interested in finding a good trade-off between machine performance and human effort. Unfortunately, we cannot easily guess what is the ideal moment to hand over the task to a machine. This is because the amount of training data that is required to build a robust model depends on a number of factors, such as the homogeneity of the data, the discrimination ability of the extracted features, the number of subjects and classes, and not least at the complexity of the recognition problem. Alternatively, instead of trying to determine a sweet spot beforehand (and possibly miss it), we could iteratively test the applicability of the strategy and stop when the performance seems promising.

Therefore, we opted to make the described strategy an integral part of our tool (see Section 6.4). This allows annotators to visually examine the results at any time and to individually decide whether more labelling is required or not. However, this means that the time it takes to run the CML strategy becomes a crucial factor. Generally, it should not take longer than a few seconds or the annotation process will be interrupted (this is especially true for the session completion step). To reach this goal, we should reuse as much information as possible. One possibility is to apply classification on a small sliding window (frames) and use a rather simple (e. g. linear) classifier. Working on frames of a fixed size means that features have to be extracted only once (or can be even pre-extracted), and do not have be recalculated in case the segment length changes. A simple classifier ensures fast training, more advanced classifiers might result in better accuracy. In our approach we presume that classifiers deliver a "confidence" value for their decisions. This might be an actually probability or e.g. in case of a linear SVM we define the normalised distance to the hyper-plane for each class as "confidence" value.

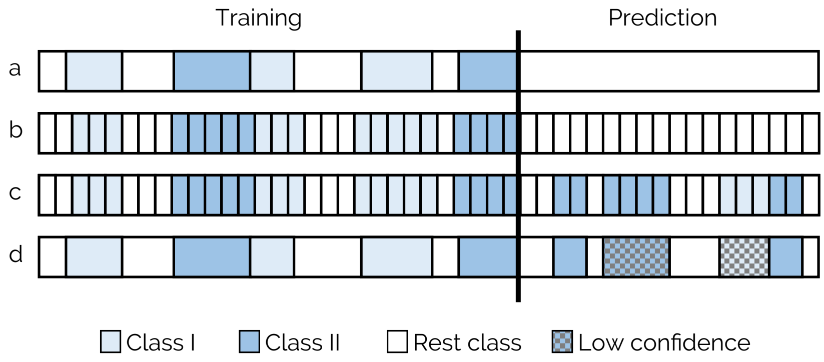


Figure 27.: Visualisation of the cooperative machine learning strategy by means of the SC step: (a) the end point of the last segment of the manual annotation defines where the training fraction ends and prediction begins. (b) labelled segments are mapped onto frames of a fixed size and frames without a label are assigned to a temporary rest class. (c) a model is build from the frames in the training fraction and used to predict the frames in the prediction fraction. (d) successive frames with the same class label are combined, the rest class is removed (as it was not in the original annotation scheme) and segments with a low confidence are highlighted.

In the following, we restrict ourselves to complete discrete annotations, i. e. we deal with multi-class problems. In case of the SC step we receive the raw signal stream (e. g. an audio signal) of the current session and a partly finished annotation composed of labelled segments with a discrete start and end point. The segments can be of variable length and there may be gaps between two successive segments. By applying the following procedure we then predict the segments for unlabelled fraction of the session (see also Figure 27):

1. If not provided, extract frame-wise features for the whole session.

2. Find the frame that coincides with the end point of the last label in the annotation and split the feature sequence into a training fraction (preceding frames) and a prediction fraction (successive frames).

3. In the training set assign frames that overlap with a labelled segment by at least 50% to the corresponding class. In case of several candidates keep the dominant one (most overlap). Assign remaining frames to a rest class.

4. Learn a classifier using all frames from the training fraction.

5. Use the classifier to label the prediction fraction by assigning to each frame the class with the highest confidence.
6. Combine successive frames belonging to the same class and keep the average confidence of the combined frames to make a statement about the confidence of the whole label. Remove frames that belong to the rest class. Optionally, apply thresholds to remove very small segments and fill small gaps to avoid unwanted micro segments.
7. Add the predicted segments to the original annotation and mark segments with a low confidence.

The ST step works in the same way with the difference that whole sessions are used to train the classifier, which is applied to predict whole sessions afterwards.

6.3 Evaluation

We now turn to some experiments in which we examine the practical effect of the proposed cooperative machine learning (CML) strategies of Section 6.2. We do this by means of a database including natural human-human interaction and simulate a situation where the detection system is applied to predict unlabelled fractions of the dataset. Using the original and predicted parts of the corpus to train a final detection model we evaluate the robustness and efficiency of the CML approach.

6.3.1 Database and Problem Description

As database we use the NOXI corpus, as introduced in Section 3.3. One purpose of NOXI is to study interruption strategies. For instance, when a listener decides to ask a question or comment to what the speaker was saying and therefore starts an attempt to take over the speech turn. The simplest way to detect such situations is by looking for spots where the voice of the two participants is overlapping. If afterwards a speaker change occurs we can assume that the interrupting party successfully took over the turn. Otherwise we can treat it as a failed attempt. However, an interposed utterance is not necessarily a signal to interrupt the speaker. It can also be an expression of approval or interest, denoted as *backchannels*. Likewise, not every speaker pause signals a floor change if, for instance, the speaker needs time to think what to say next. To bridge these pauses speakers usually utter a *filler* sound. Hence, to correctly identify speaker interruptions we have to separate backchannels and fillers from other speech parts.

In the following, we present a detection system that is trained to automatically identify backchannels and fillers in speech. First, we evaluate the system following a classic machine learning approach to measure the performance of the system. Afterwards, we examine if and to what extent the system is able to speed-up the manual annotation process in the CML loop.

6.3.2 Detection System

Though in our experiments we concentrate on the detection of speech and fillers/ backchannels, we opt for a detection system that is as generic as possible. This will allow us to apply it to other classification problems, too. Also, speed performance plays a crucial role as we do not want to interfere with the annotation process. In the following we start by describing the proposed generic detection system.

Due to its modularity and capability of fast online incremental processing we rely on the OpenSmile audio feature extraction tool (Eyben, Weninger, et al., 2013). However, we refrain from using a large statistical feature set like the ComParE (Computational Paralinguistic Evaluation) set, which assembles 6 373 features by brute-force combination of low level descriptors (LLDs) with functionals (Schuller, Steidl, et al., 2013). This kind of feature sets are usually applied on chunks of several seconds length (e. g. a whole utterance). In our scenario, however, we opt for a frame-based feature set extracted over a small moving window that can be reused across successive training steps. Also, we should keep in mind that especially in the beginning of the annotation process the size of the training sets can be small. In that case a smaller feature set will lower the risk of overfitting.

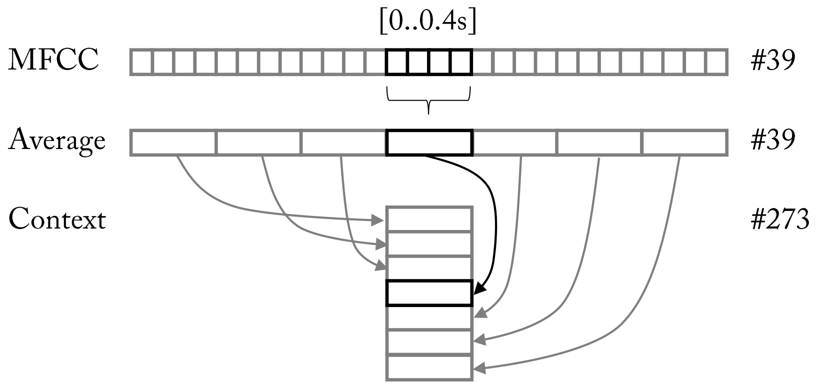


Figure 28.: Illustration of the feature extraction step. First, four MFCC frames with a dimension of 39 are averaged to reduce the sample rate of the signal to 25 Hz. Afterwards, neighbouring frames are added, here 3 frames from the left and 3 frames from the right. This results in a final feature vector of size 273.

Mel-frequency cepstral coeffcients (MFCC) provide a compact representation of the short-term power spectrum. Not only have they a long tradition in speech recognition systems (Rabiner and Juang,

1993) and speaker verification tasks (Ganchev, Fakotakis, and Kokkinakis, 2005), but have also been successfully applied in the field of social signal processing, e. g. emotional speech recognition (Lee et al., 2004; Vogt and André, 2005; Beritelli et al., 2006; Neiberg, Elenius, and Laskowski, 2006; Schuller, Batliner, et al., 2007; Kishore and Satish, 2013) and laughter detection (Kennedy and Ellis, 2004; Knox and Mirghafori, 2007; Urbain et al., 2010). For our tests, we calculate 13 Mel-frequency cepstral coefficients (including the 0th coefficient) and their first- and second-order frame-to-frame difference (delta-delta). According to standard practice we use a moving window of 25 ms with a frame step of 10 ms. Afterwards we reduce the stream to a frame step of 40 ms by averaging always four frames. This ensures that the sample rate of the feature stream is consistent with the video frame rate of 25 Hz. Though not relevant for the current study, such a sample rate comes in handy for visual features. Yet, 40 ms are small enough to detect start and end point of voiced segments sufficiently accurate. Since the length of the filler events we want to detect may be longer than 40 ms, we optionally concatenate neighbouring frames from both sides of the current frame – in the following denoted as context size n. A context of size 3, e. g. means that the current frame is extended by 3 frames from the left and 3 frames from the right. This increases the number of features by a factor of $2 \cdot n + 1$. Figure 28 illustrates the feature extraction step.

As classification model we use a linear support vector machine (SVM) provided by LIBLINEAR – a library for large linear classification (Fan et al., 2008). Since the implementation does not use kernels, training time is significantly reduced even for large input sets composed of several ten thousand samples. For multi-class classification we select a L2-regularised logistic regression solver (option -s 0) and add a bias term of 0.1 (option -B 0.1). We keep default values for all other parameters. Since we expect unbalanced class distributions, we randomly remove samples to match the size of the class with the least number of samples. Finally, features values are scaled between -1 and 1 (when we test a sample we apply the scaling derived from the training set). Confidence values are scaled in a way such that individual class scores sum to 1.

6.3.3 Results

Having established a generic classification system we will now evaluate recognition performance on the NOXI corpus (see Section 6.3.1). We pick 18 sessions (German sub-corpus) and randomly split them into a training set including two-third of the sessions summing up to nearly 7 h of audio data. The remaining 6 sessions form the test set with an overall duration of almost 3.5 h.

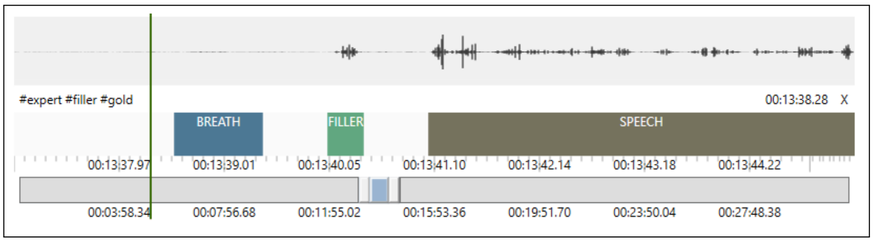


Figure 29.: Example of a manual annotation.

To evaluate the proposed detection system we need to establish a ground truth. We use NOVA to manually annotate voiced parts in the audio files. To not introduce a machine bias none of the CML strategies described in Section 6.2 are applied. Manual annotation is accomplished by three experienced annotators[2], each completing six sessions. Table 5 lists the applied annotation scheme. Since labels are assigned to voiced sounds the remaining parts implicitly define the rest class SILENCE. Because of the better audio quality we use the head set recordings. However, it turned out that the close-talk recordings tended to pick up breathing sounds, so we introduce an additional BREATH class to prevent false alarms during silenced parts. Backchannels, fillers, laughter, and other voiced sounds such as grunts and coughs, are gathered in a single class denoted as FILLER. Speech segments that are neither backchannels nor fillers are labelled as SPEECH. An example of an annotation is shown in Figure 29. We asked the raters to measure how long it took to annotate the sessions. In total the annotators spent a little more than 14 h, which results in an average time of 47 minutes per session.

Table 5.: Annotation scheme and frame number per class.

Class	Description	Train	%	Test	%
SPEECH	Speech (except filler and backchannels)	265 466	41.4	126 183	41.2
BREATH	Breathing (except unvoiced laughter)	22 918	3.6	3 929	1.2
FILLER	Backchannels, fillers, laughter, and other voiced sounds	26 665	4.2	8 592	2.8
SILENCE	Implicit rest class representing unvoiced parts	325 528	50.8	167 502	54.7
$\sum$		640 577		306 206	

Next, we split the annotations in frames of 40 ms length and extract MFCC features, which results in 946 783 frames (exact class distribution are given in Table 5). We down-sample the training set to 22 918 samples per class and train a linear SVM model. Results are summarised in Table 6. We report class-wise recognition accuracy and unweighted average (UA) recall (average across classes). For a direct comparison with the INTERSPEECH 2013 social signals paralin-

[2] Two research assistants who have been working in the field of SSP for several years and one master student who took part in an annotation course.

guistic challenge we also consider the area under the curve (AUC) measure. A 85% unweighted average AUC for the FILLER class (best case) shows that results are comparable to (Schuller, Steidl, et al., 2013) (who achieved a UAAUC of 87.6% for their development and 83.3% for the test set in the INTERSPEECH 2013 challenge on filler detection). We take this as evidence that our detection system does a reasonable job on the examined task.

Table 6.: Class-wise recall and area under the curve (in brackets) in % with respect to the context n. A context of n=5 here means that we consider the 5 frames left and right of the actual frame in the classification.
UA = Unweighted average, UAUAC = UA of AUC

n	0	1	2	5	10	15
SPEECH	64.7 (95)	67.6 (96)	69.5 (96)	73.7 (97)	74.6 (97)	74.3 (97)
BREATH	82.5 (95)	84.3 (96)	85.1 (97)	87.2 (98)	87.9 (98)	88.2 (98)
FILLER	46.6 (69)	54.1 (74)	59.1 (77)	66.1 (82)	71.9 (84)	74.1 (85)
SILENCE	82.9 (92)	83.1 (93)	83.9 (94)	85.5 (95)	84.0 (96)	82.8 (96)
UA (UAUAC)	69.2 (88)	72.3 (90)	74.4 (91)	78.1 (93)	79.6 (94)	79.8 (94)

As seen in Table 6 increasing the number of concatenated frames has a positive effect on the recognition accuracy (~10%). Especially the FILLER class benefits from a larger frame context (25% improvement), which we explain with the fact that fillers are usually short and isolated speech episodes surrounded by silence. In Figure 30 we notice a saturating effect for more than 10 frames. Also, we must not forget that in an online recognition system each additional frame that we look into the future introduces extra delay. For this reason, we decided to stick with a stacked context of 5 introducing a lag of 0.2 s, which we found still tolerable.

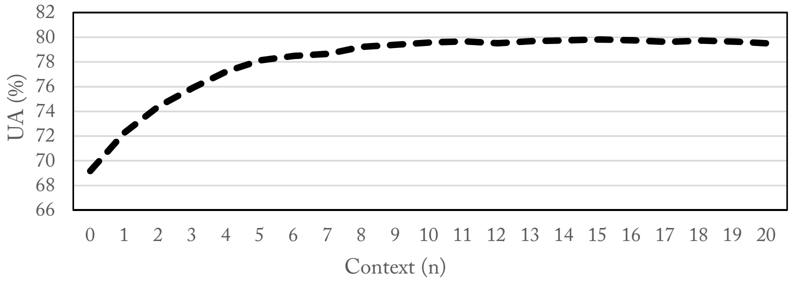


Figure 30.: Classwise UA recall in % with respect to the context size n.

To give an impression of how the system performs in terms of speed we report measurements on an Intel(R) Core(TM) i7-3930K. In our tests extracting MFCC-based features with a context of size 5 and a frame step of 0.04 s took 0.9 s for one minute of mono audio sampled at 48 kHz. Extrapolated to 10.5 h of interaction it requires less than 10 minutes to extract features for the whole German subset. Since features are reused this defines a one-time effort. Training a linear classifier on the training set (91 672 frames after class balancing) took on average 50 s. Frame-wise prediction on the test set (306 206 frames) only ~ 2.9 s. Such values suggest that the proposed detection system is fast enough to be embedded into the annotation process without causing serious interruptions (even if several hours of data are used as input / output).

6.3.4 CML Simulation

Finally, we want to know how the proposed detection system performs in combination with the proposed CML strategies. In Section 6.2 we have defined the *sweet spot* as the moment when additional annotation efforts no longer improve the stability of the classification model. Practically, this defines the ideal point to hand the task over to the machine.

To experientially determine the sweet spot for the given problem, we incrementally inject information into the training process. In the following, we simulate this procedure by splitting the original training set into two parts: we assume that n sessions have been manually labelled (subset L), whereas the remaining sessions are yet unlabelled (subset U). Now, we derive three classifiers c, c' and c'' (see Figure 31):

c Train with the labels of L.

c' Use c to predict the labels of U and retrain with the predicted labels.

c'' Before retraining inspect the predicted labels if their confidence is below a threshold t and correct them if necessary.

c' simulates the case where the annotation process is stopped at some point and the labelled fraction of the database is used to predict the remaining parts. Note that in this case *all* predicted labels are included during the final training step, i. e. no automatic selection strategies and no additional manual efforts are applied.

c'' simulates the case where parts of the prediction are inspected (here the selection is based on the class confidence). To assess the additional manual effort we measure what we call the *inspection rate* (IR), which is the fraction of frames below the confidence, and the *correction rate* (CR), which is the fraction of frames that are finally assigned a different label.

Table 7 summarises the performance of c, c' and c'' on the test set (the same as before). In each row we assume that n sessions of the original training set have been labelled (e. g. $n = 4$ means that L

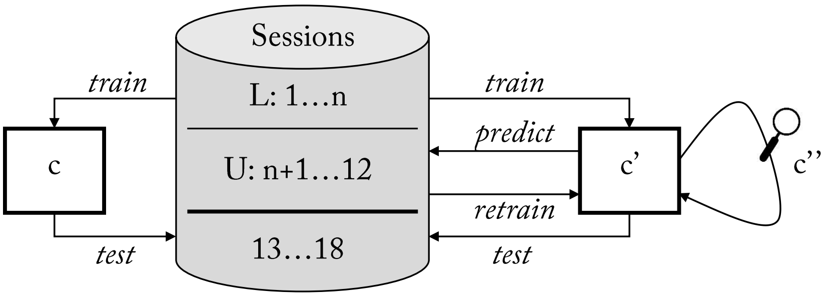


Figure 31.: In the default condition a classifier c is evaluated after training with labelled sessions (L) only. In case of c′ unlabelled sessions (U) are predicted and used to retrain the model. And in case of c″ predicted labels are reviewed and possibly corrected before retraining takes place.

consists of sessions 1 to 4 and U consists of sessions 5 to 12). Based on the results we can gain some interesting insights. Let us therefore assume we aim for a classification model that is at maximum one percent worse than the reference model trained on all sessions, i. e. has an unweighted average (UA) recall of at least 77.1% (throughout the tests we have applied a stacking context of 5).

The performance of classifier c shows that labelling ten of the twelve sessions are sufficient to yield a 77.8% recognition accuracy. Hence, to achieve our goal we can stop after labelling ten sessions and skip the last two. Now, what happens if we extend the training set with predicted labels (no selection or manually correction yet)? Checking the results of c′ we see that again ten sessions are required to achieve the desired accuracy. In fact, extending the training set with purely predicted data generally has no positive effect on the recognition performance. Although disappointing at first glance this is actually not too surprising. Obviously we cannot expect to improve a model unless we inject some new knowledge, which is not the case if we add predictions without inspection. This is as if we asked a student to revise his own test, which is pointless unless we point out some of his mistakes first.

Hence, some manual efforts are needed here. And indeed: after correcting frames with a confidence below 0.5 (that is 9% of all frames in the remaining subset) c″ yields 77.1% already after initially labelling only 6 sessions. To achieve this we actually had to review 27% of predicted frames. If we assume that the remaining six sessions make up approximately half of the frames this corresponds to $\frac{1}{8}$ of the full training set, i. e. in total we have to examine $\frac{5}{8}$ (= $\frac{1}{2} + \frac{1}{8}$) of the training data. As mentioned earlier the average time to annotate a session was 47 minutes. Hence, we can reckon a saving of approximately 3.5 hours (5.9 h instead of 9.4 h). Obviously, this significantly speeds up the annotation process.

Table 7.: Recognition results on the test when incrementally injecting information into the training process using the three classifiers c, c', c'' (see remarks in text). In case of c'' t defines the confidence threshold for inspecting predicted labels. In each row we start with n labelled sessions. Results are obtained with the detection system described earlier using a stacking context of 5.

	c	c'	c'' ($t = 0.5$)			c'' ($t = 0.75$)		
n	UA (%)	UA (%)	UA (%)	IR	CR	UA (%)	IR	CR
1	67.2	70.1	74.1	38%	14%	**77.4**	**87%**	**25%**
2	72.3	70.5	74.3	51%	17%	78.0	82%	25%
3	73.0	71.6	76.0	36%	12%	77.9	66%	18%
4	74.4	73.2	76.4	36%	12%	78.0	59%	18%
5	76.2	75.8	76.9	31%	11%	78.0	51%	16%
6	76.3	76.4	**77.1**	**27%**	**9%**	78.0	45%	14%
7	76.4	76.4	77.2	25%	9%	78.2	40%	13%
8	77.0	76.3	78.0	15%	5%	78.0	26%	7%
9	76.8	76.9	78.1	11%	4%	78.1	19%	6 %
10	**77.8**	**77.8**	78.0	5%	2%	77.9	10%	2%
11	78.1	77.9	78.1	1%	0%	78.1	4%	1%
12	78.1	78.1	78.1	-	-	78.1	-	-

Apparently, the more work we are willing to spend on the correction of predicted labels the earlier we receive a stable classification model. In fact, if we lift the correction threshold to 0.75 we observe that c'' now yields 77.4% already after the first session. However, this is achieved at the expense of a more than three times higher inspection rate (87%), which means that we have to view almost $\frac{9}{10}$ of the corpus (precisely $0.87\frac{9}{10} + \frac{1}{10}$). Hence, it can be a better strategy to complete a couple of sessions first and in return apply a smaller correction threshold afterwards leaving less data for inspection (more on that in Section 6.5).

6.4 NOVA Integration of CML Strategies

The results of the previous section encouraged us to integrate the proposed cooperative machine learning (CML) approach into our annotation tool NOVA (see Chapter 5). This way we give annotators the possibility to immediately inspect and if necessary correct predicted annotations.

In the following we will concentrate on one particular feature of NOVA that has not been discussed before in detail: the use of CML tools to speed up the annotation of large multi-modal corpora. The

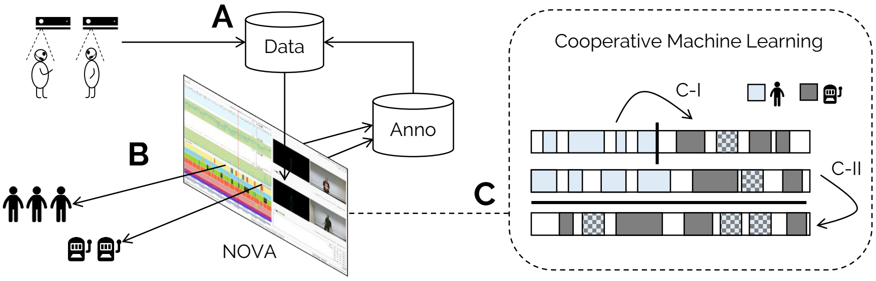


Figure 32.: CML integration in NOVA: (A) A database is populated with recordings of human interaction. (B) NOVA functions as interface to the data and provides a database to distribute and accomplish annotation tasks among human annotators. (C) At times, CML is applied to automatically complete unfinished fractions of the database: (C-I) A session-dependent model is trained on a partly annotated session and applied to complete it. (C-II) A pool of annotated sessions is used to train a session-independent model and predict labels for the remaining sessions. In both cases, confidence values guide the revision of predicted segments (here marked with a pattern).

general scheme of the integration is shown in Figure 32. It shows NOVA as a mediator between the database and several human and machine annotators. Both CML steps described in Section 6.2 are supported by the interface.

6.4.1 Machine Learning Backend

For best possible performance, tasks related to machine learning (ML) are outsourced and executed in a background process. As ML framework we use our open-source Social Signal Interpretation (SSI) framework, as introduced in Section 4.2. SSI has been successfully applied to a couple of recognition problems in the past, see e. g. (Urbain et al., 2010; Wagner, André, Lingenfelser, et al., 2011; Lingenfelser, Wagner, and André, 2011; Lingenfelser, Wagner, André, et al., 2014). Since SSI is primarily designed to build online recognition systems, a trained model can be directly used to detect social cues in real-time (Wagner, Lingenfelser, et al., 2013).

Though, SSI is developed in C++, it offers a XML interface to define feature extractors and classifiers. For instance, the definition of the MFCC features from Section 6.3.2 looks as follows:

```
<chain>
  <!-- load components -->
  <register name="audio"/>
  <meta frameStep="10ms" leftContext="15ms"/>
  <!-- apply filtering -->
  <filter>
    <item create="PreEmphasis"/>
  </filter>
   <!-- extract features -->
  <feature>
    <item create="Mfcc" option="mfcc"/>
  </feature>
</chain>
```

When applied to a stream, the signal values are first run through a pre-emphasis filter before MFCC features are extracted over a sliding window of 25 ms with a frame step of 10 ms (timings can be overwritten in NOVA). To configure the MFCC extraction (e. g. the number of coefficients) a separate option file is created (here 'mfccdd'). However, SSI supports other features sets, too. For instance, it allows to run scripts from the widely used OpenSmile toolkit (Eyben, Weninger, et al., 2013). And it provides feature sets for other types of signals. For instance, a wrapper for the OpenFace tool (Baltrusaitis, Robinson, and Morency, 2016) is available to extract of facial points and action units from video streams.

Likewise, the classification model from Section 6.3.2 is defined as follows:

```
<trainer>
  <register name="model"/>
  <!-- apply under sampling -->
  <meta balance="under"/>
  <!-- scale the features -->
  <normalize>
    <item method="Scale"/>
  </normalize>
  <!-- apply the classifier: Linear Support Vector Machine -->
  <model create="LinSVM" option="svm"/>
</trainer>
```

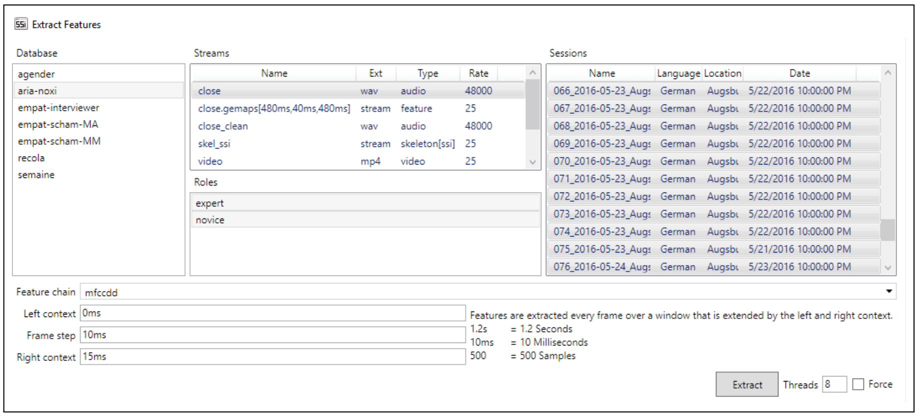
Here, SSI is configured to balance the number of class samples by removing samples from overrepresented classes and scale features into a common interval. As training model a linear SVM will be used. However, our framework also supports other classification models such as Google's neural network framework TensorFlow[3] or the popular Theano[4] library.

6.4.2 CML Walk-through

We will finish this section with a walk-through that demonstrates NOVA's CML tools. We assume that a database has been created and populated with several sessions feature audio recordings of one or more users. In our case, we work on the NOXI database described in Section 3.3 and apply the annotation scheme used during the evaluation in Section 6.3.3, i. e. we want to mark filler and breath

[3] https://www.tensorflow.org/
[4] https://github.com/Theano/

events in regular speech by assigning the labels BREATH, FILLER and SPEECH.



Extract Features

Database

- agender
- aria-noxi
- empat-interviewer
- empat-scham-MA
- empat-scham-MM
- recola
- semaine

Streams

Name	Ext	Type	Rate
close	wav	audio	48000
close.gemaps[480ms,40ms,480ms]	stream	feature	25
close_clean	wav	audio	48000
skel_ssi	stream	skeleton[ssi]	25
video	mp4	video	25

Roles

- expert
- novice

Sessions

Name	Language	Location	Date
066_2016-05-23_Augs	German	Augsbu	5/22/2016 10:00:00 PM
067_2016-05-23_Augs	German	Augsbu	5/22/2016 10:00:00 PM
068_2016-05-23_Augs	German	Augsbu	5/22/2016 10:00:00 PM
069_2016-05-23_Augs	German	Augsbu	5/22/2016 10:00:00 PM
070_2016-05-23_Augs	German	Augsbu	5/22/2016 10:00:00 PM
071_2016-05-23_Augs	German	Augsbu	5/22/2016 10:00:00 PM
072_2016-05-23_Augs	German	Augsbu	5/22/2016 10:00:00 PM
073_2016-05-23_Augs	German	Augsbu	5/22/2016 10:00:00 PM
074_2016-05-23_Augs	German	Augsbu	5/22/2016 10:00:00 PM
075_2016-05-23_Augs	German	Augsbu	5/21/2016 10:00:00 PM
076_2016-05-24_Augs	German	Augsbu	5/23/2016 10:00:00 PM

Feature chain: mfccdd

Left context: 0ms

Frame step: 10ms

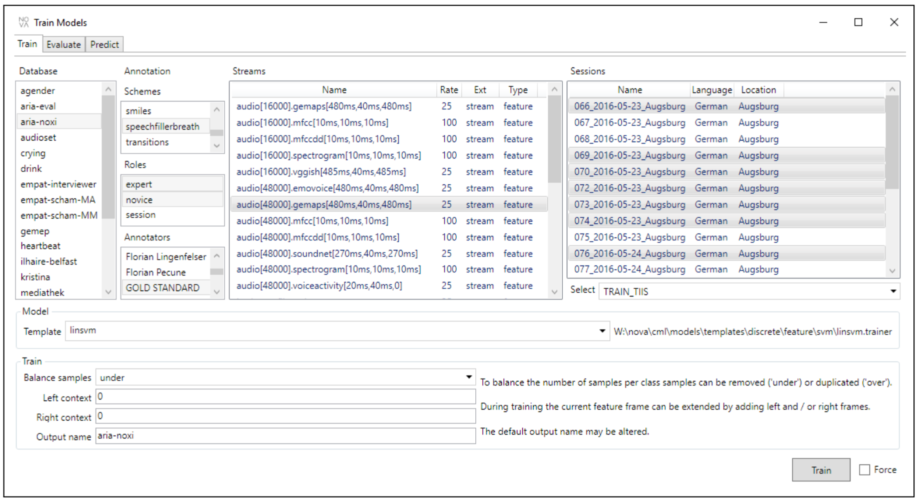
Right context: 15ms

Features are extracted every frame over a window that is extended by the left and right context.

1.2s = 1.2 Seconds
10ms = 10 Milliseconds
500 = 500 Samples

Figure 33.: Screenshot of the feature extraction dialogue. The user chooses a stream (here audio) and an according feature extraction method (here mfccdd). Feature extraction is applied for the selected roles and sessions.

As a first step, we extract MFCC features for the German sessions in the NOXI database. The dialogue is shown in Figure 33. It allows us to choose a source stream and a feature extraction method (only methods that can be applied to the selected stream will be listed). Optionally, we can overwrite the default frame step and context sizes. Extraction can be accelerated by running several sessions in parallel (here 8).



Train Models

Database

- agender
- aria-eval
- aria-noxi
- audioset
- crying
- drink
- empat-interviewer
- empat-scham-MA
- empat-scham-MM
- gemep
- heartbeat
- ilhaire-belfast
- kristina
- mediathek

Annotation

Schemes

- smiles
- speechfillerbreath
- transitions

Roles

- expert
- novice
- session

Annotators

- Florian Lingenfelser
- Florian Pecune
- GOLD STANDARD

Streams

Name	Rate	Ext	Type
audio[16000].gemaps[480ms,40ms,480ms]	25	stream	feature
audio[16000].mfcc[10ms,10ms,10ms]	100	stream	feature
audio[16000].mfccdd[10ms,10ms,10ms]	100	stream	feature
audio[16000].spectrogram[10ms,10ms,10ms]	100	stream	feature
audio[16000].vggish[485ms,40ms,485ms]	25	stream	feature
audio[48000].emovoice[480ms,40ms,480ms]	25	stream	feature
audio[48000].gemaps[480ms,40ms,480ms]	25	stream	feature
audio[48000].mfcc[10ms,10ms,10ms]	100	stream	feature
audio[48000].mfccdd[10ms,10ms,10ms]	100	stream	feature
audio[48000].soundnet[270ms,40ms,270ms]	25	stream	feature
audio[48000].spectrogram[10ms,10ms,10ms]	100	stream	feature
audio[48000].voiceactivity[20ms,40ms,0]	25	stream	feature

Sessions

Name	Language	Location
066_2016-05-23_Augsburg	German	Augsburg
067_2016-05-23_Augsburg	German	Augsburg
068_2016-05-23_Augsburg	German	Augsburg
069_2016-05-23_Augsburg	German	Augsburg
070_2016-05-23_Augsburg	German	Augsburg
072_2016-05-23_Augsburg	German	Augsburg
073_2016-05-23_Augsburg	German	Augsburg
074_2016-05-23_Augsburg	German	Augsburg
075_2016-05-23_Augsburg	German	Augsburg
076_2016-05-24_Augsburg	German	Augsburg
077_2016-05-24_Augsburg	German	Augsburg

Select: TRAIN_TIIS

Model

Template: linsvm W:\nova\cml\models\templates\discrete\feature\svm\linsvm.trainer

Train

Balance samples: under

To balance the number of samples per class samples can be removed ('under') or duplicated ('over').

Left context: 0

Right context: 0

During training the current feature frame can be extended by adding left and / or right frames.

Output name: aria-noxi

The default output name may be altered.

Figure 34.: Screenshot of the model training dialogue. The user selects a coding scheme, a role and an annotator (here Gold Standard). Sessions for which an according annotation exists are now displayed and a stream can be selected to define the input for the learning step. Finally, a model (here linsvm) is chosen and the training begins.

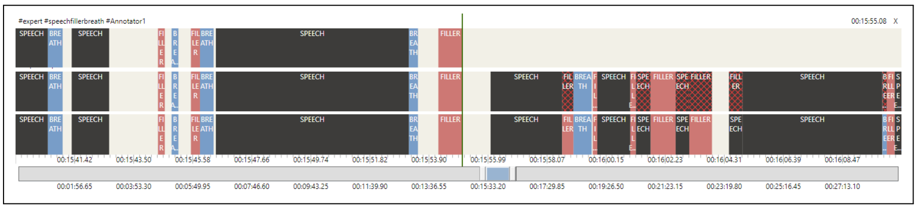
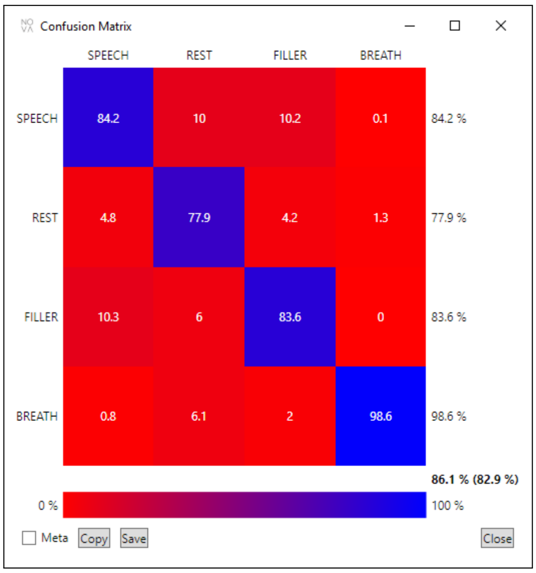


Figure 35.: Visualisation of partly finished annotation (upper tier) and the results after the tier is automatically completed (middle tier). Segments with a low confidence are marked with a red pattern. The lower tier shows the final result after manual correction.

In a next step, we can now pick an annotation scheme and apply it to the previously extracted feature streams. Figure 34 shows the interface that allows us to select the input and choose a classification model (only models are shown that fit the selected input). Optionally, we can set a left and right context to concatenate neighbouring feature frames (see Section 6.3.2). Afterwards the trained model is stored and can now be applied to predict unlabelled data.


Confusion Matrix

	SPEECH	REST	FILLER	BREATH	
SPEECH	84.2	10	10.2	0.1	84.2 %
REST	4.8	77.9	4.2	1.3	77.9 %
FILLER	10.3	6	83.6	0	83.6 %
BREATH	0.8	6.1	2	98.6	98.6 %

86.1 % (82.9 %)

Figure 36.: A confusion matrix provides information about the recognition accuracy of individual classes and to what extent they are confused with other classes. For instance, here we see that speech frames are often falsely classified as fillers and vice versa. Hence, an annotator should put attention to these classes while revising the predictions. The REST class implicitly represents *silence* in this example.

To predict annotations, both CML strategies from Section 6.2.3 are available. In case of *session transfer* a dialogue similar to the one in Figure 35 is shown. However, this time we select a previously trained model and use it to predict the selected sessions. In case of the *session completion* step, the annotation is completed by temporarily training a model using only the labels available from current tier. An example before and after the completion is shown in Figure 35. The screenshot shows that labels with a low confidence are highlighted with a pattern. This way crucial parts are quickly found and can be revised if necessary.

To assess the prediction accuracy of a model, a dialogue similar to Figure 34 is available. Here, we can pick a trained model and the sessions we want to use for evaluation (only sessions with an according annotation are listed). Additionally, we can pre-define session sets, so all users of the database may use a pre-selection of sessions for training, prediction and evaluation. The model is now applied to predict labels for the selected sessions and the output is compared to the existing annotations. The result is presented in form of a confusion matrix as shown in Figure 36. A confusion matrix provides information on the overall recognition performance, as well as, accuracies for individual classes and which class pairs are often confused. For regression problems, the Pearson correlation coefficient is shown instead.

6.5 Experiences and Discussion

In Section 6.3 we have presented a technical evaluation of our proposed cooperative machine learning (CML) strategy. Results show that CML bears great potential to significantly reduce human labelling effort. However, it does not necessarily mean that results gained in a simulation can be transferred to human annotators without further ado. Hence, in the following, we want to discuss the experiences of users who have been applying the CML strategies with the NOVA tool introduced in the previous section.

6.5.1 What gain is to Expect?

So, what exact gain can we expect when giving a tool like NOVA into the hands of human labellers? Unfortunately, a general answer to this question does probably not exist. Our experiences show that the amount of time we may save depends on a couple of variables, which may vary from one case from one case to another.

Probably, the largest uncertainty comes from the nature of the annotation problem itself and the ability of the applied machine learning (ML) techniques to cope with it. For instance, let us assume the task of labelling voiced parts in audio. If the recordings have low background noise and speech is really the only prominent signal, a simple feature like loudness may already allow us to train a robust model on few samples, yet generalising well on unseen data. In this case, the time saving (compared to a completely manual approach) can be tremendous. On the other hand, if the speech files are noisy and contain other audible sounds – possibly overlapping with speech – the problem becomes immediately harder. As a consequence not only a more sophisticated feature set (and classification model) is needed, but more manual labelling effort is required to obtain a robust model. As a consequence, less time is saved. And we may even reckon the case where the problems becomes too hard to train a reliable model at all, so that the effort to manually revise the prediction may eat up initial savings. The possibility to exchange features and classifiers in NOVA is therefore an essential precondition to adapt to the problem at hand in the best possible way.

Another point to consider is the quality of the annotation that is desired. Can we live with some false predictions? Or do we aim for a high precision, yet do not mind a high number of false negatives? This, of course, depends very much on the purpose the data is labelled for. As a special flaw social signals often lack a ground truth. And when multiple raters are employed the agreement often turns out to be low. This makes it specially difficult to estimate the quality of a prediction. In the end, it depends a lot on the assessment of the user if he or she is pleased with the automatic completion. Here, NOVA's feature to immediately visualise the results is an important tool to let raters assess the quality of automatic predictions.

Finally, comparing manual with semi-manual annotations is not as straight forward as it may seem. When observing automatic predictions we observed that on- and offset of the labels were often more precise than that of humans, which are usually rather fuzzy (unless they work at a very fine granular time scale, which is usually too time-consuming). Likewise, we found that short occurrences of a behaviour are easily overlooked by human labellers, especially as their attention drops with time. Hence, since machines show no signs of fatigue their predictions are often more consistent throughout a corpus compared to those of humans. Consequently, applying CML strategies may not just help saving time, but also lead to more accurate and stable annotations.

6.5.2 Experienced Annotators

To learn more about the general applicability of our approach, we asked three annotators we had hired for the manual annotation of the NOXI database (see Section 6.3.1) to redo some of their sessions. The task was again to annotate speech and filler events. This time, however, we explicitly told them to make use of the integrated CML tools. Afterwards we compared the new semi-automated annotation with the previous manual ones and also asked for their subjective impressions.

First of all, they reported that they were positively surprised by the accuracy of the automatically generated labels. Only little manual efforts were required to correct the predictions for the given task. In particular, they found that in some cases detections were even more precise compared to their previous manual annotation. We explain this with the fact that the human brain naturally filters out information it perceives as not relevant in the current context. For instance, we do not consciously hear short breathing sounds during an utterance since we concentrate on the content of the spoken message. And even if we force ourselves to pay special attention to certain events it may be too exhausting to accurately label each and every occurrence. This is a situation in which the semi-automated annotation can really pay off as machines – in contrast to humans – do not get tired when they have to repeat the same task over and over again. Hence, behaviour that is easy to detect but occurs too frequently to justify manual efforts should be labelled automatically.

Of course, the precise working method of a machine may not always have the desired effect. The annotators noticed that sometimes incorrect labels were introduced, too. In particular, when filler events were falsely detected at the beginning and ending of an utterance. We explain this with the fact that some fillers are indeed words like "yeah" and "okay" or at least have a very similar sound (e. g. 'uh-huh" and "hmm"). The classifier learns to label these sounds as fillers if they are surrounded by silence, which is the case at sentence boundaries. Here, often the semantic context (see Section 7.2) is required to decide whether a word is a filler or part of speech. This is a situation where the automated approach is likely to fail. However, even in that case it can still help to speed up the annotation process since it is usually faster to correct a wrongly assigned label than creating it from scratch (in NOVA hotkeys are available for this purpose).

6.5.3 Inexperienced Users

To see how inexperienced users cope with NOVA, we asked students in an introductory lecture on human-computer interaction to solve an annotation task and fill out a questionnaire afterwards. Firstly, the 14 students were divided into four groups (3,3,4,4) and a quick introduction to NOVA was given. Again we stick to recordings from the NOXI database, but simplified the annotation task to two classes: SPEECH and LAUGHTER. We then asked the students to load one of the sessions and create an empty annotation. After annotating few speech and laughter chunks they could use the session completion tool to finish the remaining part of the session. After observing some of the predicted labels they could then decide to either add more manual labels and repeat the completion step, or revise predicted labels with a low confidence. Finally, we provided a manual annotation of the session and asked them to compare it to their own semi-manual annotation.

In the questionnaire we wanted to know what they believe are the strengths and weaknesses of human versus machine coding. We also asked them open questions on how to improve the system. Interestingly, they observed that machine labels were generally more precise, but failed in specific situations e. g. when speech and laughter occurred at the same time. Despite the short time they spent with the tool, they already reported a loss of concentration and noted that this does not apply to machines. Regarding the visual guidance during the revision of a prediction, all groups agreed that highlighting labels with a low confidence helped them correct their annotations. However, we were also interested in their opinion on the visualisation of this information. Currently, labels below an adjustable confidence threshold are superimposed with an uniform pattern. Such a binary decision has the advantage that the user can quickly detect spots that require actions. On the other hand, it is not evident whether a label is weakly or strongly accepted / rejected. While two of the groups liked the binary highlighting, the other two groups preferred a more detailed visualisation, e. g. using a colour gradation. One group also mentioned that probabilities for all classes should be available to get a better understanding why a prediction failed. Further investigations are needed to understand if a more finely graduated representation into several confidence classes is preferable.

Generally, the students reported that they had no difficulties using the interface of NOVA and that the integrated CML tools helped them complete the task in less time.

6.5.4 Generalisability and Adaptation

We also investigated how our approach performs with respect to other modalities than audio. To this end, we applied OpenFace (Baltrusaitis, Robinson, and Morency, 2016) to extract visual features from the videos in the NOXI database (German sessions). The result is a 196 dimensional feature vector per frame (25 Hz), including facial landmarks, action units and gaze directions. Based on these features facial behaviour can be learned. For the following experiment, we defined the task of smile annotation. Hence, an annotation scheme containing a single label SMILE was applied. We employed an experienced annotator who has been working with NOVA before and introduced him to new CML tools. However, this time we did not set a fixed procedure as we were interested in seeing how he applied the tools to solve the problem in an explorative process. Therefore, we asked him to take notes about his experiences.

As we would expect, the annotator started to apply the session completion step after labelling smiles within the first two minutes of a session. He noted that at first the system was not able to reliably predict the smiles for the remaining session. He therefore, corrected another two minutes of the predicted smiles, removed all the predictions beyond that point and applied the completion step once again. This procedure was repeated until the prediction looked stable so that only few smiles with a low confidence had to be revised. Once the first session was finished, he trained a model and applied it to predict the smiles of the second session (session transfer step) and so on. If the prediction of a new session looked reliable, he completed the session by revising labels with a low confidence. However, for some subjects he noted that the prediction was not stable enough (possibly because no subject with similar facial expressions had been seen by the model yet) and so he decided to apply the session completion step instead. In any case, after completing a session he retrained the model including the new labels. Since the robustness of the models improved with each new session which was added to the training set, the predictions got more accurate towards the end of the corpus. This increasingly speeded up the coding process. In total, it took him less than 6 h to finish the 18 sessions (in comparison, manual filler annotation in Section 6.3.3 took more than 14 h).

6.5.5 Finding the Sweet Spot

The previous experiment also shows how it is possible to detect the moment when it is safe to hand the labelling task over to the machine (see the *sweet spot* discussion in Section 6.2.4). We noted that with time, a human annotator learns whether it is worth correcting the predicted annotations or instead adding more labels first before letting the machine complete the session. Especially at earlier stages, sometimes a model trained on few subjects may not perform well enough for unseen users. In the latter case, it may be better to continue using the session completion step. Though, there is no automatic way to predict whether session completion or session transfer should be used, the interface of NOVA allows it to quickly explore both options and pick the more promising approach. Either way, with each completed session the training set incrementally grows, improving the robustness and generalisability of the model.

At some point, when enough sessions are available, the user can apply the following strategy to assess the quality of the prediction. Train a model on a subset of the completed sessions and evaluate it on the remaining ones. The obtained confusion matrix (see Section 6.4) provides feedback about the reliability of the labels. For instance, if a class is often confused with another it may be worth to review all predictions of that class, whereas labels predicted with a high confidence may be safely skipped. Additionally we implemented a "leave-one-session-out" strategy which delivers a cross-validation approach that trains the model on all sessions except one, but evaluates on the last one. Then the step is repeated for all other sessions and the results are averaged. Generally, visually reviewing predictions in NOVA is key to find an optimal work-flow with respect to a specific task.

6.6 Conclusions

The core idea behind *cooperative machine learning* (CML) is to create a loop, in which humans start solving a task (here labelling social signals) and over time a machine learns to automatically complete the task. In conventional approaches, this involves at least two parties: an end-user, who has knowledge about the domain, and a machine learning practitioner, who can cope with the learning system. However, to make the process more rapid and focused, Amershi, Cakmak, et al. (2014) demand that more control should be given to the end-user. To this end, our tool combines a traditional annotation interface with CML functions that can be applied out of the box requiring no knowledge on machine learning. We found it important to give coders the possibility to individually decide when and how to use them in the labelling process. And to assess the reliability of automatic predictions immediate visual feedback is provided, which gives annotators the chance to adapt their strategies at times. By interactively guiding and improving automatic predictions, an efficient integration of human expert knowledge and rapid mechanical computation is achieved. The reported experiments show that even end-users with little or no background in machine learning are able to benefit from the described machine-aided techniques.

The goal of the presented work is to foster the application of *cooperative machine learning* (CML) strategies to speed up annotation of social signals in large multi-modal databases. Well described corpora that are rich of human behaviour are needed in a number of disciplines, such as social signal processing and behavioural psychology (also see Section 3.1). However, populating captured user data with adequate descriptions can be an extremely exhausting and time-consuming task. To this end, we have presented strategies and tools to distribute annotation tasks among multiple human raters (to bundle as much human efforts as possible) and automatically complete unfinished fractions of a database (to reduce human efforts where possible).

In particular, we have proposed a two-fold CML strategy to support the manual coding process (Section 6.2). Applied to a fresh database it first concentrates on completing few individual sessions. A relatively small amount of labels is sufficient to build a session-dependent model, which – though not strong enough to generalise well across the whole database – can be used to derive local predictions. Afterwards, a session-independent classification model is created to finish the remaining parts of the database. During both steps, confidence values are created to guide the inspection of the predictions.

Overall, our experiments demonstrated the potential of the CML approach in reducing human labour during the annotation process. Future work will focus on the question of how to further leverage the complementary skills of human and machines. The employment of the CML approach requires end-users to incrementally inject information into the training process until a desired system behaviour is achieved. Such a workflow necessitates a tight coordination of machine and human tasks. In particular, it would be desirable to provide end-users with guidelines on when to hand over annotation jobs to the machine. We observed that CML strategies not only have the potential to speed up coding, but can also have a positive influence on the annotator's coding style. Because of the preciseness machine-aided techniques introduce into the coding process the level-of-detail is improved while at the same time human efforts are reduced. Here, strategies to guide the attention of the annotator during inspection of the predicted labels become a crucial matter. As mentioned before Rosenthal and Dey (2010) investigated which kind of information should be provided to the user to minimise annotation errors. However, in their studies they concentrate on single images whereas in our case we deal with continuous recordings. To not overload the annotator with too many details we decided to uniformly highlight labels below an adjustable confidence threshold.

EXPLAINABLE AND CONTEXT-SENSITIVE MODELLING OF COMPLEX SOCIAL SIGNALS

7 CONTEXT-SENSITIVE ANALYSIS OF COMPLEX MULTI-MODAL SOCIAL SIGNALS

> *"Reality is not a function of the event as event, but of the relationship of that event to past, and future, events."*
>
> — Robert Penn Warren.

7.1 Motivation

This chapter is based on the publications: Baur, Mehlmann, Damian, Lingenfelser, Wagner, Lugrin, André, and Gebhard, 2015 Baur, Schiller, and André, 2017 Ritschel, Baur, and André, 2017

Previously we introduced our approach towards a more transparent understanding of how machine learning models can be improved with more data and better annotations for recognition problems that aim to map low-level features to abstract classes. In social signal processing this includes, for example, mapping facial landmarks to a specific facial expression, such as a smile, or mapping specific audio features to a voice label, such as "filler". By employing *cooperative machine learning*, we do not only get a more comprehensible concept of the machine's decisions, but we support (non-)experts in the annotation process by speeding up an otherwise tedious task. Once a session is annotated on multiple abstraction layers and for multiple modalities, we now are interested in how such abstractions correlate to each other and how we can combine them in a user model to infer and explain more complex behaviours from our observations. As discussed in Chapter 2, the complex of behaviours needs to be regarded when we want to make inferences about a person's emotions or attitudes. From a technical perspective, this includes the fusion of multiple modalities observed for an individual person, but also considering the interaction dynamics between multiple persons, as well as interaction context information.

Nowadays, machine learning approaches are most often purely data-driven as they use so-called *"black-box"* approaches that map low-level features or decisions of previous classifiers onto abstract emotion labels following statistical methods. Here we usually have no transparent concept of how the model is internally represented, e.g. how and why weights on the nodes of artificial neural networks are related. In most research areas (e.g., in behaviour analysis), the goal of creating a model is to reason about observations in the world, while creating and validating theories that aim to find causation and explanations. Then, such models are often validated in simulations, or collated with real-world observations. That means on the one hand, we have data-driven models in machine learning that do a decent job in creating predictions for a huge amount of recognition problems,

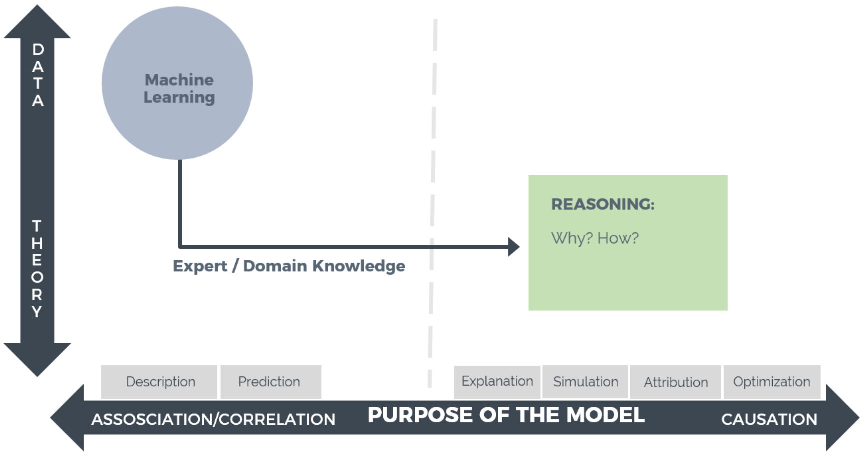


Figure 37.: Data-driven vs theory-based modelling. Machine learning approaches are purely data-driven with the goal to predict outcomes based on learned training samples. In research areas, such as behaviour analysis, most often the goal of a model is to explain correlations and causation based on theories and expert knowledge.

but deliver no transparent way to understand their decisions and not necessarily with a theory behind them. On the other hand, we have models that aim to explain interrelations of observations of the world and/or of their inner states. Such models are also called *"white-box"* approaches. An example for classical *"white-box"* approaches is a simple decision tree. A tree provides a flowchart describing how an observation should be classified by starting at the root of the tree, and following the branches until the final leaf determines the classification we predict.

The graph in Figure 37 distinguishes between the source of a model (data vs. theory driven, *y axis*), as well as between the model's purpose which goes from finding correlations and predictions, to finding causation (*x axis*). Supervised machine learning models therefore can be found on the upper left quadrant while explanatory models, such as decision trees are found on the lower right corner quadrant. In this work, we aim to combine machine learning to find correlations between abstractions of behaviours with transparent theory-based models that represents and explain real-world interrelationships.

The combination of argumentation-based models of evidence and probabilities has aroused interest in recent years (Keppens, 2012; Verheij, 2014; Vlek et al., 2015). Often, argumentation is used to analyse probabilistic reasoning. Argumentation theory describes how conclusions can be justified using models. Such models "closely follow the reasoning patterns present in human reasoning, which makes argumentation an intuitive and versatile model for common sense reasoning tasks." (Timmer et al., 2017). Argumentative reasoning in Bayesian networks proves to be of use for the interpretation of probabilistic reasoning. Explanation methods exist for a diagnostic analysis of the model itself, (e.g. in the work of Lacave, Luque, and Diez (2007) or Koiter (2006)), as well as for evidences in the network. The aim of our framework is to find relevant abstractions of behaviours that indicate a complex problem (diagnostic analysis), based on an expert-knowledge representation of the real-world (explanation of evidences). In Chapter 2 we presented various theories about the correlation of social signals and emotions or social attitudes. In order to incorporate such theories in a computational model, but further, to validate a model's relevance for a specific corpus, in this chapter we propose a hybrid approach that allows theory based modelling in combination with data-driven parameter learning.

We see the main contributions presented in this chapter as follows:

- In Section 7.2 we give an overview on various kinds of context information that we suggest should be considered when making assumptions about complex social signals.

- In Section 7.3 we present our approach to model complex social signals with the help of dynamic Bayesian networks. We incorporate human expert knowledge in the structure of the models, yet learn parameters and correlations based on concurrent observations of social signals on multiple abstraction levels and context information.

- In Section 7.4 we present a walk-through to demonstrate how the approach is applied on the use-case of inferring "conversational engagement". We introduce the annotation scheme and network structure and compare the approach with state-of-the art black box approaches. Further, we investigate how we can extend a model to a multi-person model that additionally considers interaction dynamics between multiple participants.

- In Section 7.5 we introduce a second use case to illustrate the generalisability of the approach. More precisely, we employ a transparent model for inferring emotion regulation strategies in the context of virtual job interviews. In order to provide feedback to users of such systems, it is not only important to identify critical situations, but also to provide comprehensible explanations of why the system identified a situation as such.

7.2 The Role of Context

> *"People also smile when they are miserable."*
>
> — Paul Ekman

This section is based on the publication: Baur, Schiller, and André, 2017

So far we have considered techniques for recognising human behaviours in each modality with only little attention to *context* (e.g. context that is represented by surrounding frames when training a model). Yet there are behaviours that are difficult to analyse and interpret correctly without further information about the *context* of a situation. *Context* is a wide-ranging term that has different meanings depending on the paradigm of research, application and scenario. Duranti and Goodwin (1992) noted that it seems impossible to present a single, precise and technical definition of context. Context information might appear as a single impact factor on the interaction or as a combination of multiple types of information. In this section we approach different aspects of context:

- *Temporal context:* In classical linguistics, context is "a frame that surrounds the event and provides resources for its appropriate interpretation" (Duranti and Goodwin, 1992). Wöllmer, Metallinou, et al. (2010) considered context as the temporal surroundings of an observation. In their work they successfully applied bidirectional long-short-term memory (BLSTM) neural networks to consider contextual long-range observations for the prediction of emotions. They further investigated algorithms such as multidimensional dynamic time wrapping (DTW) and asynchronous hidden-markov models to fuse mutual information from multiple modalities, while considering their temporal alignment (Wöllmer, Al-Hames, et al., 2009). An overview on algorithmic approaches, such as dynamic and canonical time wrapping in the context of facial expression analysis is given in (Panagakis, Rudovic, and Pantic, 2018).

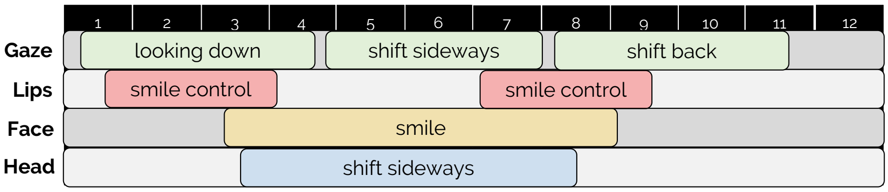


Figure 38.: A typical time series of social cues that are performed when a person is feeling "embarrassed"

When analysing complex social signals and emotions, the temporal order of behaviours is of vast importance. As an example, Keltner (1995) describes a typical time series of behaviours in multiple modalities, that represent a typical instance for the complex emotion "embarrassment" (see Figure 38). Typically, the gaze shifts towards the bottom, the lips make slight movements that often turn into a smile followed by the gaze and head shifting to the side and back. Considering such sequences of social signals adds valuable information to the interpretation, compared to the analysis of isolated single cues.

- *Interaction dynamics context:* Analysing the dynamics in human communication includes being able to investigate both, the individual multi-modal dynamics (see temporal context) as well as the interpersonal dynamics (see Section 2.3). Researchers consider interpersonal dynamics on multiple abstractions. For example, Delaherche et al. (2012) and Varni et al. (2015) consider the synchronicity of people in dyadic interactions on a signal level. Therefore, they developed a set of synchronicity measurements. Rich, Ponsleur, et al. (2010) defined state machines to automatically recognise the four interpersonal cues "mutual gaze", "directed gaze", "adjacency pairs" and "backchannels" (see Section 2.3.2.2). In their work they counted the appearance of such bi-directional cues and considered their appearance as an indicator of a person's engagement. Another aspect is the current role in a conversation. Depending on whether the user is in the role of a listener or a speaker, the same kind of behaviour might be interpreted in a completely different way. The influence of the interaction role is illustrated by the following example. Let us assume we observe a person showing a high amount of gestural activity. If the person is in the role of a listener, the observed activity could be interpreted as restlessness. On the opposite, if the person is in the role of a speaker, we might conclude that the person is actively engaged in the conversation. Salam and Chetouani (2015) classify multiple aspects of context as parts of the relationship of a social robot and a human during an interaction. More precisely, the interaction context in their definition describes how a scenario relates multiple interlocutors.

- *Discourse and domain context:* In order to improve the interpretation of social cues, the situation in which they are displayed should be taken into account. In human-agent interactions, such a situation might be triggered by the agent. For example, if a job applicant reacts to a difficult question with a laughter, it is unlikely that he or she is happy about the question. Rather, the laughter portrays embarrassment. Based on the social cues alone, it is almost impossible to distinguish between different forms of laughter. Morency (2010) points out that for humans, "knowledge about the current topic and expectations from previous utterances help guide recognition of non-verbal cues". In their work, they consider dialogue information from the interaction with a robot to disambiguate individual behaviour of a human user. Other work, e.g. Gatica-Perez (2009), defined different contexts related to actions that happen in group meetings in the working environment (e.g. seated discussions vs. white board presentations). An extensive overview on approaches towards domain adaptation is given in (Patel et al., 2015).

- *Semantic context:* The interpretation of detected social cues can be entirely altered through the semantics of accompanying verbal utterances. For example, a laughter in combination with an utterance commenting a negative event would no longer be interpreted as a sign of happiness, but rather be taken as sarcasm. By considering the semantics of accompanying spoken content, detected social cues could be interpreted more accurately. Studies further indicate that humans use semantic context for the interpretation of facial expressions (Bruce and Young, 1998; Ratner, 1989; Wallbott, 1988).

- *Environmental context:* The location and environmental surroundings may also influence the way we behave during an interactions. As an example, Zimmermann (1996) argues that the environmental surroundings directly influence our behaviours e.g. in the way we breathe or speak. In human-computer interaction and especially in ubiquitous computing, a system is called context-aware when it understands the circumstances and conditions surrounding the user. Abowd et al. (1999), define context as "any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves". They further state that context is highly dependable on the current perspective.

- *Social context:* Another aspect of context is the so called "social context". Riek and Robinson (2011) stress the importance of considering social context when creating automated behaviour analysis systems. In their definition, social context is the "environment where a particular person is situated with four factors that may influence (their) behaviour: situational context, cultural context, the person's social role context, and the environmental social norms". Such aspects may be addressed by the following questions: In what kind of situation does the conversation happen? What is the setting of the interaction? (situational context), How well do the interlocutors know each other? Do they share common knowledge? What culture or gender do they have? What is their personality like? (cultural context). How is their relationship? How is their social status? (the person's social role). What are the social norms in the location of the interaction? What are the social norms in the community of the interlocutors? (environmental social norms). Questions like these play an important role, especially when interpreting non-verbal behaviour. Some of these aspects might be difficult to retrieve in an automated manner during the interaction between multiple interlocutors. However, if it is not possible to automatically gather such context information, it could be collected upfront.

When humans interpret behaviours of other people, they consciously or unconsciously include these and similar considerations in their reasoning process. Machines that aim to correctly interpret human behaviours should consider contextual aspects in their interpretation models. Yet, besides temporal context (e.g. Wöllmer, Metallinou, et al., 2010), only little attention has been put to contextual aspects in current social signal processing research.

7.3 Modelling Approach

In order to infer complex social signals with a transparent user model, we consider dynamic Bayesian networks (DBN) (Murphy and Russell, 2002) as the modelling approach in our conceptual and technical framework. DBNs are probabilistic models that allow expressing causal relationships between nodes in a network, while at the same time considering previous observations. Even tough the parameters for such nodes and even the overall network structure may be learned with machine learning techniques, DBNs allow retracing the decisions they are making for each node or layer of nodes visually. We could think about using alternative models, such as deep end-to-end learning with artificial neural networks. While such approaches deliver promising results on audio-visual data, they only give little insight on *how* and *why* they predict behaviours the way they do (this currently is an evolving research direction in the area of explainable AI). Especially in scenarios where it is essential to know why a persons behaviour is interpreted as, e.g., "strongly disengaged", the idea is often to identify cues that led to this interpretation, providing an additional abstraction layer. While the structure of a DBN may be modelled based on a theory, our framework allows to access the annotation database introduced in Section 5.3.4 to provide a DBN with parallel observations, so it can learn correlations between concurrent behaviours, context and the complex phenomena of interest. In our framework, DBNs may also be applied in a real-time environment by updating evidences with *"events"* detected with the SSI framework (see Section 4.2). Such events might represent either specific interesting changes in our signal, e.g. a gesture starting or ending, or classification results. To this end, the previously introduced NOVA tool allows to export annotations of concurrent behaviours from the annotation database described in Section 5.3.4, so that the network will learn temporal correlations between multiple events. Additionally, the NOVA tool allows to replace events with annotations to simulate the inferences of a DBN in an offline simulation for faster predictions.

7.3.1 Related Work

Bayesian networks have been successfully applied in earlier work in the area of high-level interpretation of social signals. One of the pioneer studies is the work by Conati and Maclaren (2009). They have incorporated bio-feedback sensors into a complex emotion model, that was based on a subset of the emotions proposed by OCC theory (see Section 2.4.2.3). They employed a dynamic decision network (a generalisation of a dynamic Bayesian network) to capture many of the complex phenomena associated with appraisal theories. In particular, their model estimated student goals based on personality traits and events which represent changes in the environment (e.g., progress in the system) as well as evidence from physical feedback channels to support the model's prediction.

Bosma and André (2004) used Bayesian networks for the prediction of user intentions in unclear dialogue acts. Thereby, they considered emotional states that are derived from physiological sensors to infer the intention. An initial user study suggested that physiological evidence of emotions could be used to disambiguate dialogue acts.

Sabourin, Mott, and Lester (2011) focused, similar to Conati et al., on learners' emotions, and employed multiple variations of Bayesian networks. More specifically, they investigated the benefits of using cognitive models of learner emotions, to guide the development of Bayesian networks for prediction of student affect. Predictive models were empirically trained on data, acquired from 260 students interacting with a game-based learning environment. As a dynamic Bayesian network turned out to be the most successful model, they emphasised the importance of temporal information in predicting learner emotions. They concluded that predictive models may be used to validate theoretical models of emotion.

deRosis et al. (2011) employed dynamic Bayesian networks to investigate the relation between cognitive representations and processes. In their example of "Fear", when a child is learning how to ride a bike, they illustrated how various events may have theoretical positive or negative influence on the emotion in an expectation-based approach.

Wöllmer, Schuller, et al. (2010) combined a hierarchical dynamic Bayesian network to detect linguistic keyword features together with long short-term memory (LSTM) neural networks (see Section 4.1.3.3) which model phoneme context and emotional history to predict the affective state of the user. This way, they are combining acoustic, linguistic, and long-term context information to continuously predict the current valence and activation in a two-dimensional emotion space.

Lugrin, Frommel, and André (2018) used Bayesian networks to incorporate culture into intelligent systems by combining theory-based and data-driven approaches. Their network aims to generate nonverbal culture-dependent behaviours. While the model is structured based on cultural theories and theoretical knowledge of their influence on prototypical behaviour, the parameters of the model are learned from a multi-modal corpus recorded in the German and Japanese cultures. In their work, they aim to generate adequate behaviours for an agent to show, based on its simulated culture.

Finally, one could conclude that (dynamic) Bayesian networks have been successfully employed for some predefined contexts and applications. Especially when considering context, as it is essential in e.g. appraisal emotion models, or in specific applications, DBNs turn out to be a promising approach. In contrast to most other fusion mechanisms their structure may be actively modelled, based on existing theories, so that the structure contains valuable information implicitly, allowing to include existing knowledge in the model. This is especially useful when it is required to make assumptions why the model predicted one outcome and not another. It is worth mentioning that context information has only rarely been taken into account - or in most cases, limited to aspects like temporal context in previous research. Yet, in human communication multiple aspects of context (Section 7.2) continuously influence our behaviours. In this work, we aim to consider additional contextual aspects, like the topic of the interaction, the interpersonal dynamics and social background.

7.3.2 Bayesian Network Theory

In this section a brief introduction in the theory behind Bayesian networks (BN) is given, to provide a better understanding of how BNs work and how we use them to infer complex social signals in later sections. For a more detailed overview on advanced concepts of Bayesian networks, the work by Pearl (1985), who originally introduced Bayesian networks, is recommended.

Generally speaking, a Bayesian network (or belief network) is a graphical model for expressing probabilistic relationships among a set of variables, as well as their conditional dependencies. BNs are named after Thomas Bayes' rule for updating probabilities, based on new evidence.

Bayes' theorem describes conditional and unconditional probabilities of two events "A" and "B":

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}$$

The formula can be interpreted as follows (Pearl, 1985):

- P(A) is called the "a priori" (Fisher, 1936) probability (or "unconditional", or prior probability) of A. It is *prior*, as it takes no information about B into account. However, event B does not necessarily have to occur after event A.

- P(A | B) represents the conditional probability of event A, given event B. Another name is the "posterior" probability as it is derived from or depends upon event B.

- P(B | A) represents the conditional probability or likelihood of B given A.

- P(B) represents the "prior" probability of B and is a normalising constant in the formula.

- $\frac{P(B|A)}{P(B)}$ is named "Bayes factor" or likelihood ratio.

Bayes' theorem mathematically represents the relationship of the conditional probability of an event (A), given another event (B), and at the same time, the conditional probability of event B given event A. Initially, Bayesian networks were developed to model inference by both, bottom-up (perceptual) and top-down (semantic) combinations of evidence. BNs are, as of today, considered state-of-the-art, for reasoning with uncertainties in expert systems and artificial intelligence applications. They are, compared to e.g. artificial neural networks, capable of bi-directional inferences.

A Bayesian network is a compact representation of the so called *joint probability distribution* (JPD). This allows us to compute the posterior probabilities of any subset of variables when evidence about any other subset of variables is provided.

Therefore, a model always has to represent the *joint probability distribution*, which means that for every possible event that may appear in our network, the combination of all possible values of any variable must be represented. Bayesian networks achieve compactness by resolving the *joint probability distribution* into *local conditional distributions* (LCD) for each variable given its parents. The number of parameters in a network grows linearly when additional nodes are added, but given a discrete representation of the *conditional probability distribution* (CPD) with a *conditional probability table* (CPT), the size of a LCD grows exponentially by the number of parent nodes. Often the CPTs are further decomposed to allow fastening the network's inferences (Zhang and Poole, 1999; Butz, 2002; Poole and Zhang, 2011). One example is the so called context-specific independence that is given when a target node is "independent of certain parents, given that other parents are assigned specific values" (Yap, Tan, and Pang, 2008)

In the definition of Boutilier et al. (1996) context-specific independence is defined as follows: Let the sets of variables X, Y, Z and C be pairwise disjoint. Sets X and Z are context-specifically independent, or contextually- independent, given Y and $c \in val(C)$ (the context), if the conditional probability P(X | Y, c, Z) = P(X | Y, c) whenever $P(Y, c, Z)$ >0.

A Bayesian network consists of a directed acyclic graph (DAG) to indicate dependencies within the structure, and local probability distributions that specify probabilistic relationships. The DAG of a Bayesian network is expressed with nodes and directed links between them:

- A *node* is a representation of a variable of interest (e.g., a feature of a movement, the general occurrence of an event or the discourse of the interaction). Bayesian networks are in general able to handle both discrete and continuous values. Most implementations support discrete or discretised values as inputs for nodes. The nodes hold attributes that represent categories or ranges of continuous values.

- A *directed link* shows either statistical (purely informational) or causal dependencies between variables. The directions of the nodes define affinity relations, e.g., parent-child relationships. In a Bayesian network, a link from A to B visualises that A is the parent node of B and therefore B is A's child node. A node without parents is also called a *root node*. For root nodes, the *local probability distributions* (LPD) are unconditional, whereas for nodes with parents, the LPDs are conditional. In this case, *conditional probability tables* (CPTs) quantify the dependencies for each node, given its parents, and therefore, all its ancestors in the graph.

A relation of nodes in a Bayesian network can either be causal or non-causal. As the name implies, in non-causal Bayesian networks, no causal assumptions are made. That means, the structure does not represent any knowledge about the causal order between nodes and variables. Therefore, interpreting a non-causal BN should only be statistical and for informational purposes. An example is a simple network containing two nodes. The child node that represents the "shoe size" and its parent node representing "hair colour". In a *conditional probability table* (CPT) the conditional probabilities of the "shoe size", given by the parent node "hair colour" are provided. As mentioned before such networks only represent statistical values and do not infer any further information. Non-causal networks are still of great value for analysing behaviours in an interaction. For example, given a context parent node "question about weaknesses" and a child node "smile", one can see a direct relationship of the social cue smile in the given discourse context. The structure of a model is also non-causal when it is learned using optimisation algorithms. These algorithms aim to optimise the recognition of the target node, while as a trade-off ignore causal relationships. Compared to statistical relationships, the diagram in Figure 39 describes the causal relationships among five nodes in a simplified Bayesian network.

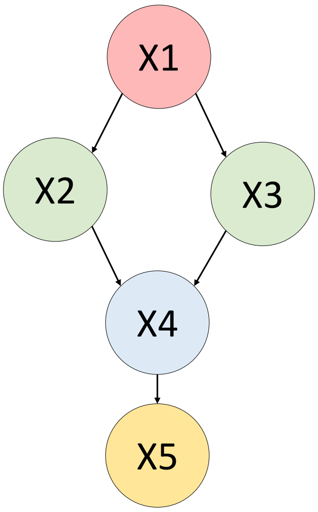


Figure 39.: A simple causal Bayesian network

We'll imagine a simplified scenario to explain the causal relationship between the five nodes representing:

- X1: the candidate is in a job interview situation
- X2: the interviewer makes a compliment
- X3: the interviewer asks a harsh question
- X4: the candidate's emotional state (happy, neutral or embarrassed)
- X5: the candidate is smiling

The absence of a direct link between being in the situation of a job interview (X1) and the candidate smiling (X5) captures our common understanding that no direct influence of the person being in a job interview and her/his smiling behaviour exists. The influence on a person smiling or not is mediated by the emotional state (being neutral, happy or embarrassed) (X4) .

Causal Bayesian networks act as direct representations of the world, rather than of reasoning processes. The direction of arrows in the DAG represents causal connections during reasoning instead of information flow (which is the case in rule-based systems or artificial neural networks). Information is propagated in any direction during the reasoning process in Bayesian networks.

To explain this, we consider our example from various perspectives:

- Given the interviewer makes a compliment (X2), there is a high chance the person is possibly happy (X4) (*prediction, simulation*).

- If we observe the person smiling (X5), this will give us evidence that the person is probably either happy or embarrassed, but not in a neutral state any more. (X4) (*diagnosis, abduction*, or *reasoning to a probable cause*).

- If we instead observe that the person is embarrassed (X4), it is more likely that the interviewer asked a harsh question (*abduction*) (X3).

- If we observe that the interviewer made a compliment (X2), this will drastically reduce the likelihood she or he also asked a harsh question in the same turn (X3) (*argumentation, explaining away*). This step is hard to model in rule-based systems and artificial neural networks, as information needs to be propagated in two directions.

By extending a Bayesian network with temporal links we speak of dynamic Bayesian networks. They additionally consider temporal relations in the inference process. That way, time series or sequences may be incorporated in a model. For example, if we want to analyse a series of events related to the mirroring of smiles, we consider the temporal alignment of smile cues of multiple interlocutors, to find correlations between their behaviours.

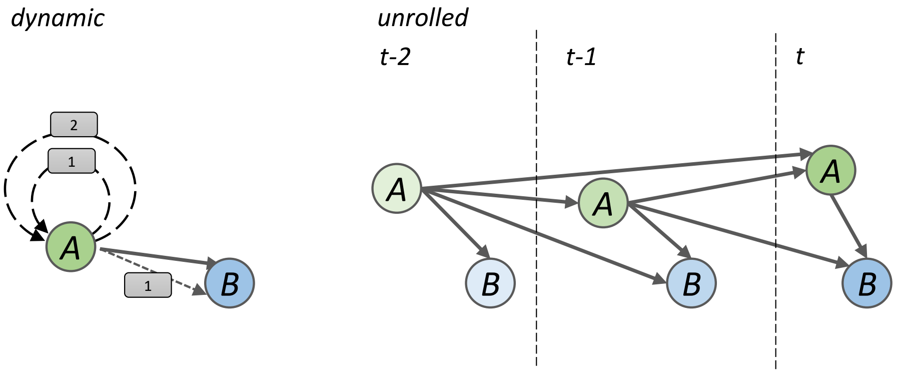


Figure 40.: A dynamic Bayesian network (left) can be unrolled to a static Bayesian network (right)

DBNs may be used to model complex multivariate time series and different regimes of behaviours, since time series often behave differently in different contexts. For each time step, a copy of the state variable is kept. State variables X_t and X_{t-1} describe the state of the world at times t and t-1 respectively. Figure 40 illustrates a simple dynamic Bayesian network. The temporal nodes illustrate that node

A has a relationship to itself in timesteps t-1 and t-2. Further, A, and A in timetep t-1 are parents of B. A dynamic Bayesian network can be unrolled to a static Bayesian network, as illustrated on the right.

Each set of evidences E_t represents the observations that are performed at time t. The *"sensor model"* $P(E_t| X_t)$ is encoded in the conditional probability distribution for the observable variables, given the state variables. The *"transition model"* $P(X_t| X_{t-1})$ relates the state at time t-1 to the state at time t. To keep track of the world state, the actual probability distribution needs to be calculated given all previous observations

7.3.3 Intermediate Summary

In our framework we combine concurrent annotations of social cues and context information to learn parameters of dynamic Bayesian networks. The structure of a network is actively modelled by a researcher or expert based on theories (as described in Chapter 2). In a real-time environment we update the DBNs with evidences based on either meaningful features or classification results. To achieve this, we apply the event mechanisms of the SSI framework (see Section 4.2). Alternatively, using the NOVA tool (see Section 5.3), we can run our model in an offline simulation, where it is updated with parallel annotations of single social cues and context to infer complex behaviours.

In the next sections we introduce two concrete use-cases where the proposed tool chain has been applied in. The goal is to illustrate a walk-through for the transparent and context-sensitive recognition of complex social signals using our approach. We exemplify this in more detail with the use-case of inferring a person's conversational engagement in the context of an information-retrieval agent system. Here, transparency is especially useful for developers of complex interaction systems and social scientists who analyse human-human and human-agent interactions. Secondly we discuss how our approach is used to recognise emotion regulation strategies in the context of virtual job interviews. Often, the goal of such systems is to provide feedback to end-users about their behaviours (e.g. on why the system found the user's engagement low in a certain situation). Therefore, using transparent approaches enables agents to provide explanations on why they made certain statements. This way, a user is enabled to comprehend and therefore actually learn from the feedback.

7.4 Exemplary Case I: Conversational Engagement

A first use-case we want to consider is the automated recognition of *conversational engagement* (as discussed in Section 2.5.2). The engagement a person is showing in an interaction is a complex social attitude that, to a wide extent, depends on the context of the situation. When we want to apply a model of engagement in an interaction system with a virtual agent or robot, we are further interested in explaining why the system predicted a person to be highly engaged - or not engaged at all. For learning the parameters of our first use-case model that aims to recognise the engagement of a person in a dyadic interaction, we rely on the NOXI corpus that has been introduced in Section 3.3. One objective of the NOXI corpus was to provide data to support research in *embodied conversational agents* (ECAs). More precisely, a main goal was to advance virtual agents with the capacity of interpreting a user's non-verbal behaviour in real time and to associate it with different engagement variations. The proposed model was developed for interactions with the ARIA-Valuspa platform (Valstar, Baur, et al., 2016). We'll first introduce the ARIA-Valsupa platform to give an impression of the overall scenario our model is applied in.

7.4.1 Scenario: The Aria-Valuspa Platform

The ARIA Valuspa platform (AVP) is a modular software platform for the creation of embodied conversational agents. It is an architecture of interconnected modules that update a virtual agent's state and generate the most relevant behaviour when interacting with a user. The high-level architecture consists of three mayor blocks: an input or *behaviour sensing* block, an *agent core* block, and a *behaviour generation* block. Blocks consist of multiple modules. For example, the behaviour sensing block consists of a automatic speech recognition module (ASR), a visual analysis module (eMax), and the audio-based paralinguistic analysis module (OpenSmile), which are wrapped in the SSI framework (see Section 4.2). Based on the multiple inputs of the behaviour sensing block, a dynamic Bayesian network is updated to infer the engagement of the user during the interaction with the agent.

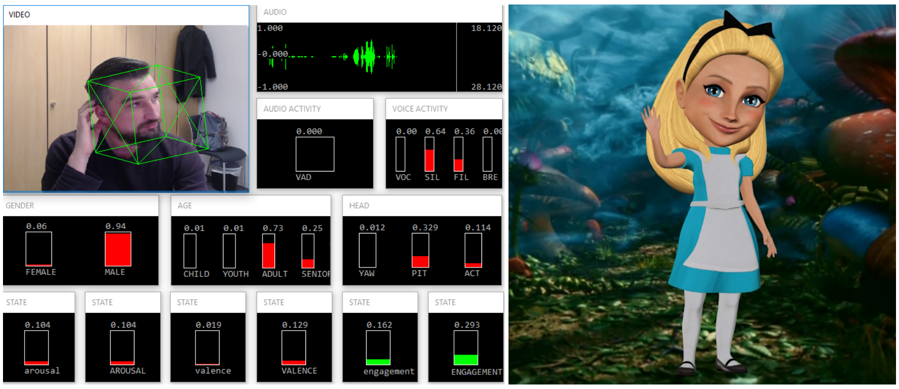


Figure 41.: The user's social signals are measured during the interaction with the virtual character Alice (right). The user's conversational engagement (green) is inferred based on observations from various modalities using a DBN.

In order to demonstrate how the ARIA framework can be used to address the reality of the user's needs, the project has delivered a number of interaction scenarios. One scenario contains a smart and interactive book reader, which we call Book-ARIA. During the project other use cases, such as, an industry one which is backed by UNILEVER and its brand PERSIL, were developed. The Book-ARIA functions as a showcase of the rich characters that can be generated with the ARIA-VALUSPA platform and how they function as interfaces for information retrieval for more complex tasks, such as, questions about the novel's content, characters, the author, etc. For the purpose of this project, the novel *Alice in Wonderland* by Lewis Caroll has been selected as an illustrative example.

The animated virtual agent "Alice" has been created by one of the project's industry partners "Cantoche". A typical interaction of a user with Alice is shown in Figure 41.

A main purpose of the Book-ARIA Scenario was to keep the user engaged in a conversation with the agent. From a dialogue management perspective, this means to firstly be aware of the user's engagement and to react adequately towards it. With the ARIA-Valuspa platform, we will demonstrate our approach to model the complex social signal "conversational engagement". We will also consider multiple aspects of context, as described in Section 7.2, like social and temporal context. In Section 7.4.5 we will complement the approach with a multi-person model that additionally considers the interaction dynamics.

7.4.2 Annotation and Network Structure

For the modelling of conversational engagement we first developed an annotation scheme that serves as a training-set for our Bayesian network. For the annotation we used the NOVA tool (as described in Chapter 5). An overview on the annotation scheme can be seen in Table 8.

Table 8.: Annotation scheme for the multi-modal behaviours and engagement annotations in NOXI.

Tier (modality)	Labels/Scores
Audio	voice activity, fillers, arousal (0..1)
head activity	nod and shake, general movements
head direction	frontal, sideways
facial expression	valence (0..1)
gaze direction	towards-interlocutor, up, down, sideways, other
gestures	continuous movement , energy of hands (0..1)
poses	arms crossed, head touches, distance to head, openness hands behind back
context	unexpected event, gender, role
engagement	strongly disengaged - strongly engaged (0-1)

- **Engagement** We defined a continuous scheme ranging from 0 to 1, meaning 0.5 represents a medium level of engagement (e.g., behaving neutral) while 0 means strongly disengaged and 1 means strongly engaged. Annotators are instructed to observe the "value that a participant in an interaction attributes to the goal of being together with the other participant(s) and to continue the interaction", following Poggis definition (Poggi, 2007). This definition has been used in several works (Castellano et al., 2009; Peters et al., 2005; Glas and Pelachaud, 2014; Sanghvi et al., 2011). It is to mention that human annotators might have different understandings of the observed behaviours.

- **Facial behaviour** We use a model trained on eMax (Valstar, Martinez, et al., 2010) features to automatically detect smiles and eyebrow movements (represented in a valence value), as well as changes in gaze direction and head activity.

- **Kinesics.** Several gestures and postures are automatically annotated based on existing recognisers. To this end, we used the skeleton tracked by the Microsoft Kinect 2 sensor as input for several recognition modules in the SSI framework. In order to extract poses like arms crossed, hands together and lean rotation we applied the FUBI (Kistler et al., 2012) plugin. For extracting information about the expressiveness, like the amount of movements, the energy and fluidity, we employed the body properties plugin (see Appendix A).

- **Paralinguistics** At first, we added the model introduced in Section 6.3.2 that decide if a person is speaking, to determine who has the turn in the conversation. Additionally the model includes filler sounds, which we consider important as they indicate whether a person is making feedback sounds like "uh-hu". We further used models to generate continuous outputs for a subject's arousal, trained on the GEMAPS (Eyben, Scherer, et al., 2016) audio feature set.

- **Context** To consider behaviours in their context, we manually labelled various context information, such as the appearance of unexpected events (in NOXI that were calls, SMS or walk-ins), but also the gender and the role (expert or novice) of the participant. Further, we automatically labelled the voice activity of each person to determine the speaker and the listener. An addition would have been to annotate cultural information as well, which we decided not to focus on. Yet we see a huge potential in this sort of context information.

The annotations are directly mapped onto the structure of a Bayesian network. In our framework, a tier represents a node and the labels represent the attributes of a node. A big advantage of Bayesian networks is that the structure has intrinsic meaning compared to other models (e.g. artificial neural networks). This way, expert knowledge may be used to model the structure based on a theory. The structure of BNs may also be learned, using various brute-force algorithms to find the "best" structure in terms of recognition of a target node. In that case, causal relations between nodes are no longer given, but arrows represent a simple relationship between nodes. The goal in this approach is to incorporate human expert and domain knowledge in the process to validate existing or new theories, we presume a model is created based on theories and domain knowledge. Figure 42 shows a simplified Bayesian network, meant to model our interpretation of *conversational engagement*. Context nodes such as the gender or role are conditional nodes, so that engagement is predicted "given" the context information, while social cues are symptoms shown by the observed person.

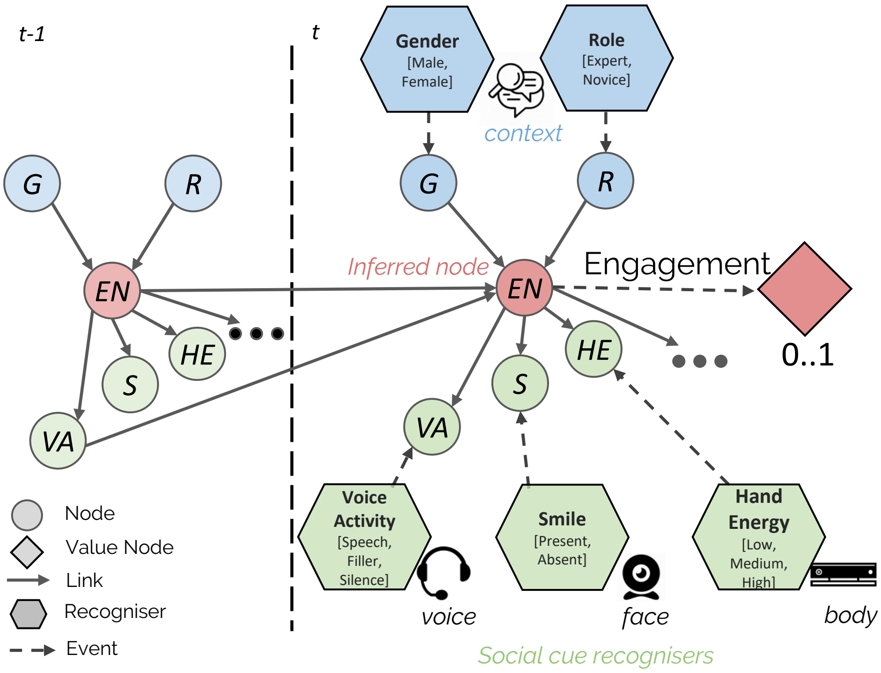


Figure 42.: Schematic of a simplified dynamic Bayesian network. The network is illustrated for the current timestep t and the previous timestep t-1. The DBN is updated with events from multiple sources in timestep t. For example, the voice channel is used to provide classification results for the recognition of speech, fillers and silence. It further determines the gender of a person (which is considered a contextual node). Additionally, smiles are recognised from video, and body activity from a Kinect Sensor is discretised based on a simple threshold mechanism. The output of the final engagement node is mapped to a continuous value between 0 and 1 using a value node.

By applying a diagnostic analysis, the influence of single nodes on their parent nodes can be analysed visually using, for example, the approach by Koiter (2006). This way, we can directly observe within our network if – for the data the network was trained on – causal influence correlates with our theory. Observations with none or only little influence on our target node can then be removed from the model (or be evaluated on additional datasets for further investigation). As an interesting addition, highly correlating features found with this bi-directional influence analysis might be used as input for less transparent recognition models, such as support vector machines or artificial neural networks, to guide the training process.

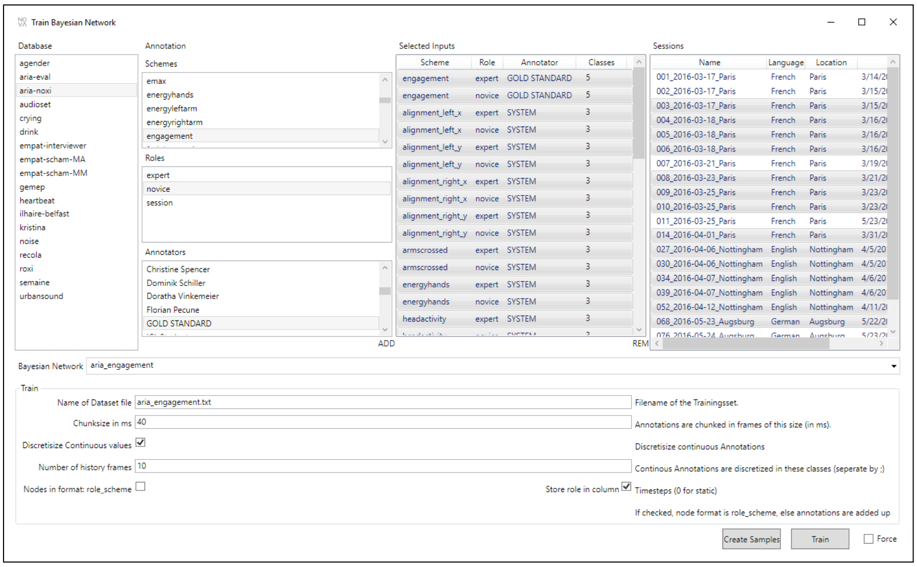
For our use-case Bayesian network we modelled the *engagement* node with five possible attributes, representing a discretisation of the engagement values ranging from very low to very high. A value of 0.5 represents medium or neutral engagement. The value node maps the probabilities back to a continuous score between 0 and 1 by weighting each attribute of the node with a factor. In our case, a high probability for very high engagement would map to a value close to 1 while a high probability for very low engagement would map to 0. We calculate the final output of the engagement_value node with the following formula: 0 * P(verylow) + 0.25 * P(low) + 0.5 * P(medium) + 0.75 * P(high) + 1,0 * P(veryhigh).

In the simplified network in Figure 42, two context nodes represent the parents of our engagement node. These are the interaction roles (expert or novice) and the gender.

The probabilities for the other nodes are updated by classification outputs of support vector machines, artificial neural networks, or other sources (e.g. state machines, or meaningful features, such as the energy of both hands passing a certain activity threshold). For example, the probability that the node *Voice Activity* has the value *FILLER* is high, if the corresponding social cue "filler" has been detected with high confidence. The model then might be applied in either, an offline simulation based on annotations, or an online scenario where evidences are updated based on observed events within the SSI framework. A light-version of the SSI online recognition pipeline can be found in Appendix C.

7.4.3 Learning Parameters with NOVA

The NOVA tool, introduced in Chapter 5 allows parameter learning in a (dynamic) Bayesian network to fuse multiple observations in a prediction model. Annotations from various annotators may be combined as training data. In Figure 43 "engagement" annotations are selected from the *gold standard user*. The *gold standard* annotations are created by combining annotations from multiple raters to gain the ground truth (see Section 5.3.5). As this will be the node of interest, no evidence will be delivered during runtime, but rather the degree of engagement will be inferred from other observations. Annotations, such as the arms-openness, facial expressions or the amount of hand movement have been automatically created using existing machine learning classifiers or represent continuous values of meaningful features. Finally, annotations from human annotators could be added to the model. In this case, speech/filler/breath/silence annotations have been added from a human annotator, which have been created using the cooperative machine learning techniques, as described in the example in Chapter 6. On the right, all available sessions, that contain annotations from all selected scheme/annotator combinations are listed.


Train Bayesian Network

Database
agender
aria-eval
aria-noxi
audioset
crying
drink
empat-interviewer
empat-scham-MA
empat-scham-MM
gemep
heartbeat
ilhaire-belfast
kristina
noise
recola
roxi
semaine
urbansound

Annotation

Schemes
emax
energyhands
energyleftarm
energyrightarm
engagement

Roles
expert
novice
session

Annotators
Christine Spencer
Dominik Schiller
Doratha Vinkemeier
Florian Pecune
GOLD STANDARD

Selected Inputs

Scheme	Role	Annotator	Classes
engagement	expert	GOLD STANDARD	5
engagement	novice	GOLD STANDARD	5
alignment_left_x	expert	SYSTEM	3
alignment_left_x	novice	SYSTEM	3
alignment_left_y	expert	SYSTEM	3
alignment_left_y	novice	SYSTEM	3
alignment_right_x	expert	SYSTEM	3
alignment_right_x	novice	SYSTEM	3
alignment_right_y	expert	SYSTEM	3
alignment_right_y	novice	SYSTEM	3
armscrossed	expert	SYSTEM	3
armscrossed	novice	SYSTEM	3
energyhands	expert	SYSTEM	3
energyhands	novice	SYSTEM	3
headactivity	expert	SYSTEM	3

Sessions

Name	Language	Location	
001_2016-03-17_Paris	French	Paris	3/14/2
002_2016-03-17_Paris	French	Paris	3/15/2
003_2016-03-17_Paris	French	Paris	3/15/2
004_2016-03-18_Paris	French	Paris	3/16/2
005_2016-03-18_Paris	French	Paris	3/16/2
006_2016-03-18_Paris	French	Paris	3/16/2
007_2016-03-21_Paris	French	Paris	3/19/2
008_2016-03-23_Paris	French	Paris	3/21/2
009_2016-03-25_Paris	French	Paris	3/23/2
010_2016-03-25_Paris	French	Paris	3/23/2
011_2016-03-25_Paris	French	Paris	5/23/2
014_2016-04-01_Paris	French	Paris	3/31/2
027_2016-04-06_Nottingham	English	Nottingham	4/5/20
030_2016-04-06_Nottingham	English	Nottingham	4/5/20
034_2016-04-07_Nottingham	English	Nottingham	4/6/20
039_2016-04-07_Nottingham	English	Nottingham	4/6/20
052_2016-04-12_Nottingham	English	Nottingham	4/11/2
068_2016-05-23_Augsburg	German	Augsburg	5/22/2

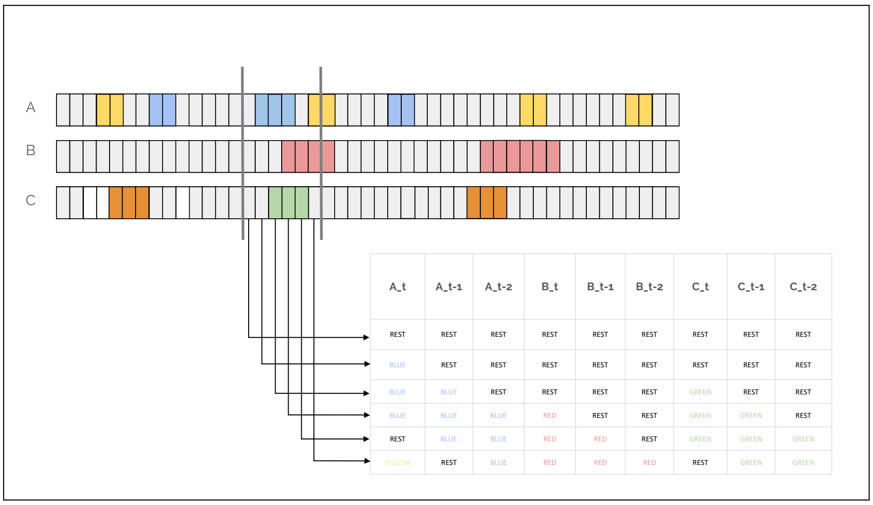
Bayesian Network aria_engagement

Train
Name of Dataset file aria_engagement.txt — Filename of the Trainingsset.
Chunksize in ms 40 — Annotations are chunked in frames of this size (in ms).
Discretisize Continuous values ☑ — Discretisize continuous Annotations
Number of history frames 10 — Continous Annotations are discretized in these classes (seperate by ;)
Nodes in format: role_scheme ☐ Store role in column ☑ — Timesteps (0 for static)
If checked, node format is role_scheme, else annotations are added up

Figure 43.: Annotations from multiple annotators can be selected to create a data sheet from training a predefined network.

For the training process, multiple options can be set. Annotations are split in frames of a fixed size, as seen in Figure 44. This way, the network can learn to find correlations between the appearance of certain events that happen at the same time (or in time sequences). Continuous annotations are discretised to the amount of classes the user sets. E.g., if we have continuous annotations for engagement ranging from 0 to 1, the tool will automatically map these to classes (here: VERYLOW, LOW, MEDIUM, HIGH, VERYHIGH). Finally, the user can set the number of time frames that should be considered for the training step.

Figure 44 illustrates this process in a simplified version. In this example, we have three annotation tiers A, B and C. Each tier contains multiple labels, here represented by colours. We split the annotations on the tiers into smaller units, e.g. to stay in the previous example into frames of 40ms. Now a table is created, containing a column for each tier at the current time t. Additionally, for each selected history timestep an additional column is created. To simplify the illustration we use two time steps here, representing t-1 and t-2. We fill the table by adding a new line for each point t in time until we reach the end of the tier.



A_t	A_t-1	A_t-2	B_t	B_t-1	B_t-2	C_t	C_t-1	C_t-2
REST	REST	REST	REST	REST	REST	REST	REST	REST
BLUE	REST	REST	REST	REST	REST	REST	REST	REST
BLUE	BLUE	REST	REST	REST	REST	GREEN	REST	REST
BLUE	BLUE	BLUE	RED	REST	REST	GREEN	GREEN	REST
REST	BLUE	BLUE	RED	RED	REST	GREEN	GREEN	GREEN
YELLOW	REST	BLUE	RED	RED	RED	REST	GREEN	GREEN

Figure 44.: A simplified illustration of the process of creating data sheets from concurrent annotations to train dynamic Bayesian networks with NOVA. For each tier (A, B and C) annotations are split in segments of a fixed size. The current observation is held in the column (A_t, B_t and C_t), as well as the previous observations (here: t-1 and t-2). Continuous annotations are discretised, e.g. in segments ranging in categories from very low to very high.

Once the data sheet is created, it is combined with a modelled Bayesian network to automatically map the nodes of the network with annotator/scheme combinations from NOVA. To learn parameters in the model, the expectation maximisation (EM) algorithm (Moon, 1996) is applied. Expectation maximisation is an iterative algorithm that aims to find the maximum a-posteriori probabilities of parameters in statistical models that depend on latent unobserved variables.

In order to train Bayesian networks with the NOVA tool, annotations are used to generate data sheets to learn the probabilities in the network. There has to be a trade-off between training the model on sheer manual annotations, which represent the "ground truth" and deliver a perfect foundation, and on the other hand automatically created annotations that represent outputs that our classifiers are actually able to predict. By employing the cooperative machine learning approach suggested in Section 6.2, we are able to adapt the outcomes of the machine already during the annotation process, so that the models reach a state where we can "trust" them to create annotations similar to a human annotator. That means in conclusion, for our Bayesian network to work as expected, we need to find social cues that are recognised reliably well and that represent the problem at hand. Of course, the complex attitude (in this case the conversational engagement) needs manual annotations to represent the ground truth. As many social phenomena are not straight-forward in terms of interpretation, it is preferable (maybe even necessary) to have multiple raters for the given problem. Here, two strategies seem feasible. The first one is to create a common annotation, based on the ratings of all annotators, as suggested in Section 5.3.5.2. The second approach is to consider each annotator independently, meaning the same observations with possibly different ratings are shown to the EM algorithm multiple times.

The Bayesian networks used in our system are visually modelled with the GeNIe tool and learned with the SMILE library*, using extracted annotations as parameters.

*https://www.bayesfusion.com

Once we learned the parameters of our (D)BN, we may use it either for statistical prediction purposes, in a simulation or in a real-time scenario, by updating the nodes with evidences received from our social signal interpretation (see Section 4.2) component, as well as external sources such as an interaction management system (see Section 7.4.5.2). We receive these evidences by using the network in a SSI pipeline, updating evidences with observations from multiple social signal recognisers(optionally also with external information).

7.4.4 Experiments

To find out wheter our theory-based transparent model can compare with other "state-of-the-art" classification feature sets and models in terms of prediction accuracy, we evaluated the model on a test set of the NOXI corpus and compared the results with other approaches, e.g. with a feature level fusion in support vector machines. The main focus of this work is not to find the "best prediction" algorithm, but to find a model that is transparent and comprehensible and at the same time delivers comparable results to black-box approaches. At first, we applied a rather "classical" approach by applying various feature sets extracted on streams from single modalities.

As we experimented with multiple variations of network structures, which in some cases took over 20 hours for training, we decided for a train/eval set split instead of a cross-validation. To this end, for all trained models we used the data of sessions 2, 3, 5, 6, 8, 9, 10, 14, 30, 34, 52, 68 and 76 (411,41 minutes, 617124 samples) . As evaluation set we picked sessions 1, 4, 7, 11, 39 and 77. (Overall duration of 219,08 minutes, 328625 samples). The sessions were picked to represent different genders (3w, 7m) and languages (FR, UK, DE). We also made sure that participants did not appear in more than one session.

The engagement annotations were created on the ratings of 4-7 different annotators, using the inter-rater agreement and merging mechanisms in the NOVA tool. The annotation for one user of the corpus can be seen on the upper tier in Figure 45. To measure the efficiency of our predictions, we employ the Pearson correlation coefficient as described in Section 5.3.5.2. A PCC of close to $|r| = 1.0$ represents a high correlation between the gold standard annotations and the predictions where |r| close to 0 represents no correlation at all. As mentioned earlier, various interpretations of PCCs exist. According to Cohen (1988) an |r| = .10 is considered a weak correlation, |r| = .30 a medium or moderate correlation and |r| = .50 a strong correlation. In other interpretations, e.g. in psychological questionnaires values up to .30 are considered a slight correlation where .50 is considered as moderate and .70 - .80 a very high correlation.

To get an impression and a baseline of how standard standard feature sets perform on our corpus for the recognition of conversational engagement, we trained linear support vector machines on the gold standard annotations of the training set and evaluated them on the test set. Please note that the correlations only relate to our problem at hand (conversational engagement on the NOXI corpus) and are no general statement about the quality of the feature sets on other recognition tasks. The weighted (on the duration of a session) average PCCs in tests can be seen in Table 9. All PCCs described in this section are considered statistically significant (p < .001) given the overall sample size of 328625, and minimum sample size of 13400 for the shortest session). The significance tests were performed using the Fisher r-to-z transformation (Fisher, 1915).

Table 9.: Average PCCs on feature sets for single modalities using a linear support vector machine

Feature set	Modality	PCC
eMAX	Video (Face)	.2914
Openface	Video (Face)	.5060
Soundnet	Audio (Voice)	.6365
Gemaps	Audio (Voice)	.6355
BodyProperties	Kinect (Skeleton)	.5963

The sets we used to create the baselines contain the eMAX (Valstar, Martinez, et al., 2010) and Openface (Baltrusaitis, Robinson, and Morency, 2016) feature sets that contain both facial landmarks and action units. The Soundnet features are based on the Soundnet (Aytar, Vondrick, and Torralba, 2016), a pre-trained deep neural network. The network is "cut off" in the 18th layer and the weights of the layers are extracted as features (A so called end-to-feature approach). GEMAPS (Eyben, Scherer, et al., 2016) is a general audio feature set designed for a wide spectrum of audio recognition problems. The BodyProperties feature set is our own implementation of features related to body movements and postures.

The goal of our experiment was to compare the DBN described in the last section with these standard feature sets to see if a manually modelled and transparent approach can compare with state-of-the-art black box approaches. Of course, the suggested model is a multi-modal one while the feature sets in Table 9 only consider a single modality (audio, face, body). We ran multiple versions of the BN in an offline simulation based on test set annotations from the NOXI corpus. A multi-modal (non-dynamic) network achieved a PCC of .6596 based on visual input events which is considered (depending on the definition) a moderate uphill (positive) linear relationship. When adding audio-based events, e.g. the continuous arousal outputs for the voiced parts the network even achieved an average PCC of .7373 which is considered a strong uphill (positive) linear relationship by almost all definitions. By extending the BN with temporal links for nodes related to body and face movement, as well as voice activity to create a Dynamic BN (DBN) the PCC improved significantly ($p < .001$) to .7443. Noteworthy, the same DBN without the context node (trained and evaluated separately for novice and expert data) for the role of a person resulted in a significantly ($p < .001$) worse PCC of .7362. One explanation could be that background context (in our scenario) has a greater impact than temporal context on the results.

Table 10.: Average PCCs on multimodal inputs

Method	**Modalities**	**PCC**
Keras FC NN	Face, Body, Voice	.6034
LSVM	Face, Body, Voice	.6253
BN	Face, Body, Context	.6596
BN	Face, Body, Voice, Context	.7373
DBN (10 timesteps)	Face, Body, Voice	.7362
DBN (10 timesteps)	Face, Body, Voice, Context	.7443

The (D)BNs we applied are created using a hybrid fusion approach where classification results for sub-recognition tasks, as well as threshold based features are used to update the evidences in the network. This makes it difficult to compare the multi-modal model with other classification models that rely on low level features. To this end we performed a feature level fusion on the Openface features, the Soundnet features and the raw skeleton tracked by the Kinect, resulting in a feature vector of dimension 820. A linear support vector machine achieved an average Pearson correlation coefficent of .6253. The best fully-connected neural network in our tests (learning rate=0.0001, batch size=32, 10 epochs) implemented in Keras achieved .6034.

Overall, the goal was to show that a hybrid fusion approach using a theory-modelled DBN can deliver comparable results to state-of-the-art black-box approaches. On our corpus it even slightly outper-

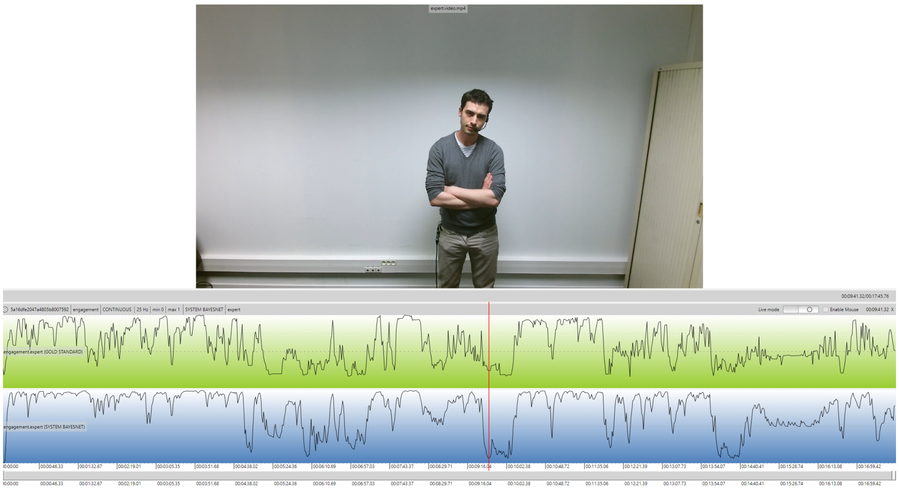


Figure 45.: The engagement output for a participant of the NOXI corpus shown in the NOVA tool. On the top tier the gold standard annotation is shown while on the bottom tier the predictions of the previously introduced dynamic Bayesian network are visualised for the whole continuous interaction

formed the other classification methods for visual features and to a wide extend when adding predictions from the audio channel and temporal links considering temporal context. We explain this with several aspects: by employing the transparent DBN we could intuitively refine our first assumptions on what influences engagement, which allowed us to incrementally add features and classification results, until the network achieved satisfying correlations with our gold standard annotation. Further, through the update mechanism on annotation/event abstraction we aimed to simulate a decision making- and reasoning process that's similar to the one of humans. To our understanding, humans will consciously or unconsciously map abstractions of behaviours on their perception of the other person. The results are summarised in Table 10. In Figure 45 we see the gold standard annotation for a participant of the NOXI corpus on the upper green tier, and the prediction of the final DBN on the lower blue tier. In the next section we'll investigate how we can extend our model with interpersonal dynamics as another context factor.

7.4.5 Multi-person Model

In interactions between humans, we might not only be interested in the observed social cues from one person but from multiple, and how people influence each other. On the one hand, this is useful if we want to analyse a person with regard to context information provided by the dynamics of the interaction. For example, a headnod does not have any deeper meaning when considered in isolation. Depending on the context, it often is used to support the other person's state-

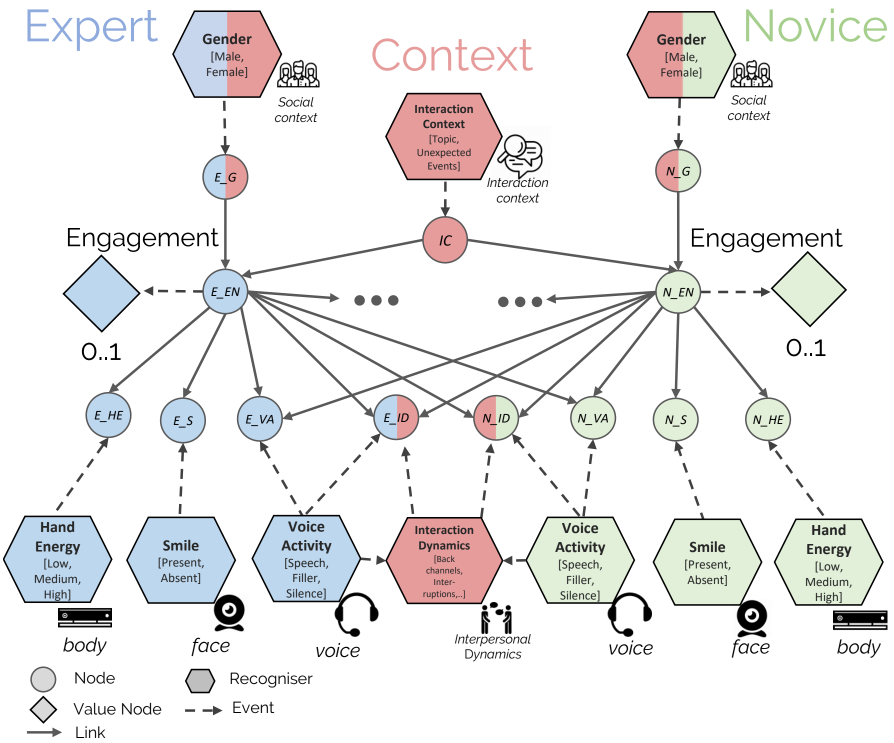


Figure 46.: Schematic of a single time slide in a simplified dynamic Bayesian network for two persons. This illustration extends Figure 42 with the social cues of a second person that are considered for the analysis. This time the role is not encoded with a node, but rather each node exists for each person separately. Additionally, contextual nodes are added that determine interpersonal cues such as backchannels or interruptions and information about the interaction, such as the "unexpected events" in the NOXI corpus.

ment, which in most cases is considered an indicator of attentiveness and involvement in the conversation. On the other hand, we might be interested in multi-user models to make assumptions about group phenomena. For example, this could help to judge a conversation as enjoyable or interesting, if all participants of the conversation show signs that indicate their engagement in the interaction. Our framework does not only allow modelling complex social attitudes of one person, but of multiple persons in parallel within the same model. For example, in a dyadic social interaction we might be interested to measure the engagement of both interlocutors, given their individual and the interpersonal cues and interaction dynamics. This way we might identify further findings, for example the parts of the conversation that were engaging for both participants. Figure 46 is an extension of Figure 42 which holds nodes for two individuals.

For each node, the role is encoded, e.g. we would now have nodes with the ids expert_voiceactivity (in the figure abbreviated to E_VA) and novice_voiceactivity (in the figure abbreviated to N_VA). Additionally to the scheme, presented in Section 7.4.2, we now also add interpersonal cues, such as backchannels (see Section 2.3.2.1) or interruptions (see Section 2.3.2.2). Such cues depend on the behaviours of both interlocutors. In the following, we'll exemplify the automated detection of such conversation transitions, which then are added as a node in the network.

7.4.5.1 Human to Human Interaction Example: Conversational Interruptions and Backchannels

This section is based on the publication: Cafaro, Wagner, Baur, Dermouche, Torres, Pelachaud, André, and Valstar, 2017

To gather annotations about interruptions and backchannel signals to train a multi-person network, we decided to apply semi-automated strategies. As described in Section 2.3 the amount of interpersonal cues is a valuable indicator of the engagement of a person. To this end, as a first step, we automatically analysed NOXI's audio recordings using the SSI framework (see Section 4.2) and, based on the voice activity detection (VAD) we automatically extracted segments of **turn transitions**.

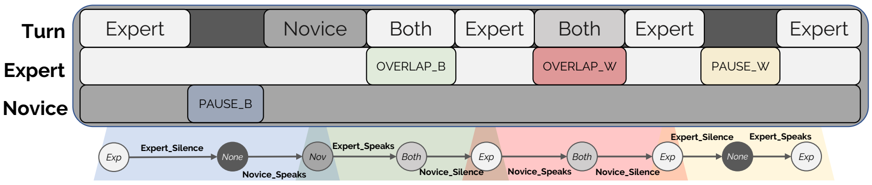


Figure 47.: Example annotations of communicative states and turn transitions tiers automatically implied from Expert and Novice's voice activity detection.

For the turn transitions we followed the ideas from Heldner and Edlund (2010) and Roger, Bull, and Smith (1988) (also see Section 2.3.2.2. In the annotation, a turn transition represents a change of the actual conversational state which can result in a switch from speech to silence and vice versa for the same speaker (i.e. within turn) or between the two speakers (i.e. between turns). Therefore, the following turn transitions are automatically extracted in two separate tiers for each, the expert and the novice. A **pause within** turns (*PAUSE_W*) is a silence between two turns of the same speaker without speaker change. A **pause between** turns (*PAUSE_B*) is a speaker switch from expert to novice (or vice-versa) with a silence in between. A **perfect** turn transition (*PERFECT*) happens with a speaker change without a silence nor an overlap in between. A high amount of "perfect" transitions indicates a fluent conversation. An **overlap between** turns (*OVERLAP_B*) is marked when an overlap between speakers occurs that leads to a change in the turn. Often, this behaviour indicates an interruption, where the interrupting person tries to take over the turn. Whereas an **overlap within** a turn (*OVERLAP_W*) is a speaker's overlap without any speaker change, e.g. in a backchannel or short comment.

In order to obtain the labels, we ran a parallel finite state machine that used state transitions for labelling the turn transitions (see Figure 47). Therefore, a conversation state that goes from EXPERT_speechs to SILENCE and then to EXPERT again, triggers a PAUSE_W (pause within turns) label in the turn transitions. Figure 47 shows an example annotation. In the turn transition tiers, the labels are always assigned to the tier corresponding to the previous listener. The first pause between turns (*PAUSE_B*), for instance, is assigned to the novice tier because there is a transition that goes from the expert to the novice through a pause in the expert's speech. The expert's tier obtains an overlap between turns because while the novice is speaking the expert interrupts to claim the turn. We add the automatically generated annotations to the training set of a dynamic Bayesian network, besides regular abstractions of social cues. This way we are able to learn time-series of turn coordination behaviours and their correlations to the conversational engagement. For example we learn if an overlap within (which in most cases represents a backchannel) has influence on both the speaker and the listener's engagement.

7.4.5.2 Human to Agent Interaction Example: Interaction Context

A multi-person model must not necessarily include multiple actual persons, but can also include a computational agent. For example in the Aria Valuspa Platform, introduced in the beginning of this section, one interlocutor is represented by the virtual agent Alice. Compared to human-human interactions, in human-agent interactions context information can be implicitly logged and used for the analysis of the user. In most cases the behaviours of agents or robots are coordinated by an underlying interaction management system. In various research projects, a couple of dialogue and interaction management tools were developed. Prominent examples are Disco (Rich and Sidner, 2012), Flipper (ter Maat and Heylen, 2011) and VisualScenemaker (Gebhard, Mehlmann, and Kipp, 2012; Mehlmann and André, 2012).

We will employ the VisualSceneMaker (VSM) here to illustrate basic concepts of logging context information in interaction management tools. To actually make use of information about the state of the interaction, the interaction manager needs to provide such information to the signal interpretation component, respectively the reasoning model. Modelling concepts of VisualSceneMaker intuitively support the recording of information about the agents' behaviour, the various kinds of dialogue context knowledge and the progress of the interaction.

Based on the modelling concepts of VisualSceneMaker, various mechanisms may be used to provide our interpretation system with information about the agents' behaviours and context knowledge.

In VSM, an author may specify arbitrary logging directives in a scene. Figure 48 shows an example in which an author uses log direc-

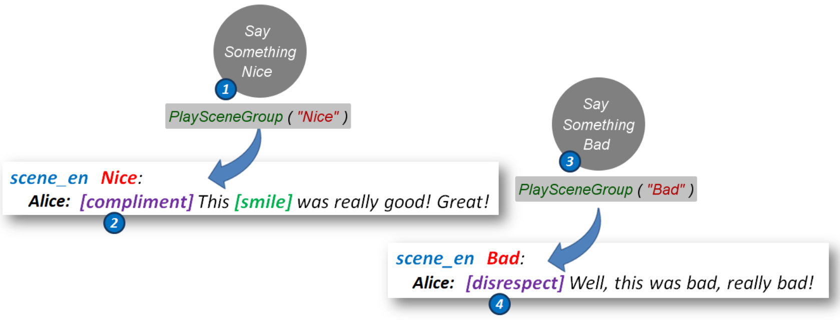

Say Something Nice
1
PlaySceneGroup ("Nice")
scene_en Nice:
Alice: [compliment] This [smile] was really good! Great!
2
Say Something Bad
3
PlaySceneGroup ("Bad")
scene_en Bad:
Alice: [disrespect] Well, this was bad, really bad!
4

Figure 48.: An author may use logging directives within a scene to provide NOVA with arbitrary meta information.

tions directly in the textual description of a scene to explicitly create events whenever the agent provokes a certain emotional reaction in the user. These emotion eliciting events and the user's prompt emotional reactions are later automatically analysed to find out if they have achieved their purpose by e.g. upsetting the user. In the example on the left side in Figure 48, three kinds of information will be logged a) The content of what the agent is saying (with start and end time) b) The occurrence of a non-verbal behaviour (smile) and c) meta information that this sentence has the purpose to be a compliment. Tags are defined in a dictionary and depend on the character rendering engine. In the illustration, purple tags deliver meta information about the content, while green tags are concurrent with animation commands for the agent.

By logging behaviours in interactions between humans and agents, we are enabled to use similar techniques compared to the ones from human-human interactions to analyse the interaction dynamics. While a scene is scheduled on the character engine, the scheduling algorithm notifies the start and end event of each gesture animation, facial expression or synthesis of a spoken utterance to the Sceneplayer, which then automatically forwards these events to the dynamic Bayesian network. When a user is performing a specific social cue, a possible question for the analysis of this behaviour is, if this social cue is a response to a stimulus. For example, does the user nod his head because she or he is trying to encourage the interlocutor in their current topic?

The example in Figure 49 illustrates how the system combines behaviour of the user, the agent, as well as the context to detect interpersonal cues. According to the agent's scene script, the system is provided with the meta information that the agent's actual utterance is a compliment. The agent is looking at the user's face directly before starting to speak. If the user returns the gaze within a certain timespan, by looking at the agent's face, the system will label a successful mutual gaze. Similarly, the system recognises the mirroring of a smile. For example, as the agent shows a smile and the user

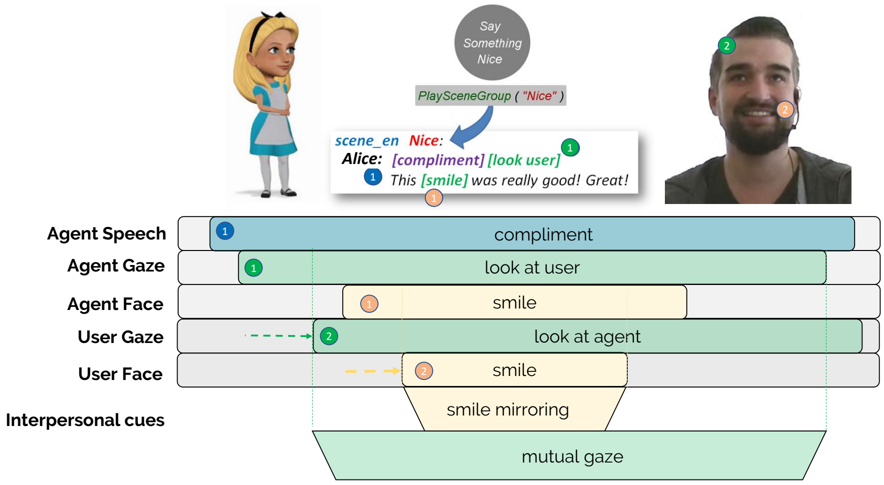


Figure 49.: The Sceneplayer implicitly records the beginning and the end of utterances and nonverbal behaviours. This allows putting them in relation to the user's social cues which indicate interpersonal cues, such as mirroring or mutual gaze.

smiles back within a predefined timespan, as visualised in Figure 49, this interpersonal behaviour (mirroring) will also be recognised by considering the temporal alignment. A high amount of interpersonal cues increase the dynamic of a conversation and are considered a reflection of high engagement.

7.4.5.3 Multi-person Network Experiments

We extended our evaluation from Section 7.4.4 with the presented multi-person model, which contains information of both interlocutors and high-level abstractions of interpersonal cues. In particular, the network contains information about "mutual gaze" and the turn transitions described in Section 7.4.5.1. Here we used the same training and test set, but with additional labels for the interpersonal cues. The dynamic multi-person network achieved a PCC of .7545, which again is significantly better ($p < .001$) (using a Fisher r-to-z transformation (Fisher, 1915)) than the best DBN (.7443) in the previous experiments that was trained on the cues of a single person and context information about the role and gender of a person. We conclude that adding additional context information about the interpersonal dynamics (here mutual gaze and turn transitions) actually leads to improvements in the correct and adequate interpretation of complex behaviours.

7.5 Exemplary Case II: Emotion Regulation Strategies

"Shame is the most powerful, master emotion. It's the fear that we're not good enough."

— Brene Brown

We now want to turn to another exemplary complex behaviour. In particular, we are interested in regulated emotions (see Section 2.4.2.3) in the context of job interviews. The goal of this work is to lay the basis for a deeper cooperative human-machine analysis of social and emotional signals and their connection to cognitive processes. This means, we need to provide transparent concepts and explanations of the predictions of a recognition system. On the one hand, this helps researchers to create better models to explain correlations between social signals and context information, and on the other hand, it provides job candidates (or users of a job interview simulation system) with constructive and comprehensible feedback so that they are able to learn from the predictions. In this section, we introduce a first example exploitation of our model in an automated debriefing session, where a virtual agent is presenting feedback to a user about their performance in a virtual job interview, with regards to the regulated emotion "shame". The agent here needs to be capable of explaining its observations and inferences about the users performance, because this way, the user might eventually accept feedback and learn from it. The overall user model we describe in this section is entitled MARSSI (Model of Appraisal, Regulation and Social Signal Interpretation). As in the previous section, we first like to introduce the original scenario the MARSSI model was designed for.

7.5.1 Scenario: Virtual Job Interview Platform

This section is based on the publication: Baur, Mehlmann, Damian, Lingenfelser, Wagner, Lugrin, André, and Gebhard, 2015

This section describes the TARDIS project, respectively its successor EmpaT. These systems are job interview simulations, where a human interacts with a virtual character to improve their skills. The original motivation for the TARDIS project was the issue in the rising number of young people *not in employment, education or training* (NEET). NEETs often have underdeveloped socio-emotional and interaction skills (MacDonald, 2008; Hammer, 2000), such as a lack of self-confidence and sense of their own strengths. This affects their performance in various critical situations, such as job interviews. To address this issue, many European countries have specialised inclusion centres, meant to aid young people with secure employment through coaching by professional practitioners. One problem of this approach is that it is very expensive and time-consuming. Considering this, technology-enhanced solutions present themselves as viable and advantageous alternatives to the existing human to human coaching practices. Job interviews are used by the potential future employer as a means to determine whether the interviewee is suited for the company's needs. To make an assessment, interviewers heavily rely on social cues, i.e., actions, conscious or unconscious, of the interviewee that have a specific meaning in a social context, such as a job interview.

7.5.1.1 Real-Time Interaction System

The *TARDIS social cue training game* is an approach to support young adults in job interviews. It employs gaming techniques and methods to motivate adolescents and young adults to improve their social skills. For the game we use the virtual character Gloria as seen in Figure 50, which has been developed by the project partner Charamel GmbH. The scenario is set up in a virtual space modelled like a typical office environment (see Figure 50, right side).

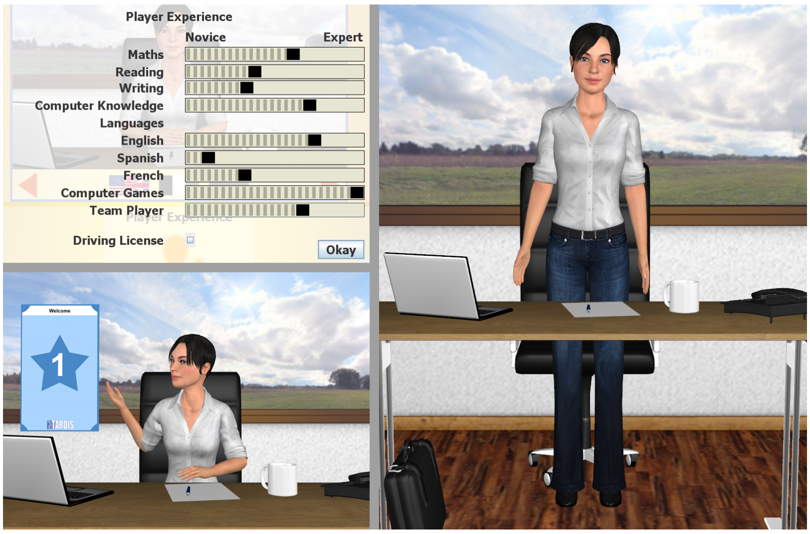


Figure 50.: Player experience, scenario, and welcome game phase.

The game is structured similarly to a job interview. It features three interview phases, namely *Welcome*, *Company Presentation* and *Strength and Weaknesses*. Prior to the *Welcome* phase, the user is given a short introduction into how the system works. At the start of the game, the user is also asked to provide information about general skills and background (see Figure 50, top left side). This information is used throughout the game to adapt the flow of the interview to the user's profile and the context-sensitive recognition of the user's complex emotional states.

While playing the game, the participant is asked to adapt to specific social task situations, which are related to the game phase. The *Welcome* phase is related to the social task of presenting oneself. The subsequent phase of *Company Presentation* is related to the task of care-fully listening and the last phase (*Strength and Weaknesses*) is related to a conversation about the user's profile. The user is expected to adapt her or his behaviour to each phase. Which type of behaviour is appropriate to a specific phase, is described on physical game cards (see Figure 51). Each game card contains several social cues, which the user should or should not perform. For example, the *Welcome* card instructs the user to 1) smile, 2) hold eye contact, 3) use open gestures, 4) speak loudly and 5) to not freeze up. These social cues have been identified by experts, e.g. social workers and job recruiters.

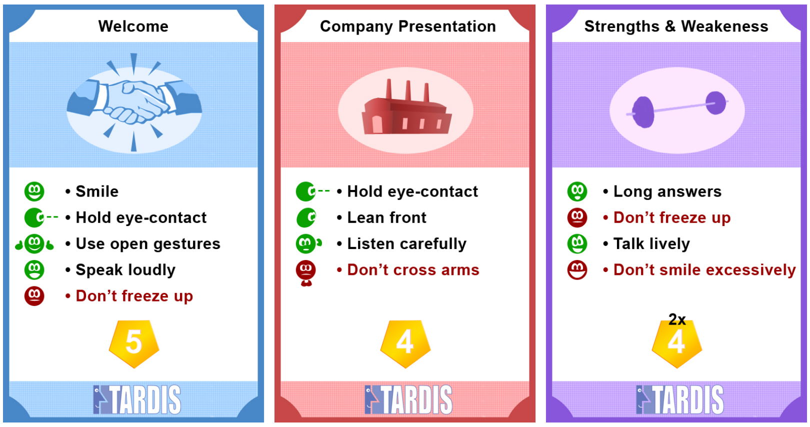


Figure 51.: Social cue action cards for each game phase.

These cards are given to the user prior to the interaction. During the game, the virtual agent informs the user before each phase which game card is relevant in the upcoming phase and only proceeds with the interview once the user confirms having read the game card (Figure 50, bottom left side). Furthermore, the social cues are also displayed directly on the game screen using graphical symbols. These symbols change in appearance depending on whether the user performed the underlying social cue or not. For example, if the user shows an appropriate amount of smiling during the *Welcome* phase, a smile symbol gets highlighted on the screen.

To encourage adequate behaviours, the system also scores the users based on their performance. More precisely, every time a user behaves in compliance with the game card, i.e., performs a requested social cue, she or he receives a point towards the total score. Some of the cues have to be performed (or not performed) for the whole duration of the interview phase, e.g., *do not freeze up*.

The TARDIS job-interview game was extended in a follow-up project, named EmpaT. The virtual environment was enhanced with a large office building and the recognition modules for social cues were updated. To recognise complex social cues, such as the coping and regulation strategies, we built the MARSSI model, which will be described in more detail in this section.

7.5.1.2 In-the-wild User Study

This section is based on the publication: Baur, Mehlmann, Damian, Lingenfelser, Wagner, Lugrin, André, and Gebhard, 2015

To evaluate the impact of the training and coaching system on actual users of such a system, we conducted a study at a secondary modern school (Mittelschule Stadtbergen, Augsburg, Germany) over the course of three days using the TARDIS game application. Within the TARDIS game, the NOVA tool serves to analyse the learner's social cues when interacting with a virtual recruiter during a virtual job interview training, with respect to dialogue dynamics and context information. In this study, we additionally employed an earlier version of the dynamic Bayesian network, described in the first use-case, to recognise a person's engagement. The continuous predictions of the model served to support the coaches to guide the post-interview sessions conduced with our experimental group. For example, coaches looked at the parts where the system recognised that the participants appeared to be disengaged, and based on the recognised abstractions of social cues they replayed the scenes to the pupils and gave immediate feedback on how to improve in such situations.

The participants of the study were 19 pupils (10 male and 9 female) in their final or pre-final graduation year, aged between 13 and 16 (mean = 14.37; SD = 0.94) which were categorised by their teachers as being at risk of exclusion. Most of them already started looking for employment. Two professional career counsellors from the Career Service of the University of Augsburg volunteered to support us in the study. The main objective of the system was to evaluate the impact of the TARDIS game and the NOVA tool (as a transparent coaching interface) on the pupils.

On the first day, each student participated in a job interview led by one of the professional career trainers. Career trainers were instructed to be as objective as possible and to focus on the non-verbal behaviour of the participants. After each interview both, career trainers and pupils, filled in questionnaires. The purpose of these first interviews was to establish a baseline, regarding the job interview performance of the pupils, before their interaction with the system. The questionnaires contained the following questions:

- *Questionnaire A:* On a seven-point Likert scale career trainers rated the pupil's 1) overall performance, 2) recommendation for the job based on their behaviour, 3) appropriate usage of smiles, 4) appropriate usage of eye contact, 5) appropriate usage of gestures, as well as the pupil's 6) nervousness 7) interest and 8) focus.

- *Questionnaire B:* Pupils rated on a seven-point Likert scale whether they thought they 1) performed well in the interview, 2) were nervous, 3) used a lot of filler words such as "er" or "uhm", 4) were not focused, 5) were aware of their non-verbal behaviour and 6) performed appropriate non-verbal behaviour.

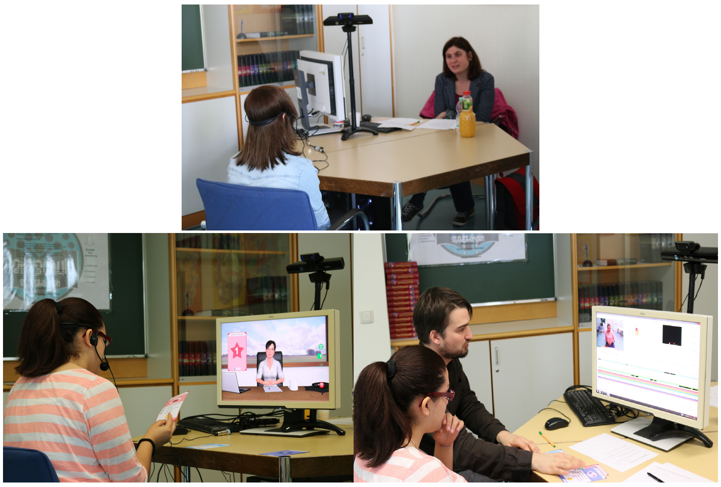


Figure 52.: Training job interviews have been performed on day 1 and 3 with professional career counsellors (top). The bottom pictures illustrate the setup of the system with a participant acting with the TARDIS Game and during the debriefing session with the NOVA user interface

On the second day, pupils were randomly assigned to either the control group (CG) or the experimental group (EG) ($N(CG) = 9$, 5 female, 4 male; $N(EG) = 10$, 4 female, 6 male).

The EG interacted with the TARDIS Game and the NOVA interface. Each training lasted for about 15 minutes, split between game interaction (see Figure 52, bottom left) and debriefing. During the session, their non-verbal behaviour was recorded and analysed by the system. A debriefing phase followed each interaction with the game. In this phase, a researcher assisted the pupils in reviewing the interaction, using the NOVA tool as coaching interface (see Figure 52, bottom right).

Pupils of the CG were reading a printed job interview guide for the same amount of time the EG interacted with the system. The written guide was published by a renowned German youth advisory institution which the school regularly cooperates with for their employment preparation classes.

On the third day, a second set of mock job interviews with the professional career trainers was conducted with each participant. Pupils of both groups (EG and CG) were brought to the career trainers in random order, who were unaware to which group the pupils were assigned during the second day. After each interview, career trainers and pupils filled in the same questionnaire they answered during day 1. This allowed us to make a direct comparison of the participants' performance between day 1 and 3.

Comparing the two groups after the third day revealed interesting insights. Pupils that interacted with the training system were rated better by the career counsellors on all dimensions compared to the CG. An independent two-tailed t-test with Bonferroni-Holm (Holm,

1979) error adjusted significance levels α yielded statistically significant differences for the career counsellors' ratings on overall performance ($p = 0.004$, $\alpha = 0.006$) (α represents the significance level). Also, a strong trend was found for the recommendation dimension ($p = 0.012$, $\alpha = 0.007$). To evaluate the improvement of performances for each group individually, we compared the results within groups between day one and three. Pupils of the EG were rated better on all dimensions on the third day, compared to the first. Performing paired two-tailed t-tests (again with Bonferroni-Holm error adjusted significance levels) revealed significant differences for the dimensions recommendation ($p = 0.005$, $\alpha = 0.006$), overall performance ($p = 0.006$, $\alpha = 0.007$), nervousness ($p = 0.006$, $\alpha = 0.007$), eye contact ($p = 0.007$, $\alpha = 0.010$) and smiles ($p = 0.012$, $\alpha = 0.013$). The pupils' self reports revealed significant differences on the nervousness dimension ($p = 0.001$ $\alpha = 0.008$), with participants rating themselves being less nervous on the third day compared to the first day. An overview on the results and standard deviations can be found in Figure 53.

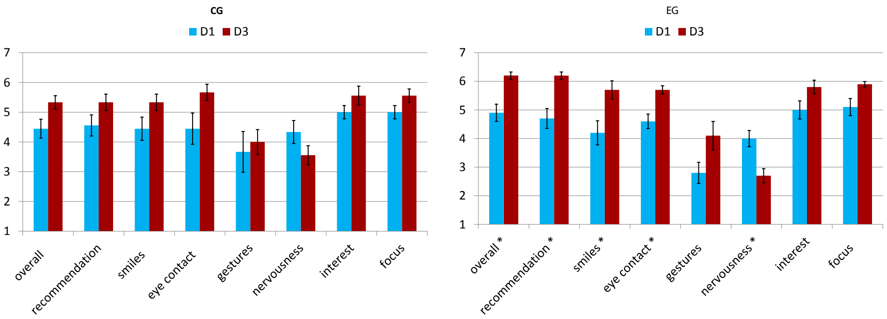


Figure 53.: Practitioners' ratings of CG (left) and EG (right) across day one and three (including SD). Dimensions marked with * present significant differences between the two days.

For the experimental group, we further asked the pupils about their impressions right after interacting with the system on day 2:

- *User Experience Questionnaire:* Pupils rated on a seven-point Likert scale whether they 1) found the video self reflection with the NOVA interface useful, 2) had the impression they learned from the self-reflection 3) would use the training system for job preparation, 4) found the gaming cards helpful 5) had fun playing the game

The participants stated they had fun playing the game with a mean result of 5.6 which we consider a good result for a training and learning environment. The pupils further rated the helpfulness of the game cards high (mean = 6.4) which suggests that direct guidelines for specific conversational topics are useful to pupils, which we assumed by designing the application. They rated they would use the system at home to prepare for a real job interview with mean = 6.1. The post interview with the NOVA tool also received high ratings for helpfulness (mean = 5.7) and learning effect (mean = 5.5). The study showed that the transparent analysis of interactions with a training system helped participants to actually learn from the machine's predictions. Here, NOVA was applied as explanatory interface that supported the coaches, but also the pupils in reflecting on both appropriate but also inappropriate behaviours that appeared during the interaction. More details on the study can be found in (Baur, Mehlmann, et al., 2015) and (Damian, Baur, et al., 2015).

7.5.2 The MARSSI Model

This section is based on the publication: Gebhard, Schneeberger, Baur, and André, 2018

Based on the experiences with the TARDIS system, in its successor project EmpaT, we extended the automated recognition of social signals in the context of the virtual job interview simulation. In this work, we focus on situations where people are challenged by a job interviewer. The goal is to create transparent models for the recognition of emotion regulation strategies. In particular, we focus on emotional coping strategies for the appraisal of shame-eliciting signals. We call the overall model MARSSI (Model for Appraisal, Regulation and Social Signal Interpretation). The theories behind our approach are based on the Orthony, Clore, and Collins (OCC) model (see Section 2.4.2.3) but we extend them by Moser and von Zeppelin (2005)'s functional emotion classification. Modelling emotions based on the OCC Model using dynamic Bayesian networks, respectively dynamic decision networks, was proposed in (Conati, 2002). In a later study Conati and Maclaren (2009) showed that by incorporating affective signals (Electromyography (EMG) from physiological sensors) the accuracy of such a user model can be significantly improved. Based on this work, the possibilities of quickly creating new models for specific social cues, as suggested in Chapter 6 and the training of a DBN using annotations from parallel tracks, we decided to created a model that relies on observation of social cues from multiple modalities, as well as surrounding context information, such as the discourse of the conversation (see Section 7.2). Further, the aspect of transparency is of vast importance here again, as the final results are used by the ALMA emotion simulation component (Gebhard, Kipp, et al., 2003; Gebhard, 2005) to generate textual explanations, which are presented by a virtual agent job recruiter in a debriefing session, right after the interaction. In MARSSI we consider both, appraisal rules, as well as regulation rules:

- An *Appraisal rule* defines how a situation is judged. With regard to cognitive appraisal theories, the situation is the elicitor of emotion. Thy are also used to model how the user could appraise a situation (also see Section 2.4.3). Multiple appraisals are allowed. We rely on the OCC appraisal theory (Ortony, Clore, and Collins, 1988), e.g., *GoodActSelf → {agency=self, praiseworthiness=1.0}*. MARSSI extends the appraisal notation with a confidence value, representing how likely the appraisal fits the detected social signals. The value is computed by a DBN, which is updated by various social signal recognisers.

- A *Regulation rule* defines how an internal emotion is regulated by changing the current appraisal information, triggering a reappraisal process that elicits a regulated emotion. They are also used to model how a user might regulate internal emotions. Multiple regulations are allowed. In MARSSI we extend the original OCC rules by processing regulation rules (also see Section 2.4.3.1).

  We created rules for the structural emotion *shame* following Nathanson's regulation theory (Figure 5). All rules contain *situational change rules* (marked with *sit_chg*) and corresponding OCC appraisal information:

  1. *AttackOther → {sit_chg:object self → object other; agency = other, praiseworthiness = -1.0}* regulates shame with reproach (negative praiseworthiness) by shifting the appraisal focus from one own's flaw to a blameworthy action of the person who is responsible for the shame experience.
  2. *Withdrawal → {sit_chg:other as actor → self as actor; agency = self, desirability = -1.0}* regulates shame with distress, elicited by a negative desirability, but replacing the person who is responsible for the shame experience with oneself, to the purpose of having control over the situation. A similar withdrawal rule might include a negative likelihood to elicit the regulated emotion fear.
  3. *Avoidance → {sit_chg:action → opposite of action|denial of action|...; agency = self, desirability = 1.0}* regulates shame with joy, elicited by a positive desirability of the imagined positive event, in which the shame action has not happened. A similar avoidance rule might have negative desirability and negative likelihood to elicit the regulated emotion fear.
  4. *AttackSelf → {sit_chg:other as actor → self as actor, action → intellectualisation of action; agency = self, liking = -1.0}* regulates shame with disgust, elicited by a negative liking and the transformation of the shameful action into an own "ugly" character feature that is less intense and can be changed by oneself in the future. In this case, the person who is responsible for the shame experience is replaced with oneself implicates having control over the situation.

  All regulated emotions of the shame regulation rules are situational emotions which are most likely communicated (non-

)verbally (e.g., Nathanson, 1994), hence become communicative emotions. Note that each regulation rule's OCC variable holds the maximal value (e.g., 1.0 or -1.0). It's sign determines the type of emotion. It's value may be used to calculate an emotion's intensity. For the proof-of-concept illustrated here, we are interested in the type only. Each rule holds a confidence value, computed by social signal classifiers, representing a value how likely the regulation fits the detected social signals.

Social cues in MARSSI are conceptually related to appraisal and regulation information expressed as communicative emotions. We employ models that are able to detect *time series of social signals* as they occur in the situation of emotion regulation. We focus on single social cues for the head gaze, specific gestures, and posture changes for the following appraisal and regulation information:

1. *BadEvent*: the user expresses anger directed towards the situation - away from the dialogue partner.

2. *BadActOther*: the user expresses anger towards the dialogue partner.

3. *BadActSelf*: the user shows a facial expression of shame (e.g., blushing), head/gaze points downwards, posture is slumped down. For all shame regulation classifiers the regulation takes time and might be accompanied by:

   3.1. *BadActSelf → AttackOther*: a lean forward posture and/or gestures that take up room, expressing anger towards the dialogue partner.

   3.2. *BadActSelf → Avoidance*: a lean back posture, gaze and head aversion and expressing joy towards the dialogue partner.

   3.3. *BadActSelf → Withdrawal*: few body movements, gaze/aversion, and expressing fear away from the dialogue partner.

   3.4. *BadActSelf → AttackSelf*: expresses disgust away from the dialogue partner, head/gaze is mainly pointed downwards.

To this end, the models for recognising single social cues included in MARSSI are trained using NOVA with a cooperative workflow (as exemplified in Chapter 6). To fuse sequences of multiple social signals, we again employ *dynamic Bayesian networks*. Besides automatically and cooperatively created annotations, for each situation, human experts manually annotate the perceived emotion regulation strategies.

During run-time, a confidence value, computed by the output of the non-verbal interpretation of the appraisal and regulation strategy is forwarded to the emotion simulation component, updating the possibilities of each modelled appraisal and regulation information. The recognition phase lasts as long as the listener handles the question or the comment.

The next section describes how we applied MARSSI to empower a virtual empathic agent to give transparent feedback to a user of a job interview simulation system.

7.5.3 Example Simulation

In this section we will illustrate the recognition of the structural emotion *"shame"*, elicited in job candidates. To train our DBN we recorded and annotated the EmpaT shame-eliciting corpus described in Section 3.4.2. Based on this data, we trained the DBN in a 50:50 split (based on sessions containing different users). To this end, we employed several social signal processing algorithms to generate labels for single social cues on multiple modalities of both, the interviewer and the candidate. Here again, some cues are calculated based on single, meaningful features, such as the energy of the motion vectors of both hands of a participant or the overall movement of the hands, head touches, and the openness of the body posture (for more details, see Appendix A).

For more complex cues, such as subtle smiles, we employed a linear support vector machine to train models based on manual annotations on the training subset of our corpus. For cues related to the head and face, we extracted OPENFACE (Baltrusaitis, Robinson, and Morency, 2016) features.

Analogously, we repeated this step for other modalities, such as the paralinguistic channel, by training a model to detect spoken words, fillers and silence, as well as models to detect the level of arousal from the audio modality based on GEMAPS (Eyben, Scherer, et al., 2016) features. Here again, a human annotator interactively corrected the annotations if necessary. After each session the models have been retrained as proposed in Chapter 6.

To find the ground truth of the observed emotion regulation strategy, we additionally labelled time segments including the duration of each question and the candidate's answer, with 1) the type of question as additional discourse context information (see Section 7.2) and 2) with the rating of human labellers for the classes related to regulation cues (e.g., AttackOther, AttackSelf, Avoidance, Withdrawal, and None). Thereby, we employed three annotators and merged their rating via majority vote with the NOVA tool.

Based on these semi-automated annotations we created a training set for our dynamic Bayesian network. It contains the concurrent appearance of the ground truth labels for the shame emotion regulation strategy, the topic of the conversation as context information and the observed social cues. We trained the dynamic Bayesian network using the expectation maximisation algorithm, to learn both the distribution of the single cues in our corpus, but also their influence on the shame regulation strategies. An instantiation of a typical time series of social signals for the regulation strategy "avoidance" on parallel tracks is illustrated in Figure 54. Here, as a reaction to the question *"Before we begin, let me ask you a short question: where did you find your outfit? It really doesn't suit you."*, the user smiles, while unconsciously looking to the side, and touching her blouse.

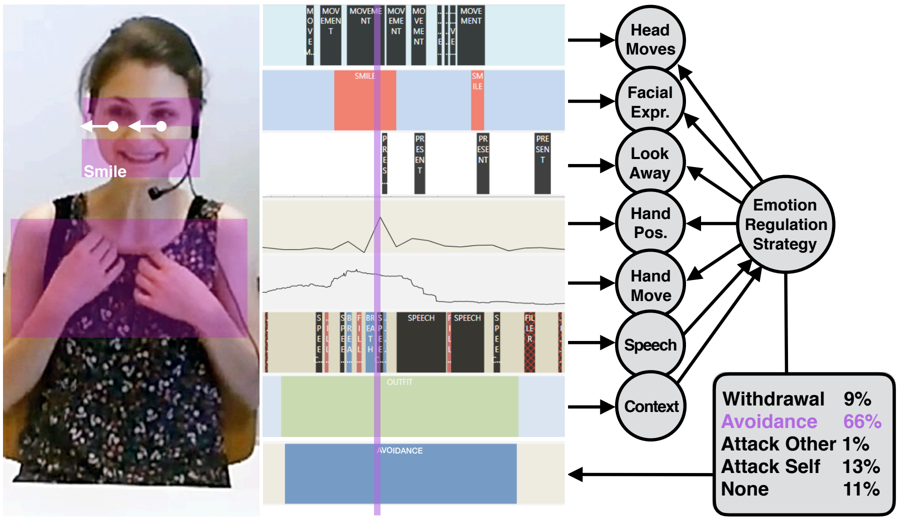
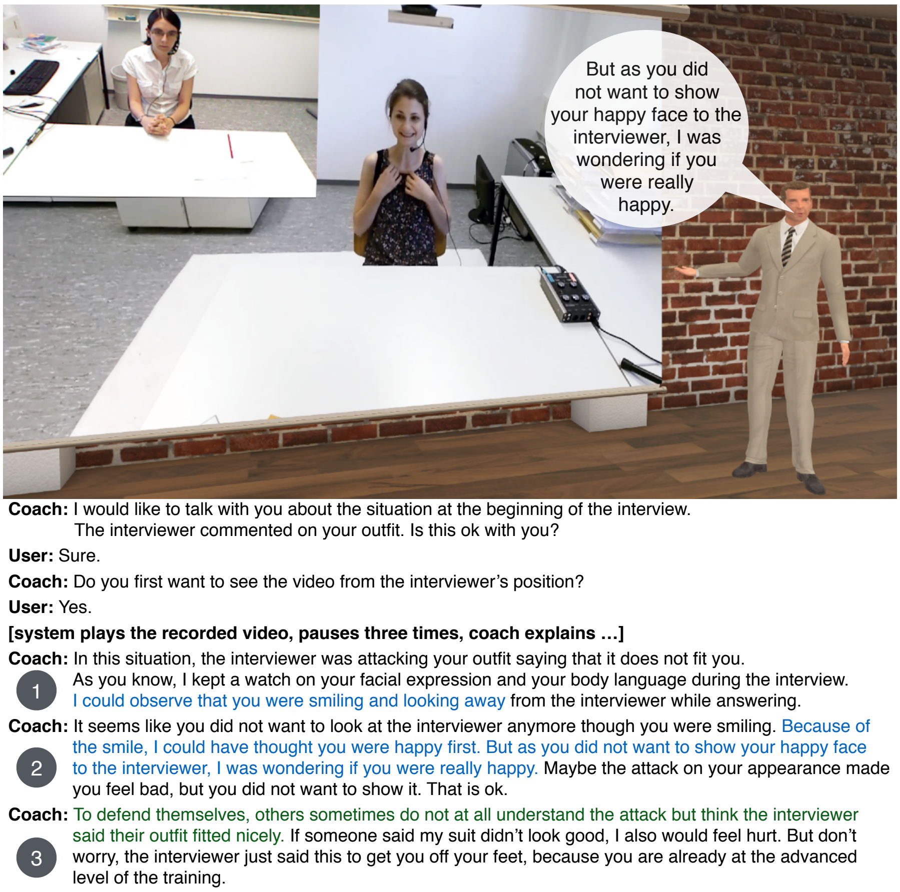


Figure 54.: Recognised and annotated cues deliver additional abstraction layers that are fed in a dynamic Bayesian network that infers the current shame regulation strategy and predicts it in real-time.

In a final step, we used ALMA's rule-based cognitive modelling and updated the rules with the outputs of the single signal classifiers and the networks prediction probabilities for the inferred emotion regulation strategy, to simulate user emotions in a debriefing session with our interactive virtual character Tom (see Figure 55). He has the role of a coach discussing the user's (non-verbal) reaction to the interviewer's question. Tom is embedded in a 3D virtual environment (Figure 55) capable of performing social cue-based interaction with the user. He utilises MARSSI's knowledge of the appraisal and the regulation strategies, in order to generate an empathic reaction. Currently, the reaction is based on the detected appraisal or regulation with the highest confidence value. The aim is, to support the user's self-reflection by explaining to her what MARSSI discovered from the social signals. We elucidate this with the previous example of the regulation strategy "avoidance" as a proof of concept application.

In general, our coach Tom (Figure 55, right) would first explain what sequences of social cues MARSSI has detected and how such cues can generally be interpreted. Afterwards, he would subtly explain the connection to the underlying structural emotion. We want to outline a possible interaction between a user and the coach in the example situation where the interviewer asked the question about the outfit, where MARSSI detected the following rule *Avoidance* → {*sit_chg:action* → *opposite of action*|*denial of action*|...; *agency* = *self*, *desirability* = 1.0} . This rule regulates shame with joy, elicited by a desirable imagined positive event in which the shame action has not happened.


But as you did not want to show your happy face to the interviewer, I was wondering if you were really happy.
1
2
3

Coach: I would like to talk with you about the situation at the beginning of the interview. The interviewer commented on your outfit. Is this ok with you?

User: Sure.

Coach: Do you first want to see the video from the interviewer's position?

User: Yes.

[system plays the recorded video, pauses three times, coach explains ...]

Coach: In this situation, the interviewer was attacking your outfit saying that it does not fit you. As you know, I kept a watch on your facial expression and your body language during the interview. I could observe that you were smiling and looking away from the interviewer while answering.

Coach: It seems like you did not want to look at the interviewer anymore though you were smiling. Because of the smile, I could have thought you were happy first. But as you did not want to show your happy face to the interviewer, I was wondering if you were really happy. Maybe the attack on your appearance made you feel bad, but you did not want to show it. That is ok.

Coach: To defend themselves, others sometimes do not at all understand the attack but think the interviewer said their outfit fitted nicely. If someone said my suit didn't look good, I also would feel hurt. But don't worry, the interviewer just said this to get you off your feet, because you are already at the advanced level of the training.

Figure 55.: A virtual coach discusses prominent situations during the Virtual Job Interview Simulation

As seen in Tom's explanation in Figure 55, he does not directly address the structural emotion. Especially in those cases, where the underlying structural emotion might be *shame*, the subtle approach is extremely important. Since *shame* is the emotion that is connected to the evaluation of the self, the coach has to be very sensitive, so that the user is still able to preserve herself (Scheff and Retzinger, 2000; Lewis, 2008).

7.6 Conclusions

For recognising complex, multi-modal behaviours, the current state-of-the-art approach is to either design or learn multi-modal features and train a model and/or fusion algorithm based on annotated samples. A classifier, such as a deep neural network will eventually learn specific constellations that belong a representation of the given problem when fed with a big enough amount of training data. If such a classifier is applied later in a real-time scenario, the predictions might, or might not be fitting to the situation. This is, for example because the model might be over-trained on the shown data, or because new situations apply that have not been seen in the training corpus. The big problem here is, that we can not retrace why the model made decisions in one or another direction, because machine learning classifiers nowadays appear as a "black box" to humans. In some cases, black-box AIs even learn to "cheat" to perform an optimal job of achieving the preprogrammed goals on the training data without representing the more complex implicit desires of the system designers. (Nguyen, Yosinski, and Clune, 2015)

Yet, while computer scientists in most cases work solely on data-driven approaches in the area of social signal processing, many reasoning models exist in other disciplines that aim to explain the complex relationships between the inner state of a person, and the observed behaviours. In this chapter, we investigated dynamic Bayesian networks as meta-fusion for recognisers of "simpler" recognition problems (such as detecting a smile, or a certain posture), expert and domain knowledge can actively be integrated into a model, so existing theories of causal relationships can be represented or new theories may be validated. While we probably are not as much interested in why a machine learning model maps a set of facial features to a smile, or why our audio classifier maps an audio segment to a voice activity (pure data-driven ML approaches are doing very well in recognising these), the overall causal relationships of predictions of such classifiers are not as clear, as even humans often are not consent about the correct interpretation.

In our experiments, dynamic Bayesian networks turn out to be a promising approach as they allow handling uncertainties and inputs ranging from simple features and classification model outputs to external sources such as logs from an interaction management component and other contextual information. DBN parameters may be estimated by learning them from a corpus of data or according to subjective experience or common sense. In other words, DBNs allow combining observations from real data with expert knowledge and theories to build appropriate models for predicting complex behaviours. To complete the toolchain of empowering the human in the model building process, the NOVA tool allows to adapt either manual, or automated annotations (or even annotations that have been created cooperatively between the human and the machine, as described in Chapter 6) to be used directly as learning input for dynamic Bayesian networks. When learning parameters from the data we are further able to find correlations of specific single behaviours and their influence on complex behaviours visually, as we can directly simulate the impact of certain observations on our predictions. Therefore DBNs are a valuable tool for finding critical incidents in the interaction which is especially helpful in scenarios such as social coaching.

Further, we described two exemplary use cases that applied the tool chain proposed in this chapter. To this end, we first introduced a use case for the recognition of *conversational engagement*. Engagement is a complex social behaviour that largely depends on the surrounding context. For example, measurements of engagement in a conversation depend on the interaction dynamics between two or more participants. In this chapter we described the annotation process for our model, an evaluation and comparison with other classification models, and an extension for multiple persons so that interpersonal cues are considered in the recognition and reasoning process. Our experiments showed that the model delivers comparable results to traditional approaches with support vector machines and artificial neural networks, while in our particular case they even outperformed them. All classifiers in our tests have been trained on almost seven hours of data, yet a large scale study remains part of our future work.

As a second use-case, we presented the MARSSI model which relates appraisal- and emotion regulation rules with social signals. It allows defining possible, plausible relations between communicative emotions (e.g. emotional expressions) and sequences of social signals to individual appraisal and regulation strategies. The latter can be triggered by elicited structural emotions, such as *shame*, which was the focus of our work. We used a corpus-based approach to create our social cue models in the context of job interviews. Using MARSSI, we were able to model appraisal and regulation strategies that might occur in an applicant during a job interview. In a debriefing session, we used this knowledge combined with predictions on various abstraction layers for analysing each individual's social cues and for computing confidence values for modelled regulation strategies. An empathic virtual agent in the role of a job interview coach explains the regulation strategy with the highest confidence value. He addresses the possible elicited structural emotion shame, while explaining further details about the detected social cues. To perform this, an emotion simulation model is updated with the final prediction of the regulation classes, as well as event-level abstractions of single social cues that led to the assumptions in the model to create explanatory feedback to the user.

CONTRIBUTIONS

8 CONTRIBUTIONS AND OUTLOOK

"Learn how to see. Realise that everything connects to everything else"

— Leonardo DaVinci.

8.1 Contributions

The main contributions of this thesis can be categorised into conceptual, technical and empirical contributions.

8.1.1 Conceptual Contributions

On a conceptual level, the main contribution is the introduction of a novel approach for including the human in the machine-learning loop in the context of both human social interactions and human-computer interaction (HCI) scenarios. To this end we designed the NOVA framework (see Chapter 5) to act as an interface between the user, the data, and machine learning algorithms. NOVA connects the three aspects in a way that models can improve as the annotation of data gets accelerated, while at the same time, users get a clearer concept of how the models work an when they can be trusted. Our most important concern was to include the strategies in a graphical user interface that allows (non-)expert users to apply the following two strategies:

- 1. For the task of creating machine learning models, which is usually reserved for machine learning experts, this dissertation suggests new concepts and tools that allow human annotators to cooperate with a 'machine annotator'. This way, the human annotator gets a more transparent idea of how and why an ML model predicts certain labels, and when it achieves the stage where it works reliable enough to be trusted. In a two-folded approach human annotators can either work on a session from scratch to create person- or session-depended models, or rely on existing models that have been previously trained on other sessions, as described in Chapter 6.
- 2. When it comes to the interpretation of more complex social signals, such as emotions or social attitudes, the proposed techniques from Chapter 7 allow us to incorporate theories and domain knowledge, in a combination with a data-driven learning algorithm. In Chapter 2 we introduced a literature overview with findings and theories from social sciences and behavioural

psychology about social signals, their relation and interpretation in terms of complex emotions and interpersonal attitudes. To this end we created interfaces to NOVA's annotation database to directly learn the parameters of theory-designed dynamic Bayesian networks. Based on the abstraction of annotation segments, respectively events, and the reasoning capacities of DBNs, decisions of our model become interpretable and understandable to humans. In Section 7.4 we exemplified this process with a model of *conversational engagement*. Our model contains social cues from various modalities, but also cues that can be observed between multiple interlocutors. Additionally we consider temporal, situational and interpersonal context. This approach is especially preferable in scenarios where it is essential to know why a model decided for its output. As second use case, in the context of the job interview simulation described in Section 7.5, we designed the MARSSI model that aims to automatically analyse the parts of an interactions where the participant showed cues associated with emotion regulation strategies to cope with unpleasant *shame* situations. MARSSI is the first computational model that considers both, appraisals and regulations in the analysis process. Parallel to the DBN, the observed social signals and the network's predictions are additionally fed in a theory-based emotion simulation component to generate textual explanations of the machine's predictions.

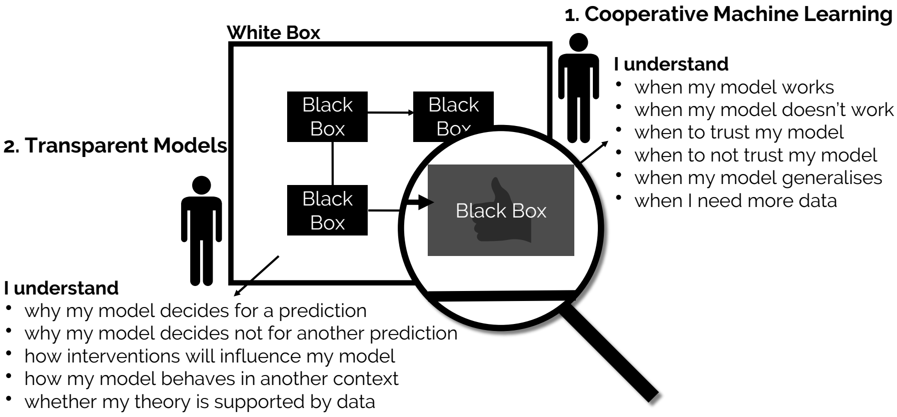

1. Cooperative Machine Learning
White Box
I understand
• when my model works
• when my model doesn't work
• when to trust my model
• when to not trust my model
• when my model generalises
• when I need more data
Black Box
Black Box
Black Box
Black Box
2. Transparent Models
I understand
• why my model decides for a prediction
• why my model decides not for another prediction
• how interventions will influence my model
• how my model behaves in another context
• whether my theory is supported by data

Figure 56.: By applying cooperative machine learning, humans get a clearer idea of how robust their "black box" models work and when they can be trusted in terms of prediction reliability. In a second step, outcomes of such models are combined in dynamic Bayesian networks, which are considered "white box" models. Their reasoning process can be understood intuitively by humans, allowing them to interpret causal relationships and decisions.

Figure 56 illustrates how the two concepts, that are directly integrated in the NOVA tool, help end users to get a more transparent view on their machine learning models in terms of trust, reliability and interpretability.

8.1.2 Technical Contributions

The goal of this thesis is to support researchers, both from artificial intelligence research, as well as from psychology and related research areas in improving their daily work-flow. This is done by providing research tools for all the aspects elaborated on a conceptual level in the last section.

- As a major contribution, a novel open-source annotation and analysis tool named NOVA was implemented during this dissertation. NOVA goes beyond the state of the art by incorporating machine learning tools, a collaborative and flexible workflow and capabilities to learn –and predict with– high-level interpretation models. The NOVA tool has been successfully used in various projects, for example the EU funded projects TARDIS, ARIA-Valuspa, KRISTINA, the BMBF founded projects EmpaT, SenseEmotion, and Glassistant and in various third party projects such as The hybrid Agent MARCO: a multimodal autonomous robotic chess opponent (Becker-Asano et al., 2014) or EmoGest: investigating the impact of emotions on spontaneous co-speech gestures (Bergmann, Böck, and Jaecks, 2014).

NOVA is available for download at `https://github.com/hcmlab/nova`

- To improve the recording of new, natural and rich corpora containing multi-person interactions we extended the SSI Framework with additional sensor devices, such as the Microsoft Kinect sensor, various eye tracking devices and physiological sensors. Also, an implementation of dynamic Bayesian networks, as well as multiple feature sets have been added as plugins to the SSI framework. We created interfaces between NOVA and SSI for a seamless integration of the cooperative machine learning workflow.

SSI is available for download at `https://github.com/hcmlab/ssi`

NOVA and its cooperative machine learning tools implemented in the SSI Framework are publicly available as an open-source project on Github since July 2016. Over 800 commits have been made to the repository.

8.1.3 Empirical Contributions

As a first empirical contribution, new corpora have been created and made available to the research community (also see Chapter 3).

- The NOXI corpus containing over 50 hours of synchronised multi-person and multi-modal data was recorded and annotated for social cues related to engagement, interruptions, valance and arousal. We created the setting and recordings in three different locations and created the tools for the annotation of the corpus. Most of the concepts suggested in this thesis have been empirically tested on this corpus with the help of technical- and user evaluations.

- The MMLI corpus is the first corpus of this richness in different laughter contexts, containing various data sources (motion capture, depth, audio, video, physiological), a large spectrum of captured modalities and that is synchronised across multiple participants.
- The EmpaT shame-eliciting job interview corpus is one of the first corpora that are designed to investigate emotions based on an appraisals and regulations.

Additionally performed multiple technical and user evaluations during this thesis

- We evaluated our cooperative machine learning approach based on the NOXI corpus. To prove the usefulness of the CML approach, we have presented results for a realistic use-case based on a database featuring natural interactions between human dyads. For our experiments in Section 6.3 we selected the task of detecting fillers in speech. Fillers are an important cue if one aims to study turn taking and interruption strategies. A fast and general audio detection system in combination with a linear classification model has been applied to more than 10 hours of natural conversations yielding an average recognition performance of almost 80 % (four classes: speech, breath, filler and silence). In a simulation we proved that labelling efforts can be significantly reduced using the proposed system. If applied in combination with a revision of instances with a low confidence value, manual inspection was reduced to $\frac{5}{8}$ of the database. In our case, this corresponds to a saving of approximately 3.5 hours (5.9 hours instead of 9.4 hours)
- Experiences with different groups of users show that the CML approach was also positively perceived from an end-user's point of view who were impressed by the system's accuracy. The feedback we obtained made us aware of different styles of annotation adopted by the end-user and the machine. While a machine is able to annotate social signals much faster and more consistently than humans can do, human raters still bring a better understanding for the application in which the models to be trained will eventually be applied. Furthermore, human raters do not just look at the behaviours to be labelled, but also reason about the context in which they occur. Being presented with the

results of an automated labelling process might influence human labellers in a positive manner. Nevertheless, one should be aware of the risk that a machine-like style of annotation might not always result in better systems. This is in particular true when social signals are analysed where raters usually disagree on the labels and no objective ground truth can be established. In order to benefit from the complementary skills of machines and human raters, annotation tools like NOVA are needed that aim for a smooth integration of human intelligence and resources.

- We further evaluated our theory-designed dynamic Bayesian network with other state of the art classification models. For the problem for recognising conversational engagement on the NOXI corpus we achieved a Person correlation coeffiecent of .7443 which is considered a strong uphill linear relationship between the gold standard annotations and the predictions and delivers comparable or even superior results to state-of-the-art black-box approaches. By extending our DBN to a multi-person network the results significantly improved ($p < .001$) to a PCC of .7545 which lets us conclude that considering interpersonal cues as context information has a high potential for improving the recognition of complex social signals.

- Besides the evaluations of the concepts proposed in this thesis we carried out studies with the systems they were applied in. The job interview study described in Section 7.5.1.2 demonstrated the capabilities of automated behaviour analysis in interactions with virtual agents. In this study we carried out guided feedback sessions. The transparent visualisation of the automatically detected behaviours in the NOVA tool and the indicator of the detected engagement, turned out to be very useful for the participants. While both, the experiment and control groups were rated better by the career counsellors on the final day of the study, only the EG showed significant improvements on the dimensions in terms of overall performance, recommendation for the job, smiles, eye contact and nervousness. As the goal of any job interview training technique is to increase the user's chances for employment, we consider these results encouraging.

Overall, the empirical findings showed that transparent approaches that include the human on multiple abstraction layers, do not only add value towards classical machine learning as an end in itself, but enrich the value of the predictions, help to reason about problems and theories of the world, and finally and foremost help people when they are embedded in useful applications.

8.2 Outlook on Future Work

> *"Our future is a race between the growing power of technology and the wisdom with which we use it."*
>
> — Stephen Hawking.

Explainable and interpretable techniques that include the human in designing social signal processing systems are still an emerging research area, and even though algorithms are capable of recognising human behaviours with high precision for many tasks today already, with new corpora in various contexts, new sensor devices, and new approaches to interpret such behaviours, we will most probably see a boost in the area in the next decades. This thesis suggests concepts and tools that will help researchers in the years to come to create and analyse new corpora, and improve the understanding of human behaviours and how social cues indicate complex emotions and interpersonal attitudes. Especially given a growing number of researchers are employing techniques to speed up the annotation process, by applied methods like the cooperative machine learning approach suggested in this thesis, the community will gain access to more annotated training samples which is still considered the largest bottleneck for this research area.

8.2.1 Cooperative Machine Learning

In our future work, we plan to extend the current CML workflow by automatically generating recommendations in which order sessions in a database should be processed. Poignant et al. (2016) suggest the use of hierarchical clustering to select prototypical examples and prioritise them during the coding process. However, it is not straightforward to adapt their techniques to continuous recordings. Alternatively, in our case we can make use of the confidence values generated during label prediction. Using the average value the following strategy is conceivable: every time a session is finished, a model is built to predict remaining sessions and pick the one with the lowest score to complete next. This way we ensure that manual efforts get spent on data that has a high potential to improve the learner in the next iteration. Further, another aspect would be to integrate concepts like in the approach of (Zhang, Coutinho, Zhang, et al., 2015a) which are summarised as *dynamic active learning* (DAL) where the most reliable annotators are chosen first to be handed the annotation task to.

In general, the hope is that non machine learning experts get easier access to the tools they need without in-depth knowledge in programming and machine learning. By making the progress of machine learning more transparent with comprehensive annotation strategies, human users not only serve as pure information provider to ML algorithms but rather are part of the learning process, getting information about the current performance quality of their models and about the parts where their model needs to be improved on.

8.2.2 NOVA's User Interface

NOVA's graphical user interface is currently written in C# WPF and therefore limited to the Windows System Platform. This on the one hand comes some advantages, e.g. the use of fast video codecs and certain Windows libraries. On the other hand, this binds users to a single platform. In the future the proposed concepts might be transferred on other platforms, e.g. by converting NOVA to XAMARIN which enables multi-platform usage for Windows, Mac, Android and iOS, or as a Web-application which can be accessed via browser.

8.2.3 Context-sensitive User Modelling

We proposed a tool-chain for building transparent models based on a hybrid approach that combines a theory-based modelling approach and a data-driven approach and exemplified this process with the use cases of "conversational engagement" and "emotion regulation strategies". An interesting addition would be to include additional context information. Imaginable are further inputs such as physiological data from self-tracking gadgets, or environmental context input such the measurements of smart-home and internet of things (IoT) devices. Of course additional complex social signals need to be investigated to clarify the generalisability of the suggested approach. This also includes an extended large-scale evaluation of the model's prediction capabilities, especially compared to existing black-box approaches (e.g. deep end-to-end learning or other purely statistical models).

8.2.4 Distribution of Models

Decentralised blockchain based approaches like singularitynet.io are likely to commercially and scientifically boost the field of SSP, by offering marketplaces for models related to AI. Social signal processing models created with the tools suggested in this thesis also offer a great commercial value for such marketplaces. If we think of virtual or physical social agents that are likely to be part of our lives in the future, a general human-like understanding of the user's attitudes and emotions will give added value to the experience. The consideration of context in user models wherever possible helps interpreting cues correctly. Transparency hereby is of vast importance in terms of trust and comprehensibility.

8.2.5 Explainable Artificial Intelligence

The need of a system to be transparent towards users has been pointed out in this thesis. For modelling complex social phenomena, we introduced a two-folded approach. First, by training "black-box" recognition models with high accuracy, such as support vector machines or artificial neural networks, using the suggested cooperative machine learning approach, users get involved in the machine learning process at an early stage. Already during annotation, they get an impression of cases where their models deliver accurate results, but also where they struggle to perform. This of course does not provide them with any information about the inner processes and decisions of their models, yet visually reveals parts where a human needs to inject new information for the model to improve. When we, as humans, explain our decisions and judgements, we use abstractions of what we consider important. For example, in a statement like: "the child is happy because I told a joke and now it is smiling" we abstract the constellation of action units in the face to the social cue "smile". In the explanation why we think the child is happy it does not matter how we came to the conclusion that the child was smiling - we probably don't even think about it. In such cases using "black-box" machine learning approaches is advantageous. They allow automating tasks that are complex and hard to describe with rules, and they most often deliver a performance improvement over rule-based approaches. What we on the other hand do explain is the influence of the variables "told a joke" and "the child is smiling" on the emotion "happy". When we ask a system why it predicts the child to be happy, it should be able to respond with such with these variables. Therefore, in the second step of our approach, we employ dynamic Bayesian networks for modelling correlations between sequences of single social cues (as results of "black-box" models) and context information. We actively model the structure of a DBN based on expert knowledge and theories, while parameters are learned with statistical methods. When the model is updated in a real-time scenario or simulation with observations, we can visually follow the model's decisions. By being able to comprehend the processes of the model on this abstraction level, we consider the overall model a "white-box" approach. Especially in cases where a model is judging a person or situation, the judgements are only valuable if the system can explain them to other stakeholders.

In our future work we are further interested in investigating methods to be able to "open up" black-box models as well. This would especially be useful for the early cooperative machine learning step. Here, if a model would provide more insights on why it is - or it is not - confident of a prediction, the human annotator would not need to rely on his or her impression and observations, but rather on information given by the model. Especially in the field of image classification, researchers are lately interested in methods to visualise a model's decisions. For example the LRP (Layer-wise Relevance Propagation) toolbox (Lapuschkin et al., 2016; Park et al., 2017; Samek, Wiegand, and Müller, 2017) or the LIME system (Ribeiro, Singh, and Guestrin, 2016) allow visualising relevant parts of an image that had the largest influence on a decision of a neural network in an image classification task. We further plan to integrate algorithms that identify training samples that are similar to the machine's predictions (e.g. as in the explicable-boundary-tree-explainer by Wu et al. (2018)). This way, the machine could even explain the origin of its prediction to illustrate the decisions towards and end user.

APPENDIX

A

APPENDIX A: BODY EXPRESSIVENESS FEATURES

As an addition to Section 2.2.1.3, the formulas for calculating various expressivness values are listed below:

- Energy/Power (EN) represents the dynamic properties of a movement (e.g. weak versus strong). It is calculated from the motion vectors' first derivative in all three dimensions where $\vec{m}()$ is the motion of the specified joint relative to the torso joint and n is the number of frames considered for the calculation.

$$EN = \sqrt{\sum_{i=0}^{n}((\vec{m}(i).x^2 + \vec{m}(i).y^2 + \vec{m}(i).z^2)/3)/n}$$

- Fluidity (FL) differentiates smooth movements from jerky ones. This feature aims to capture the continuity between movements. It is calculated as the sum of the variance (Var) of both hands' motion vectors' norms ($\vec{l}, \vec{r}$) (respectively feet for leg postures).

$$FL = Var(\sum_{i=0}^{n} \vec{l}(i)/n) + Var(\sum_{i=0}^{n} \vec{r}(i)/n)$$

- Spatial extent (SE) is modelled as the space occupied for gesticulation in front of a person. It calculates as the maximum Euclidean distance hands' position (l,r) (respectively feet for leg postures).

$$SE = max(d(\mid r(i) - l(i) \mid))$$

- Overall activation (OA) represents the quantity of the movement (passive versus active). It is calculated as the sum of the motion vectors' norm of both hands (respectively feet for leg postures).

$$OA = \sum_{i=0}^{n} \mid \vec{r}(i) \mid + \mid \vec{l}(i) \mid$$

B

APPENDIX B: INSTRUCTIONS FOR PARTICIPANTS OF THE NOXI CORPUS

B.1 Instructions for the Expert

We are creating a corpus of screen-mediated interactions for the H2020 project ARIA-VALUSPA. In this recording session you are going to discuss the topic that you have chosen with a person interested in it. Please be aware that we are not evaluating your level of expertise on this topic nor the quality of your interaction.

Your goal is to first introduce the topic and then simply follow up the discussion with the other interlocutor. To get started, for example, you could briefly discuss your expertise on the chosen topic or why you are interested in it, or since when you got interested in that topic, etc.

Please keep in mind that you are free to structure the discourse in your own way. The same applies for the whole interaction. For example, and not limited to these, both you and the interlocutor can ask questions, express opinions (e.g. appreciations, disagreement), interrupt each other or even refuse to provide an answer or discuss about a requested subtopic. Please try to be as natural as possible and do not be afraid of expressing your emotions (e.g. anger against a specific point), your opinions (e.g. strong (dis)agreement), or not knowing an answer for a specific question (we are not evaluating your expertise).

During the recording please stand within the area of the markers on the floor. Please turn off or put in silent mode your phone. You will be using headphones and a close talk microphone and your interlocutor will be displayed on the screen in front of you.

In sum, this is the procedure of the recording session:
(1) You start talking about your topic
(2) The session continues (for about 10-30 minutes) with a discussion with the interlocutor
(3) The session ends when either you or the interlocutor is tired/bored/satisfied
(4) Once finished take off the equipment and exit the room.

B.2 Instructions for the Novice

We are creating a corpus of screen-mediated interactions for the H2020 project ARIA-VALUSPA (http://aria-agent.eu). In this recording session you are going to discuss the topic that you have chosen with a person who is an expert on this. Please be aware that we are not evaluating your level of expertise on this topic nor the quality of your interaction.

Your goal is to simply discuss the chosen topic. You will first hear from the other interlocutor a brief introduction, then you are free to structure the interaction in your own way. For example, and not limited to these, both you and the interlocutor can ask questions, express opinions (e.g. appreciations, disagreement), interrupt each other or even refuse to provide an answer. Please try to be as natural as possible and do not be afraid of expressing your emotions (e.g. anger against a specific point), your opinions (e.g. strong (dis)agreement), or not knowing an answer for a specific question (we are not evaluating your expertise).

Please be aware that during the recording at any time you might (or not) receive a phone call or an sms on your mobile phone. This will be controlled by us and it will be part of the recordings. Therefore feel free to answer the call or read the text message when you receive it.

During the recording please stand roughly within the area of the markers on the floor. You will be using headphones and a close talk microphone and your interlocutor will be displayed on the screen in front of you.

In sum, this is the procedure of the recording session:
(1) The other interlocutor starts talking about the topic
(2) The session continues (for about 10-30 minutes) with a discussion with the interlocutor
(3) The session ends when either you or the interlocutor is tired/bored/satisfied
(4) Once finished take off the equipment and simply exit the room.

C

APPENDIX C: EVENT-BASED ONLINE PIPELINE TO UPDATE THE ENGAGEMENT MODEL

As an addition to Section 7.4 we give an example here of how we can employ our model in a real-time seario using an SSI XML Pipeline. For illustration purposes we will limit the example on a Microsoft Kinect 2 sensor to calculate the overall expressiveness of the user's hand movements, as well as paralinguistics based on the use case recognition problem from Chapter 6.

Especially for allowing non-experts to write and edit SSI pipelines with a simple text editor, SSI supports an XML based language, in which each component of a pipeline is represented by an XML element. The components of an element are identified through unique class names and option files can be used to adjust their parameters, e. g. the device id of a sensor or the cut-off frequency of a low-pass filter. Elements are connected via pin names that tell the interpreter from which source(s) a component receives input.

At first we need to register the single SSI components we want to use in our pipeline

```
<register>
<load name="microsoftkinect2" />
<load name="audio" />
<load name="bodyfeatures" />
<load name="opensmile" />
<load name="model" />
<load name="bayesnet" />
...
</register>
```

In this instance we make use of the Microsoft Kinect 2 sensor, which is implemented in the microsoftkinect2.dll plug-in and the microphone which is is implemented in the audio.dll. We add additional plug-ins such as the bodyfeatures.dll which contains algorithms to calculate expressiveness features as described in Section 2.2.1.3, the opensmile.dll which loads the implementation of the opensmile (Eyben, Wöllmer, and Schuller, 2010) library for audio processing, and the model.dll which contains various machine learning algorithms. Additionally we load the bayesnet.dll component which includes the implementation of our bayesian network inference engine.

Next we may set some options of the framework itself, for example we want to show the console, or if we want to wait for the sync command of another pipeline to run distributed pipelines on multiple machines.

```
<framework console="true" sync="true"
```

```
slisten="$(waitforsync)" sport="6666"/>
```

Note that some of the options contain the syntax of $(..). This represents a placeholder for options that may be outsourced to an external option file. In this case if in an external file the variable "waitforsync" is set to "true", the pipeline will wait for a message on port 6666. If it is set to "false" the pipeline will start immediately

Next we need to add our sensors to the pipeline. The Microsoft-Kinect2 sensor is found in the microsoftkinect2 plugin. Besides options, such as the samplerate (sr) and the numbers of persons to be tracked the various providers are defined. A provider might here be for example the rgb image of the sensor which is given the pin "video" or the actual skeleton channel which is named "skel" here.

```
<sensor create="MicrosoftKinect2:kinect"
sr="$(sr)" trackNearestPersons="1">
<output channel="rgb" pin="video"/>
<output channel="depth" pin="depth"/>
<output channel="au" pin="au" />
<output channel="head" pin="head" />
<output channel="skeleton" pin="skel" />
<output channel="face3d" pin="face3d" />
</sensor>
```

The single outputs of the sensor can then be processed. For example the three streams skel, head and face3d are combined in a new skeleton type that contains more information and is compatible to further skeleton processing plugins.

```
<transformer create="SkeletonConverter:skelconvert">
<input pin="skel,head,face3d" frame="1"/>
<output pin="skel_converted"/>
</transformer>
```

The advantage of a combined skeleton type is that algorithms can be applied to outputs from other sensors that provide similar skeleton data. For example a motion capture suit, or a third party depth camera may be converted in the same way, which allows algorithms in further pipeline steps to treat the data analogously.

Our converted skeleton can then for example be transformed by the EnergyMovement component that calculates the movement energy on the raw skeleton. Note that each input frame is calculated but also the last 24 frames are taken into account, so that, given a sample-rate of 25 hz, we calculate the energy on 1 second of movement but return a result for each given frame. The output of this component contains the energy for each joint of the input skeleton.

```
<transformer create="EnergyMovement">
<input pin="skel_converted" frame="1" delta="24"/>
<output pin="energy"/>
</transformer>
```

Additionally we might filter the calculated energy with a Butterworth filter to smooth the signal. Note that the previously calculated stream, that we gave the output pin "energy", is used as input here.

```
<transformer create="Butfilt" norm="true" low="0.1" high="1.0">
<input pin="energy" frame="5" />
<output pin="energy_filtered"/>
</transformer>
```

Finally we want to apply a simple threshold based classification of our current energy feature. Therefore we first select the dimension in our filtered energy stream where the hands can be found, and then create events that are registered in SSI's event system. These events will be fed in the BN later.

```
<transformer create="Selector" indices="6">
<input pin="energy_filtered" frame="1"/>
<output pin="energyHands"/>
</transformer>

<consumer create="ThresTupleEventSender" address="energyhands@body"
classes="LOW, MEDIUM, HIGH" thres="0.0, 0.33, 0.66, 1.0">
<input pin="energyHands" frame="0.5s" />
</consumer>
```

Besides the Kinect2 Sensor we might add additional sensors to our pipeline. An example is a microphone which is implemented in the audio plugin. Analogous to the Kinect2 Sensor it has an output provider, (yet here it is a single one), providing the raw audio signal at a sample-rate of 48000 hz:

```
<sensor create="Audio:mic" option="options/audio" scale="true" sr="48000">
<output channel="audio" pin="mic"/>
</sensor>
```

Next, Mel Frequency Cepstral Coefficients (MFCCS) are calculated on the raw audio stream using the capabilities of the opensmile library which provides a new stream, containing 13 MFCCs as audio features. The features are calculated every 40 ms, so that our new streams has a sample-rate of 25 hz.

```
<transformer create="OSMfccChain" option="../options/mfccdd">
<input pin="mic" frame="0.04s"/>
<output pin="mfccs" />
</transformer>
```

The extracted features are fed into a classifier that loads a pretrained model named speechfillerbreath. An event is created every frame (25 per second) containing the predicted probabilities for the classes that are contained in the model. The model used here contains the classes SPEECH, FILLER; BREATH; and REST (which represents silence). This model was trained using the Cooperative Machine Learning techniques as described in Chapter 6.

```
<object create="Classifier" trainer="$(model:speechfillerbreath)"
address="speechfillerbreath@audio">
<input pin="mfccstream" frame="1" />
</object>
```

Finally, events created by either the simple threshold event component or the classifier can be used to be fed in other components. Here, the Bayesnet component registers to listen to events from both components. Additionally we add the path to the pre-trained Network and set options e.g. to convert the output to a continuous value, and to print the current state of all nodes in the network. Finally the output of the network is send to an event monitor that simply outputs the current value of the user's engagement based on various multi-modal observations.

```
<object  create="Bayesnet"
sname="bnet"
path="BN.xdsl"
print="true"
monitored_nodes="engagement"
continuousOutput="true"
<listen address="speechfillerbreath,energyhands@audio,body"/>
</object>

<object create="EventMonitor:monitor" all="true" title="Bayesnet">
<listen address="@bnet" span="10000"/>
</object>
```

The output of the network of course might be used as an input to another system, e.g. a dialog management system that controls the actions of a virtual agent or social robot.

BIBLIOGRAPHY

Abowd, Gregory D., Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles (1999). "Towards a Better Understanding of Context and Context-Awareness." In: *Handheld and Ubiquitous Computing, First International Symposium, HUC'99, Karlsruhe, Germany, September 27-29, 1999, Proceedings*. Ed. by Hans-Werner Gellersen. Vol. 1707. Lecture Notes in Computer Science. Springer, pp. 304–307 (cit. on p. 136).

Adam, James et al. (1902). *The Republic of Plato: Books VI-X and Indexes*. Vol. 2. University Press (cit. on p. 34).

Alibali, Martha W, Sotaro Kita, and Amanda J Young (2000). "Gesture and the process of speech production: We think, therefore we gesture." In: *Language and cognitive processes* 15.6, pp. 593–613 (cit. on p. 21).

Allwood, Jens (1993). "Feedback in second language acquisition." In: *Adult Language Acquisition. Cross Linguistic Perspectives* 2, pp. 196–236 (cit. on p. 30).

Ambady, Nalini, Mark Hallahan, and Robert Rosenthal (1995). "On judging and being judged accurately in zero-acquaintance situations." In: *Journal of Personality and Social Psychology* 69.3, p. 518 (cit. on p. 54).

Amershi, Saleema, Maya Cakmak, W. Bradley Knox, and Todd Kulesza (2014). "Power to the People: The Role of Humans in Interactive Machine Learning." In: *AI Magazine* 35.4, pp. 105–120 (cit. on pp. 5, 101, 102, 127).

Amershi, Saleema, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Y. Simard, and Jina Suh (2015). "ModelTracker: Redesigning Performance Analysis Tools for Machine Learning." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pp. 337–346 (cit. on p. 103).

Amershi, Saleema, James Fogarty, Ashish Kapoor, and Desney S. Tan (2009). "Overview based example selection in end user interactive concept learning." In: *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, Victoria, BC, Canada, October 4-7, 2009*. Ed. by Andrew D. Wilson and François Guimbretière. ACM, pp. 247–256 (cit. on p. 103).

Argyle, Michael (2013). *Bodily communication*. Routledge (cit. on pp. 16, 44).

Aubrey, Andrew J., A. David Marshall, Paul L. Rosin, Jason Vandeventer, Douglas W. Cunningham, and Christian Wallraven (2013). "Cardiff Conversation Database (CCDb): A Database of Natural

Dyadic Conversations." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2013, Portland, OR, USA, June 23-28, 2013*, pp. 277–282 (cit. on p. 51).

Aytar, Yusuf, Carl Vondrick, and Antonio Torralba (2016). "SoundNet: Learning Sound Representations from Unlabeled Video." In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 892–900 (cit. on p. 153).

Baltrusaitis, Tadas, Peter Robinson, and Louis-Philippe Morency (2016). "OpenFace: An open source facial behavior analysis toolkit." In: *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*, pp. 1–10 (cit. on pp. 118, 124, 153, 170).

Bänninger-Huber, Eva (1996). *Mimik-übertragung-interaktion: die untersuchung affektiver prozesse in der psychotherapie*. Huber (cit. on p. 40).

Bänziger, Tanja, Marcello Mortillaro, and Klaus R Scherer (2012). "Introducing the Geneva Multimodal Expression Corpus for experimental research on emotion perception." In: *Emotion* 12.5, pp. 1161–1179 (cit. on p. 59).

Barrett, Lisa Feldman (2006). "Solving the emotion paradox: Categorization and the experience of emotion." In: *Personality and social psychology review* 10.1, pp. 20–46 (cit. on p. 36).

Barrett, Lisa Feldman (2011). "Was Darwin wrong about emotional expressions?" In: *Current Directions in Psychological Science* 20.6, pp. 400–406 (cit. on p. 36).

Battocchi, Alberto, Fabio Pianesi, and Dina Goren-Bar (2005). "DaFEx: Database of Facial Expressions." In: *Intelligent Technologies for Interactive Entertainment, First International Conference, INTETAIN 2005, Madonna di Campiglio, Italy, November 30 - December 2, 2005, Proceedings*. Ed. by Mark T. Maybury, Oliviero Stock, and Wolfgang Wahlster. Vol. 3814. Lecture Notes in Computer Science. Springer, pp. 303–306 (cit. on p. 51).

Baur, Tobias, Gregor Mehlmann, Ionut Damian, Florian Lingenfelser, Johannes Wagner, Birgit Lugrin, Elisabeth André, and Patrick Gebhard (2015). "Context-Aware Automated Analysis and Annotation of Social Human-Agent Interactions." In: *TiiS* 5.2, 11:1–11:33 (cit. on pp. 4, 83, 131, 161, 164, 167).

Baur, Tobias, Dominik Schiller, and Elisabeth André (2017). "Modeling User's Social Attitude in a Conversational System." In: *Emotions and Personality in Personalized Services - Models, Evaluation and Applications*. Ed. by Marko Tkalcic, Berardina De Carolis, Marco de Gemmis, Ante Odic, and Andrej Kosir. Human-Computer Interaction Series. Springer, pp. 181–199 (cit. on pp. 131, 134).

Beattie, Geoffrey W (1981). "Interruption in conversational interaction, and its relation to the sex and status of the interactants." In: *Linguistics* 19.1-2, pp. 15–36 (cit. on p. 33).

Becker-Asano, Christian, Eduardo Meneses, Nicolas Riesterer, Julien Hué, Christian Dornhege, and Bernhard Nebel (2014). "The hybrid agent MARCO: a multimodal autonomous robotic chess opponent." In: *Proceedings of the second international conference on Human-agent interaction, HAI '14, Tsukuba, Japan, October 29-31, 2014*, pp. 173–176 (cit. on p. 179).

Bekoff, Marc (2013). *Why dogs hump and bees get depressed: The fascinating science of animal intelligence, emotions, friendship, and conservation*. New World Library (cit. on p. 14).

Benecke, Cord (2002). "Mimischer Affektausdruck und Sprachinhalt." PhD thesis. Saarland University (cit. on p. 40).

Bergmann, Kirsten, Ronald Böck, and Petra Jaecks (2014). "EmoGest: Investigating the Impact of Emotions on Spontaneous Co-speech Gestures." In: *Multimodal Corpora: Combining applied and basic research targets*. Ed. by Jens Edlund, Patrizia Paggio, and Dirk Heylen. Reykjavik, Iceland (cit. on p. 179).

Beritelli, Francesco, Salvatore Casale, Alessandra Russo, Salvatore Serrano, and Donato Ettorre (2006). "Speech Emotion Recognition Using MFCCs Extracted from a Mobile Terminal based on ETSI Front End." In: *International Conference on Signal Processing*. Vol. 2 (cit. on p. 111).

Binkofski, Ferdinand and Giovanni Buccino (2006). "The role of ventral premotor cortex in action execution and action understanding." In: *Journal of Physiology-Paris* 99.4, pp. 396–405 (cit. on p. 29).

Birdwhistell, Ray L (1952). *Introduction to kinesics: An annotation system for analysis of body motion and gesture*. Department of State, Foreign Service Institute (cit. on p. 19).

Bishop, Christopher M. and Nasser M. Nasrabadi (2007). "*Pattern Recognition and Machine Learning*." In: *J. Electronic Imaging* 16.4, p. 049901 (cit. on p. 70).

Boholm, Max and Jens Allwood (2010). "Repeated head movements, their function and relation to speech." In: *Proceedings of LREC workshop on multimodal corpora advances in capturing coding and analysing multimodality*. Citeseer, pp. 6–10 (cit. on p. 31).

Bohus, Dan and Eric Horvitz (2009). "Models for Multiparty Engagement in Open-World Dialog." In: *Proceedings of the SIGDIAL 2009 Conference, The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 11-12 September 2009, London, UK*, pp. 225–234 (cit. on p. 44).

Bosma, Wauter and Elisabeth André (2004). "Exploiting emotions to disambiguate dialogue acts." In: *Proceedings of the 9th International Conference on Intelligent User Interfaces, IUI 2004, Funchal, Madeira, Portugal, January 13-16, 2004*, pp. 85–92 (cit. on p. 138).

Bostock, David (2000). *Aristotle's Ethics*. Oxford University Press (cit. on p. 34).

Boutilier, Craig, Nir Friedman, Moisés Goldszmidt, and Daphne Koller (1996). "Context-Specific Independence in Bayesian Networks." In: *UAI '96: Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence, Reed College, Portland, Oregon, USA, August 1-4, 1996*, pp. 115–123 (cit. on p. 140).

Broekens, Joost, Marcel Heerink, and Henk Rosendal (2009). "Assistive social robots in elderly care: a review." In: *Gerontechnology* 8.2 (cit. on p. 4).

Bruce, Vicki and Andy Young (1998). *In the eye of the beholder: the science of face perception.* Oxford University Press (cit. on p. 136).

Brunner, Lawrence J (1979). "Smiles can be back channels." In: *Journal of Personality and Social Psychology* 37.5, p. 728 (cit. on pp. 26, 31).

Bull, Peter (1987). *Posture and Gesture*. Vol. 16. Pergamon Books (cit. on p. 24).

Burbidge, Robert, Jem J. Rowland, and Ross D. King (2007). "Active Learning for Regression Based on Query by Committee." In: *Intelligent Data Engineering and Automated Learning - IDEAL 2007, 8th International Conference, Birmingham, UK, December 16-19, 2007, Proceedings*, pp. 209–218 (cit. on p. 104).

Burkhardt, Felix, Astrid Paeschke, M. Rolfes, Walter F. Sendlmeier, and Benjamin Weiss (2005). "A database of German emotional speech." In: *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*. ISCA, pp. 1517–1520 (cit. on p. 51).

Busso, Carlos, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth Narayanan (2008). "IEMOCAP: interactive emotional dyadic motion capture database." In: *Language Resources and Evaluation* 42.4, pp. 335–359 (cit. on p. 51).

Butz, CJ (2002). "Exploiting contextual independencies in web search and user profiling." In: *Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on*. Vol. 2. IEEE, pp. 1051–1056 (cit. on p. 140).

Cafaro, Angelo, Nadine Glas, and Catherine Pelachaud (2016). "The Effects of Interrupting Behavior on Interpersonal Attitude and Engagement in Dyadic Interactions." In: *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016*, pp. 911–920 (cit. on pp. 33, 34).

Cafaro, Angelo, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres, Catherine Pelachaud, Elisabeth André, and Michel F. Valstar (2017). "The NoXi database: multimodal recordings of mediated novice-expert interactions." In: *Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI 2017, Glasgow, United Kingdom, November 13 - 17, 2017*, pp. 350–359 (cit. on pp. 52, 157).

Calero, Henry H. (1979). *Winning the Negotiation*. New York: Dutton Adult (cit. on p. 23).

Camurri, Antonio, Paolo Coletta, Giovanna Varni, and Simone Ghisio (2007). "Developing Multimodal Interactive Systems with EyesWeb XMI." In: *The Seventh International Conference on New Interfaces for Musical Expression, NIME 2007, New York City, USA, June 6-10, 2007*, pp. 305–308 (cit. on p. 77).

Caridakis, G., A. Raouzaiou, K. Karapouzis, and S. Kollias (2006). "Synthesizing Gesture Expressivity Based on Real Sequences." In: *Workshop on multimodal corpora: from multimodal behaviour theories to usable models, LREC Conference Genoa, Italy* (cit. on p. 24).

Caridakis, George, Ginevra Castellano, Loı c Kessous, Amaryllis Raouzaiou, Lori Malatesta, Stylianos Asteriadis, and Kostas Karpouzis (2007). "Multimodal emotion recognition from expressive faces, body gestures and speech." In: *Artificial Intelligence and Innovations 2007: from Theory to Applications, Proceedings of the 4th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI 2007), 19-21 September 2007, Peania, Athens, Greece*. Ed. by Christos Boukis, Aristodemos Pnevmatikakis, and Lazaros Polymenakos. Vol. 247. IFIP. Springer, pp. 375–388 (cit. on p. 51).

Cashdan, Elizabeth (1998). "Smiles, speech, and body posture: How women and men display sociometric status and power." In: *Journal of Nonverbal Behavior* 22.4, pp. 209–228 (cit. on p. 44).

Castellano, Ginevra, André Pereira, Iolanda Leite, Ana Paiva, and Peter W. McOwan (2009). "Detecting user engagement with a robot companion using task and social interaction-based features." In: *Proceedings of the 11th International Conference on Multimodal Interfaces, ICMI 2009, Cambridge, Massachusetts, USA, November 2-4, 2009*. Ed. by James L. Crowley, Yuri A. Ivanov, Christopher Richard Wren, Daniel Gatica-Perez, Michael Johnston, and Rainer Stiefelhagen. ACM, pp. 119–126 (cit. on p. 147).

Chang, Chih-Chung and Chih-Jen Lin (2011). "LIBSVM: A library for support vector machines." In: *ACM TIST* 2.3, 27:1–27:27 (cit. on p. 59).

Chapman, A. (1983). "Humor and laughter in social interaction and some implications for humor research." In: *Handbook of humor research, Vol. 1*. Ed. by P.E. McGhee and J.H. Goldstein, pp. 135–157 (cit. on p. 61).

Chartrand, Tanya L and John A Bargh (1999). "The chameleon effect: the perception–behavior link and social interaction." In: *Journal of personality and social psychology* 76.6, p. 893 (cit. on p. 29).

Chartrand, Tanya L and Rick Van Baaren (2009). "Human mimicry." In: *Advances in experimental social psychology* 41, pp. 219–274 (cit. on p. 29).

Chen, Nan-Chen, Rafal Kocielnik, Margaret Drouhard, Vanessa Peña-Araya, Jina Suh, Keting Cen, Xiangyi Zheng, and Cecilia R. Aragon

(2016). “Challenges of Applying Machine Learning to Qualitative Coding.” In: *CHI 2016 workshop on Human Centred Machine Learning* (cit. on p. 100).

Cheng, Justin and Michael S. Bernstein (2015). “Flock: Hybrid Crowd-Machine Learning Classifiers.” In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015*, pp. 600–611 (cit. on p. 102).

Ciresan, Dan C., Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber (2012). “Multi-column deep neural network for traffic sign classification.” In: *Neural Networks* 32, pp. 333–338 (cit. on p. 4).

Cohen, Jacob (1960). “A coefficient of agreement for nominal scales.” In: *Educational and psychological measurement* 20.1, pp. 37–46 (cit. on p. 93).

Cohen, Jacob (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum Associates, Hillsdale (cit. on pp. 95, 153).

Comeau, Joey (2014). *a softer world*. URL: `http://asofterworld.com/index.php?id=1186` (cit. on p. 3).

Conati, Cristina (2002). “Probabilistic Assessment of User’s Emotions in Educational Games.” In: *Applied Artificial Intelligence* 16.7-8, pp. 555–575 (cit. on p. 167).

Conati, Cristina and Heather Maclaren (2009). “Modeling User Affect from Causes and Effects.” In: *User Modeling, Adaptation, and Personalization, 17th International Conference, UMAP 2009, formerly UM and AH, Trento, Italy, June 22-26, 2009. Proceedings*, pp. 4–15 (cit. on pp. 138, 167).

Cooke, Martin, Jon Barker, Stuart Cunningham, and Xu Shao (2006). “An audio-visual corpus for speech perception and automatic speech recognition.” In: *The Journal of the Acoustical Society of America* 120.5, pp. 2421–2424 (cit. on p. 51).

Cortes, Corinna and Vladimir Vapnik (1995). “Support-Vector Networks.” In: *Machine Learning* 20.3, pp. 273–297 (cit. on p. 72).

Cowie, Roddy, Ellen Douglas-Cowie, and Cate Cox (2005). “Beyond emotion archetypes: Databases for emotion modelling using neural networks.” In: *Neural Networks* 18.4, pp. 371–388 (cit. on p. 51).

Cowie, Roddy, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder (2000). “’FEELTRACE’: An instrument for recording perceived emotion in real time.” In: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion* (cit. on p. 85).

Cowie, Roddy, Gary McKeown, and Ellen Douglas-Cowie (2012). “Tracing Emotion: An Overview.” In: *IJSE* 3.1, pp. 1–17 (cit. on p. 85).

Cowie, Roddy, Catherine Pelachaud, and Paolo Petta (2010). “Emotion-oriented systems: the Humaine handbook.” In: (cit. on p. 29).

Cronbach, Lee J (1951). “Coefficient alpha and the internal structure of tests.” In: *psychometrika* 16.3, pp. 297–334 (cit. on p. 95).

Dael, Nele, Marcello Mortillaro, and Klaus R Scherer (2012a). "Emotion expression in body action and posture." In: *Emotion* 12.5, p. 1085 (cit. on p. 21).

Dael, Nele, Marcello Mortillaro, and Klaus R Scherer (2012b). "The body action and posture coding system (BAP): Development and reliability." In: *Journal of Nonverbal Behavior* 36.2, pp. 97–121 (cit. on p. 24).

Damasio, Antonio (2010). *Self comes to mind: Constructing the conscious brain*. Vintage (cit. on pp. 34, 35).

Damasio, Antonio R (2001). "Emotion and the human brain." In: *Annals of the New York Academy of Sciences* 935.1, pp. 101–106 (cit. on p. 34).

Damian, Ionut, Tobias Baur, Birgit Lugrin, Patrick Gebhard, Gregor Mehlmann, and Elisabeth André (2015). "Games are Better than Books: In-Situ Comparison of an Interactive Job Interview Game with Conventional Training." In: *Artificial Intelligence in Education - 17th International Conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings*, pp. 84–94 (cit. on p. 167).

Damian, Ionut, Michael Dietz, Frank Gaibler, and Elisabeth André (2016). "Social signal processing for dummies." In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016*, pp. 394–395 (cit. on p. 77).

Delaherche, Emilie, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen (2012). "Interpersonal Synchrony: A Survey of Evaluation Methods across Disciplines." In: *IEEE Trans. Affective Computing* 3.3, pp. 349–365 (cit. on p. 135).

deRosis, Fiorella, Cristiano Castelfranchi, Peter Goldie, and Valeria Carofiglio (2011). "Cognitive Evaluations and Intuitive Appraisals: Can Emotion Models Handle Them Both?" In: ed. by Roddy Cowie, Catherine Pelachaud, and Paolo Petta, pp. 459–481 (cit. on p. 138).

Dibeklioglu, Hamdi, Zakia Hammal, and Jeffrey F. Cohn (2018). "Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding." In: *IEEE J. Biomedical and Health Informatics* 22.2, pp. 525–536 (cit. on p. 51).

Dong, Miaobo and Zengqi Sun (2003). "On human machine cooperative learning control." In: *Proceedings of the 2003 IEEE International Symposium on Intelligent Control*, pp. 81–86 (cit. on p. 100).

Douglas-Cowie, Ellen, Nick Campbell, Roddy Cowie, and Peter Roach (2003). "Emotional speech: Towards a new generation of databases." In: *Speech Communication* 40.1-2, pp. 33–60 (cit. on p. 50).

Douglas-Cowie, Ellen, Roddy Cowie, Cate Cox, Noam Amier, and Dirk Heylen (2008). "The Sensitive Artificial Listner: an induction technique for generating emotionally coloured conversation." In:

LREC Workshop on Corpora for Research on Emotion and Affect. Ed. by L. Devillers, J.-C. Martin, R. Cowie, E. Douglas-Cowie, and A. Batliner. Paris, France: ELRA, pp. 1–4 (cit. on p. 51).

Douglas-Cowie, Ellen, Roddy Cowie, and Marc Schröder (2000). "A New Emotion Database: Considerations, Sources and Scope." In: *ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*. Belfast: Textflow, pp. 39–44 (cit. on p. 51).

Douglas-Cowie, Ellen, Roddy Cowie, et al. (2007). "The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data." In: *Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings*. Ed. by Ana Paiva, Rui Prada, and Rosalind W. Picard. Vol. 4738. Lecture Notes in Computer Science. Springer, pp. 488–500 (cit. on p. 51).

Drummond, Kent (1989). "A backward glance at interruptions." In: *Western Journal of Communication (includes Communication Reports)* 53.2, pp. 150–166 (cit. on p. 33).

Duchenne de Boulogne, GBA (1849). "Recherches faites à l', orde des galvanisine sur l', état de la contractilité et de la sensibilité électromusculaires dans les paralysies des membres supérieurs." In: *CR Acad Sci (Paris)* 29, p. 667 (cit. on p. 26).

Duncan, Starkey (1972). "Some signals and rules for taking speaking turns in conversations." In: *Journal of personality and social psychology* 23.2, p. 283 (cit. on pp. 31, 32).

Duranti, Alessandro and Charles Goodwin (1992). *Rethinking context: Language as an interactive phenomenon*. 11. Cambridge University Press (cit. on p. 134).

Eagly, Alice H. and Shelly Chaiken (1998). "Attitude structure and function." In: *The handbook of social psychology, 4th Edition, Vol. 1*. Ed. by Susan T. Fiske, Daniel T. Gilbert, and Gardner Lindzey. New Yorkk: McGraw-Hill, pp. 269–322 (cit. on p. 42).

Eerekoviae, Aleksandra (2014). "An insight into multimodal databases for social signal processing: acquisition, efforts, and directions." In: *Artif. Intell. Rev.* 42.4, pp. 663–692 (cit. on p. 50).

Efron, David (1941). "Gesture and environment." In: (cit. on p. 20).

Ekman, P. and W. V. Friesen (1969a). "Nonverbal leakage and clues to deception." In: *Psychiatry* 32.1, pp. 88–106 (cit. on pp. 17, 20, 28).

Ekman, Paul (2003). *Emotions revealed: recognizing faces and feelings to improve communication and emotional life*. New York: Times Books (cit. on p. 26).

Ekman, Paul (2009). "Lie catching and microexpressions." In: *The philosophy of deception*, pp. 118–133 (cit. on p. 26).

Ekman, Paul, Richard J Davidson, and Wallace V Friesen (1990). "The Duchenne smile: Emotional expression and brain physiology: II." In: *Journal of personality and social psychology* 58.2, p. 342 (cit. on p. 26).

Ekman, Paul and Wallace V Friesen (1969b). “The repertoire of nonverbal behavior: Categories, origins, usage, and coding.” In: *Semiotica* 1.1, pp. 49–98 (cit. on p. 18).

Ekman, Paul and Wallace V Friesen (1971). “Constants across cultures in the face and emotion.” In: *Journal of personality and social psychology* 17.2, p. 124 (cit. on p. 35).

Ekman, Paul and Wallace V Friesen (1981). “The repertoire of nonverbal behavior: Categories, origins, usage, and coding.” In: *Nonverbal communication, interaction, and gesture*, pp. 57–106 (cit. on pp. 17, 18).

Ekman, Paul, Wallace V Friesen, and Joseph C Hager (1978). “Facial action coding system (FACS).” In: *A technique for the measurement of facial action. Consulting, Palo Alto* 22 (cit. on p. 26).

Eyben, Florian, Klaus R. Scherer, et al. (2016). “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing.” In: *IEEE Trans. Affective Computing* 7.2, pp. 190–202 (cit. on pp. 147, 153, 170).

Eyben, Florian, Felix Weninger, Florian Groß, and Björn W. Schuller (2013). “Recent developments in openSMILE, the munich open-source multimedia feature extractor.” In: *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*. Ed. by Alejandro Jaimes, Nicu Sebe, Nozha Boujemaa, Daniel Gatica-Perez, David A. Shamma, Marcel Worring, and Roger Zimmermann. ACM, pp. 835–838 (cit. on pp. 59, 110, 118).

Eyben, Florian, Martin Wöllmer, and Björn W. Schuller (2010). “Opensmile: the munich versatile and fast open-source audio feature extractor.” In: *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*. Ed. by Alberto Del Bimbo, Shih-Fu Chang, and Arnold W. M. Smeulders. ACM, pp. 1459–1462 (cit. on pp. 77, 193).

Eyben, Florian, Martin Wöllmer, Michel François Valstar, Hatice Gunes, Björn W. Schuller, and Maja Pantic (2011). “String-based audiovisual fusion of behavioural events for the assessment of dimensional affect.” In: *Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), Santa Barbara, CA, USA, 21-25 March 2011*. IEEE Computer Society, pp. 322–329 (cit. on p. 51).

Fails, Jerry Alan and Dan R. Olsen Jr. (2003). “Interactive Machine Learning.” In: *Proceedings of the 8th International Conference on Intelligent User Interfaces*. IUI '03. Miami, Florida, USA: ACM, pp. 39–45 (cit. on p. 101).

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin (2008). “LIBLINEAR: A Library for Large Linear Classification.” In: *Journal of Machine Learning Research* 9, pp. 1871–1874 (cit. on p. 111).

Feldman Barrett, Lisa and James A Russell (1998). “Independence and bipolarity in the structure of current affect.” In: *J. Pers. Soc. Psychol.* 74.4, p. 967 (cit. on p. 36).

Fisher, Ronald A (1915). “Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population.” In: *Biometrika* 10.4, pp. 507–521 (cit. on pp. 153, 160).

Fisher, Ronald A (1936). “The use of multiple measurements in taxonomic problems.” In: *Annals of human genetics* 7.2, pp. 179–188 (cit. on p. 140).

Fleiss, Joseph L and Jacob Cohen (1973). “The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability.” In: *Educational and psychological measurement* 33.3, pp. 613–619 (cit. on p. 93).

Flutura, Simon, Johannes Wagner, Florian Lingenfelser, Andreas Seiderer, and Elisabeth André (2016). “MobileSSI: asynchronous fusion for social signal interpretation in the wild.” In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016*, pp. 266–273 (cit. on p. 77).

Foa, Uriel G (1961). “Convergences in the analysis of the structure of interpersonal behavior.” In: *Psychological Review* 68.5, p. 341 (cit. on p. 44).

Fontaine, Johnny JR, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth (2007). “The world of emotions is not two-dimensional.” In: *Psychological science* 18.12, pp. 1050–1057 (cit. on p. 36).

Fontaine, Johnny JR, Klaus R Scherer, and Cristina Soriano (2013). *Components of emotional meaning: A sourcebook*. OUP Oxford (cit. on p. 39).

Freedman, Norbert and Stanley P Hoffman (1967). “Kinetic behavior in altered clinical states: approach to objective analysis of motor behavior during clinical interviews.” In: *Percept. Mot. Skills* (cit. on p. 20).

Frijda, Nico H (1986). “The emotions: Studies in emotion and social interaction.” In: *Paris: Maison de Sciences de l’Homme* (cit. on pp. 41, 42).

Frijda, Nico H (2010). “Impulsive action and motivation.” In: *Biological psychology* 84.3, pp. 570–579 (cit. on p. 39).

Gallaher, Peggy E (1992). “Individual differences in nonverbal behavior: Dimensions of style.” In: *Journal of Personality and Social Psychology* 63.1, p. 133 (cit. on p. 25).

Ganchev, Todor, Nikos Fakotakis, and George Kokkinakis (2005). “Comparative evaluation of various MFCC implementations on the speaker verification task.” In: *in Proc. of the SPECOM-2005*, pp. 191–194 (cit. on p. 111).

Gardner, Rod (2001). *When listeners talk: Response tokens and listener stance*. Vol. 92. John Benjamins Publishing (cit. on p. 30).

Gatica-Perez, Daniel (2009). "Automatic nonverbal analysis of social interaction in small groups: A review." In: *Image Vision Comput.* 27.12, pp. 1775–1787 (cit. on p. 136).

Gebhard, Patrick (2005). "ALMA: a layered model of affect." In: *4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2005), July 25-29, 2005, Utrecht, The Netherlands*, pp. 29–36 (cit. on p. 167).

Gebhard, Patrick, Michael Kipp, Martin Klesen, and Thomas Rist (2003). "Adding the Emotional Dimension to Scripting Character Dialogues." In: *Intelligent Agents, 4th International Workshop, IVA 2003, Kloster Irsee, Germany, September 15-17, 2003, Proceedings*, pp. 48–56 (cit. on p. 167).

Gebhard, Patrick, Gregor Mehlmann, and Michael Kipp (2012). "Visual SceneMaker - a tool for authoring interactive virtual characters." In: *J. Multimodal User Interfaces* 6.1-2, pp. 3–11 (cit. on p. 158).

Gebhard, Patrick, Tanja Schneeberger, Tobias Baur, and Elisabeth André (2018). "MARSSI: Model of Appraisal, Regulation, and Social Signal Interpretation." In: *17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS2018), At Stockholm, Sweden* (cit. on pp. 62, 167).

Girard, Jeffrey M (2014). "CARMA: Software for continuous affect rating and media annotation." In: *Journal of Open Research Software* 2.1, e5 (cit. on p. 85).

Girard, Jeffrey M and Aidan G C Wright (2016). "DARMA: Dual Axis Rating and Media Annotation." In: *Manuscript submitted for publication* (cit. on p. 85).

Glas, Nadine and Catherine Pelachaud (2014). "Politeness versus Perceived Engagement: an Experimental Study." In: *The 11th International Workshop on Natural Language Processing and Cognitive Science*. Venice, Italy (cit. on p. 147).

Glas, Nadine and Catherine Pelachaud (2015). "Definitions of engagement in human-agent interaction." In: *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015, Xi'an, China, September 21-24, 2015*, pp. 944–949 (cit. on p. 45).

Goleman, Daniel (1998). *Working with emotional intelligence*. Bantam (cit. on p. 42).

Goodall, Noah J (2016). "Can you program ethics into a self-driving car?" In: *IEEE Spectrum* 53.6, pp. 28–58 (cit. on p. 5).

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio (2014). "Generative Adversarial Nets." In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2672–2680 (cit. on p. 50).

Gray, Jeffrey Alan and Neil McNaughton (2003). *The neuropsychology of anxiety: An enquiry into the function of the septo-hippocampal system*. 33. Oxford university press (cit. on p. 39).

Greenwald, Anthony G and Mahzarin R Banaji (1995). "Implicit social cognition: attitudes, self-esteem, and stereotypes." In: *Psychol. Rev.* 102.1, p. 4 (cit. on p. 43).

Grimm, Michael, Kristian Kroschel, and Shrikanth Narayanan (2008). "The Vera am Mittag German audio-visual emotional speech database." In: *Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, ICME 2008, June 23-26 2008, Hannover, Germany*, pp. 865–868 (cit. on p. 51).

Gross, James J (2002). "Emotion regulation: Affective, cognitive, and social consequences." In: *Psychophysiology* 39.3, pp. 281–291 (cit. on p. 38).

Gunderman, Richard B (2011). "Emotional intelligence." In: *Journal of the American College of Radiology* 8.5, pp. 298–299 (cit. on p. 42).

Gunes, Hatice and Maja Pantic (2010). "Dimensional Emotion Prediction from Spontaneous Head Gestures for Interaction with Sensitive Artificial Listeners." In: *Intelligent Virtual Agents, 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings*. Ed. by Jan M. Allbeck, Norman I. Badler, Timothy W. Bickmore, Catherine Pelachaud, and Alla Safonova. Vol. 6356. Lecture Notes in Computer Science. Springer, pp. 371–377 (cit. on p. 51).

Gunning, David (2017). "Explainable artificial intelligence (xai)." In: *Defense Advanced Research Projects Agency (DARPA)* (cit. on p. 5).

Hadar, Uri, Timothy J Steiner, Ewan C Grant, and F Clifford Rose (1984). "The timing of shifts of head postures during conservation." In: *Human Movement Science* 3.3, pp. 237–245 (cit. on p. 32).

Hall, Lynne E., Sarah Woods, Ruth Aylett, Lynne Newall, and Ana Paiva (2005). "Achieving Empathic Engagement Through Affective Interaction with Synthetic Characters." In: *Affective Computing and Intelligent Interaction, First International Conference, ACII 2005, Beijing, China, October 22-24, 2005, Proceedings*, pp. 731–738 (cit. on p. 44).

Hammer, Torild (2000). "Mental health and social exclusion among unemployed youth in Scandinavia. A comparative study." In: *International journal of social welfare* 9.1, pp. 53–63 (cit. on p. 161).

Han, Wenjing, Eduardo Coutinho, Huabin Ruan, Haifeng Li, Björn Schuller, Xiaojie Yu, and Xuan Zhu (2016). "Semi-Supervised Active Learning for Sound Classification in Hybrid Learning Environments." In: *PLOS ONE* 11.9, pp. 1–23 (cit. on pp. 104, 105).

Hantke, Simone, Florian Eyben, Tobias Appel, and Björn W. Schuller (2015). "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing." In: *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015,* *Xi'an, China, September 21-24, 2015*. IEEE Computer Society, pp. 891–897 (cit. on pp. 85, 102).

Harré, Rom (1986). *The social construction of emotions*. Blackwell (cit. on p. 36).

Hartmann, Björn, Maurizio Mancini, and Catherine Pelachaud (2005). "Implementing Expressive Gesture Synthesis for Embodied Conversational Agents." In: *Gesture in Human-Computer Interaction and Simulation, 6th International Gesture Workshop, GW 2005, Berder Island, France, May 18-20, 2005, Revised Selected Papers*. Ed. by Sylvie Gibet, Nicolas Courty, and Jean-François Kamp. Vol. 3881. Lecture Notes in Computer Science. Springer, pp. 188–199 (cit. on p. 24).

Heldner, Mattias and Jens Edlund (2010). "Pauses, gaps and overlaps in conversations." In: *Journal of Phonetics* 38.4, pp. 555–568 (cit. on pp. 33, 157).

Heldner, Mattias, Anna Hjalmarsson, and Jens Edlund (2013). "Backchannel relevance spaces." In: *Nordic Prosody XI, Tartu, Estonia*. Peter Lang Publishing Group, pp. 137–146 (cit. on p. 31).

Hinton, Geoffrey E., Simon Osindero, and Yee Whye Teh (2006). "A Fast Learning Algorithm for Deep Belief Nets." In: *Neural Computation* 18.7, pp. 1527–1554 (cit. on p. 74).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory." In: *Neural Computation* 9.8, pp. 1735–1780 (cit. on p. 75).

Holm, Sture (1979). "A simple sequentially rejective multiple test procedure." In: *Scandinavian journal of statistics*, pp. 65–70 (cit. on p. 165).

Hoque, Mohammed (Ehsan), Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard (2013). "MACH: my automated conversation coach." In: *The 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '13, Zurich, Switzerland, September 8-12, 2013*. Ed. by Friedemann Mattern, Silvia Santini, John F. Canny, Marc Langheinrich, and Jun Rekimoto. ACM, pp. 697–706 (cit. on p. 4).

Hume, David (2012). "Emotions and moods." In: *Organizational behavior*, pp. 258–297 (cit. on p. 35).

Iacoboni, Marco (2009). "Imitation, empathy, and mirror neurons." In: *Annual review of psychology* 60, pp. 653–670 (cit. on p. 30).

Iwasaki, Shoichi (1997). "The Northridge earthquake conversations: The floor structure and the 'loop'sequence in Japanese conversation." In: *Journal of Pragmatics* 28.6, pp. 661–693 (cit. on p. 31).

James, William T. (1932). "A Study of the Expression of Bodily Posture." In: *The Journal of General Psychology* 7.2, pp. 405–437 (cit. on p. 22).

Jiang, Bihan, Michel François Valstar, and Maja Pantic (2011). "Action unit detection using sparse appearance descriptors in space-time video volumes." In: *Ninth IEEE International Conference on Auto-*

matic Face and Gesture Recognition (FG 2011), Santa Barbara, CA, USA, 21-25 March 2011, pp. 314–321 (cit. on p. 51).

Kamar, Ece, Severin Hacker, and Eric Horvitz (2012). "Combining human and machine intelligence in large-scale crowdsourcing." In: *International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2012, Valencia, Spain, June 4-8, 2012 (3 Volumes)*. Ed. by Wiebe van der Hoek, Lin Padgham, Vincent Conitzer, and Michael Winikoff. IFAAMAS, pp. 467–474 (cit. on p. 102).

Kang, Sin-Hwa, Jonathan Gratch, Candy L. Sidner, Ron Artstein, Lixing Huang, and Louis-Philippe Morency (2012). "Towards building a virtual counselor: modeling nonverbal behavior during intimate self-disclosure." In: *International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2012, Valencia, Spain, June 4-8, 2012 (3 Volumes)*. Ed. by Wiebe van der Hoek, Lin Padgham, Vincent Conitzer, and Michael Winikoff. IFAAMAS, pp. 63–70 (cit. on p. 4).

Kasap, Zerrin, Maher Ben Moussa, Parag Chaudhuri, and Nadia Magnenat-Thalmann (2009). "Making Them Remember - Emotional Virtual Characters with Memory." In: *IEEE Computer Graphics and Applications* 29.2, pp. 20–29 (cit. on p. 44).

Keerthi, S. Sathiya and Chih-Jen Lin (2003). "Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel." In: *Neural Computation* 15.7, pp. 1667–1689 (cit. on p. 73).

Keltner, Dacher (1995). "Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame." In: *Journal of personality and social psychology* 68.3, p. 441 (cit. on p. 134).

Kendon, Adam (1967). "Some functions of gaze-direction in social interaction." In: *Acta psychologica* 26, pp. 22–63 (cit. on p. 31).

Kendon, Adam (1972). "Some relationships between body motion and speech." In: *Studies in dyadic communication* 7.177, p. 90 (cit. on p. 32).

Kendon, Adam (1980). "Gesticulation and speech: Two aspects of the process of utterance." In: *The relationship of verbal and nonverbal communication* 25.1980, pp. 207–227 (cit. on p. 21).

Kendon, Adam, Thomas A Sebeok, and Jean Umiker-Sebeok (1981). *Nonverbal communication, interaction, and gesture: selections from Semiotica*. Vol. 41. Walter de Gruyter (cit. on pp. 14, 19).

Kennedy, Lyndon and Daniel P. W. Ellis (2004). "Laughter Detection in Meetings." In: *Proc. NIST Meeting Recognition Workshop*. Montreal, pp. 118–121 (cit. on p. 111).

Keppens, Jeroen (2012). "Argument diagram extraction from evidential Bayesian networks." In: *Artif. Intell. Law* 20.2, pp. 109–143 (cit. on p. 132).

Kim, Bongjun and Bryan Pardo (2017). "I-SED: An Interactive Sound Event Detector." In: *Proceedings of the 22nd International Conference* *on Intelligent User Interfaces, IUI 2017, Limassol, Cyprus, March 13-16, 2017*, pp. 553–557 (cit. on p. 102).

Kipp, Michael (2013). "ANVIL: The Video Annotation Research Tool." In: *Handbook of Corpus Phonology*. Oxford, UK: Oxford University Press (cit. on p. 85).

Kishore, Krishna K.V. and Krishna P. Satish (2013). "Emotion recognition in speech using MFCC and wavelet features." In: *International Conference on Advance Computing Conference (IACC)*, pp. 842–847 (cit. on p. 111).

Kistler, Felix, Birgit Endrass, Ionut Damian, Chi Tai Dang, and Elisabeth André (2012). "Natural interaction with culturally adaptive virtual characters." In: *J. Multimodal User Interfaces* 6.1-2, pp. 39–47 (cit. on pp. 59, 147).

Kitayama, Shinobu, Hazel Rose Markus, and Masaru Kurokawa (2000). "Culture, emotion, and well-being: Good feelings in Japan and the United States." In: *Cognition & Emotion* 14.1, pp. 93–124 (cit. on p. 36).

Knapp, Mark L, Judith A Hall, and Terrence G Horgan (2013). *Nonverbal communication in human interaction*. Cengage Learning (cit. on pp. 18, 26, 28).

Knox, Mary Tai and Nikki Mirghafori (2007). "Automatic laughter detection using neural networks." In: *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*. ISCA, pp. 2973–2976 (cit. on p. 111).

Kobayashi, Hiromi and Shiro Kohshima (2008). "Evolution of the human eye as a device for communication." In: *Primate origins of human cognition and behavior*. Springer, pp. 383–401 (cit. on p. 27).

Koiter, Joost R (2006). "Visualizing inference in Bayesian networks." In: *Delft University of Technology* (cit. on pp. 133, 149).

Kotsiantis, Sotiris B. (2007). "Supervised Machine Learning: A Review of Classification Techniques." In: *Emerging Artificial Intelligence Applications in Computer Engineering - Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pp. 3–24 (cit. on p. 74).

Krauss, Robert M, Yihsiu Chen, and Rebecca F Gotfexnum (2000). "13 Lexical gestures and lexical access: a process model." In: *Language and gesture* 2, p. 261 (cit. on p. 20).

Lacave, Carmen, Manuel Luque, and Francisco Javier Diez (2007). "Explanation of Bayesian Networks and Influence Diagrams in Elvira." In: *IEEE Trans. Systems, Man, and Cybernetics, Part B* 37.4, pp. 952–965 (cit. on p. 133).

Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data." In: *biometrics*, pp. 159–174 (cit. on p. 94).

Lapuschkin, Sebastian, Alexander Binder, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek (2016). "The LRP Toolbox for Artificial Neural Networks." In: *Journal of Machine Learning Research* 17, 114:1–114:5 (cit. on p. 185).

Lazarus, Richard S and Elizabeth Alfert (1964). "Short-circuiting of threat by experimentally altering cognitive appraisal." In: *The Journal of Abnormal and Social Psychology* 69.2, p. 195 (cit. on p. 39).

Lazarus, Richard S and Susan Folkman (1984). "Coping and adaptation." In: *The handbook of behavioral medicine*, pp. 282–325 (cit. on p. 39).

Le Cun, Yann, D Touresky, G Hinton, and T Sejnowski (1988). "A theoretical framework for back-propagation." In: *The Connectionist Models Summer School*. Vol. 1, pp. 21–28 (cit. on p. 73).

Lee, Chul Min, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan (2004). "Emotion recognition based on phoneme classes." In: *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*. ISCA (cit. on p. 111).

Levow, Gina-Anne and Susan Duncan (2012). "Contrasting Cues to Verbal and Non-Verbal Backchannels in Multi-lingual Dyadic Rapport." In: *INTERSPEECH*, pp. 835–838 (cit. on p. 31).

Lewis, Michael (2008). "Self-Conscious Emotions: Embarrassment, Pride, Shame, and Guilt." In: *Handbook of Emotions*. Ed. by Michael Lewis, Jeannette M. Haviland-Jones, and Lisa Feldmann Barrett. New York: The Guilford Press, pp. 742–756 (cit. on p. 172).

Lieberman, Philip, Shirley Fecteau, Hugo Théoret, Ricardo R Garcia, Francisco Aboitiz, Ann MacLarnon, Robin Melrose, Tobias Riede, Ian Tattersall, and Philip Lieberman (2007). "The evolution of human speech: Its anatomical and neural bases." In: *Current Anthropology* 48.1, pp. 39–66 (cit. on p. 28).

Lingenfelser, Florian, Johannes Wagner, and Elisabeth André (2011). "A systematic discussion of fusion techniques for multi-modal affect recognition tasks." In: *Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, Alicante, Spain, November 14-18, 2011*. Ed. by Hervé Bourlard, Thomas S. Huang, Enrique Vidal, Daniel Gatica-Perez, Louis-Philippe Morency, and Nicu Sebe. ACM, pp. 19–26 (cit. on p. 117).

Lingenfelser, Florian, Johannes Wagner, Elisabeth André, Gary McKeown, and William Curran (2014). "An Event Driven Fusion Approach for Enjoyment Recognition in Real-time." In: *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*. Ed. by Kien A. Hua, Yong Rui, Ralf Steinmetz, Alan Hanjalic, Apostol Natsev, and Wenwu Zhu. ACM, pp. 377–386 (cit. on p. 117).

Lotfian, Reza and Carlos Busso (2017). "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings." In: *IEEE Transactions on Affective Computing* PP.99, pp. 1–1 (cit. on p. 103).

Lugrin, Birgit, Julian Frommel, and Elisabeth André (2018). "Combining a Data-Driven and a Theory-Based Approach to Generate Culture-Dependent Behaviours for Virtual Characters." In: *Advances in Culturally-Aware Intelligent Systems and in Cross-Cultural Psychological Studies*. Ed. by Colette Faucher. Cham: Springer International Publishing, pp. 111–142 (cit. on p. 139).

MacDonald, Robert (2008). "Disconnected youth? Social exclusion, the 'Underclass'& economic marginality." In: *Social Work & Society* 6.2, pp. 236–248 (cit. on p. 161).

MacWhinney, Brian (2007). "The TalkBank Project." In: *Creating and digitizing language corpora*. Springer, pp. 163–180 (cit. on p. 52).

Mayer, John D, Peter Salovey, and David R Caruso (2004). "Emotional Intelligence: Theory, Findings, and Implications"." In: *Psychological inquiry* 15.3, pp. 197–215 (cit. on p. 42).

Mayor, Oscar, Quim Llimona, Marco Marchini, Panagiotis Papiotis, and Esteban Maestre (2013). "repoVizz: a framework for remote storage, browsing, annotation, and exchange of multi-modal data." In: *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, pp. 415–416 (cit. on p. 85).

McCrae, Robert R. and Jr. Costa Paul T. (1997). "Personality trait structure as a human universal." In: *Am. Psychol.* 52.5, pp. 509–516 (cit. on p. 58).

McKeown, Gary, William Curran, Johannes Wagner, Florian Lingenfelser, and Elisabeth André (2015). "The Belfast storytelling database: A spontaneous social interaction database with laughter focused annotation." In: *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015, Xi'an, China, September 21-24, 2015*. IEEE Computer Society, pp. 166–172 (cit. on pp. 52, 54).

McKeown, Gary and Ian Sneddon (2014). "Modeling continuous self-report measures of perceived emotion using generalized additive mixed models." In: *Psychol. Methods* 19.1, p. 155 (cit. on p. 84).

McKeown, Gary, Michel François Valstar, Roderick Cowie, and Maja Pantic (2010). "The SEMAINE corpus of emotionally coloured character interactions." In: *Proceedings of the 2010 IEEE International Conference on Multimedia and Expo, ICME 2010, 19-23 July 2010, Singapore*. IEEE Computer Society, pp. 1079–1084 (cit. on p. 51).

McNeill, David (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press (cit. on p. 20).

McNeill, David (2006). "Gesture: a psycholinguistic approach." In: *The encyclopedia of language and linguistics, Elsevier Amsterdam; Boston* (cit. on pp. 20, 21).

Mehlmann, Gregor Ulrich and Elisabeth André (2012). "Modeling multimodal integration with event logic charts." In: *International Conference on Multimodal Interaction, ICMI '12, Santa Monica, CA, USA, October 22-26, 2012*. Ed. by Louis-Philippe Morency, Dan Bohus, Hamid K. Aghajan, Justine Cassell, Anton Nijholt, and Julien Epps. ACM, pp. 125–132 (cit. on p. 158).

Mehrabian, Albert (1969). "Significance of posture and position in the communication of attitude and status relationships." In: *Psychological Bulletin* 71.5, p. 359 (cit. on pp. 22, 24).

Mehrabian, Albert et al. (1971). *Silent messages*. Vol. 8. Wadsworth Belmont, CA (cit. on p. 13).

Mehrabian, Albert and James A Russell (1974). *An approach to environmental psychology.* the MIT Press (cit. on p. 36).

Messinger, Daniel S, Alan Fogel, and K Laurie Dickson (2001). "All smiles are positive, but some smiles are more positive than others." In: *Developmental Psychology* 37.5, p. 642 (cit. on p. 26).

Metallinou, Angeliki and Shrikanth Narayanan (2013). "Annotation and processing of continuous emotional attributes: Challenges and opportunities." In: *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013, Shanghai, China, 22-26 April, 2013*, pp. 1–8 (cit. on p. 84).

Meudt, Sascha, Lutz Bigalke, and Friedhelm Schwenker (2012). "Atlas - Annotation tool using partially supervised learning and multiview co-learning in human-computer-interaction scenarios." In: *11th International Conference on Information Science, Signal Processing and their Applications, ISSPA 2012, Montreal, QC, Canada, July 2-5, 2012*, pp. 1309–1312 (cit. on p. 86).

Mlodinow, Leonard (2012). *Subliminal: How your unconscious mind rules your behavior*. Vintage, p. 117 (cit. on p. 13).

Molcho, Samy (2001). *Alles über Körpersprache*. München: Willhelm Goldmann Verlag (cit. on p. 22).

Montepare, Joann, Elissa Koff, Deborah Zaitchik, and Marilyn Albert (1999). "The use of body movements and gestures as cues to emotions in younger and older adults." In: *Journal of Nonverbal Behavior* 23.2, pp. 133–152 (cit. on p. 22).

Moon, Todd K (1996). "The expectation-maximization algorithm." In: *IEEE Signal processing magazine* 13.6, pp. 47–60 (cit. on p. 151).

Morency, Louis-Philippe (2010). "Modeling Human Communication Dynamics [Social Sciences]." In: *IEEE Signal Process. Mag.* 27.5, pp. 112–116 (cit. on p. 135).

Moresi, Sofie Michèle Joseph (2009). "Preparing for action: a behavioral and pupillometric study." In: (cit. on p. 27).

Morris, Desmond (1997). *Bodytalk. Körpersprache, Gesten und Gebärden*. München: Heyne Verlag (cit. on p. 23).

Morris, William N and Nora P Reilly (1987). "Toward the self-regulation of mood: Theory and research." In: *Motivation and emotion* 11.3, pp. 215–249 (cit. on p. 35).

Moser, Ulrich and Ilka von Zeppelin (2005). In: chap. Die Entwicklung des Affektsystems, pp. 161–221 (cit. on pp. 41, 167).

Moskowitz, Debbie S (1993). "Dominance and friendliness: On the interaction of gender and situation." In: *Journal of Personality* 61.3, pp. 387–409 (cit. on p. 44).

Murata, Kumiko (1994). "Intrusive or co-operative? A cross-cultural study of interruption." In: *Journal of Pragmatics* 21.4, pp. 385–400 (cit. on pp. 33, 34).

Murphy, Kevin Patrick and Stuart Russell (2002). *Dynamic bayesian networks: representation, inference and learning*. University of California, Berkeley (cit. on p. 137).

Nathanson, Donald L. (1994). *Shame and Pride: Affect, Sex, and the Birth of the Self*. WW Norton & Company (cit. on pp. 40, 41, 169).

Navarro, Joe and Marvin Karlins (2008). *What Every BODY is Saying: An Ex-FBI Agent's Guide to Speed-Reading People*. New York: William Morrow Paperbacks (cit. on p. 23).

Neiberg, Daniel, Kjell Elenius, and Kornel Laskowski (2006). "Emotion recognition in spontaneous speech using GMMs." In: *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*. ISCA (cit. on p. 111).

Nguyen, Anh Mai, Jason Yosinski, and Jeff Clune (2015). "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 427–436 (cit. on p. 173).

Nicolaou, Mihalis A., Hatice Gunes, and Maja Pantic (2010). "Audio-Visual Classification and Fusion of Spontaneous Affective Data in Likelihood Space." In: *20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23-26 August 2010*. IEEE Computer Society, pp. 3695–3699 (cit. on p. 51).

Niewiadomski, Radoslaw, Maurizio Mancini, Tobias Baur, Giovanna Varni, Harry J. Griffin, and Min S. H. Aung (2013). "MMLI: Multimodal Multiperson Corpus of Laughter in Interaction." In: *Human Behavior Understanding - 4th International Workshop, HBU 2013, Barcelona, Spain, October 22, 2013. Proceedings*, pp. 184–195 (cit. on p. 61).

Niewiadomski, Radoslaw, Maurizio Mancini, Giovanna Varni, Gualtiero Volpe, and Antonio Camurri (2016). "Automated Laughter Detection From Full-Body Movements." In: *IEEE Trans. Human-Machine Systems* 46.1, pp. 113–123 (cit. on p. 62).

Ortony, Andrew, Gerald L. Clore, and Allan Collins (1988). *The Cognitive Structure of Emotions*. Cambridge, MA: Cambridge University Press (cit. on p. 167).

Ortony, Andrew, Gerald L Clore, and Allan Collins (1990). *The cognitive structure of emotions*. Cambridge university press (cit. on pp. 36, 37).

Oviatt, Sharon L., Kevin Hang, Jianlong Zhou, and Fang Chen (2015). "Spoken Interruptions Signal Productive Problem Solving and Domain Expertise in Mathematics." In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015*, pp. 311–318 (cit. on p. 33).

Panagakis, Yiannis, Ognjen Rudovic, and Maja Pantic (2018). "Learning for Multi-modal and Context-sensitive Interfaces." In: *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition* 2, in press (cit. on p. 134).

Pantic, Maja, Anton Nijholt, Alex Pentland, and Thomas S. Huang (2008). "Human-Centred Intelligent Human-Computer Interaction (HCI^2): how far are we from attaining it?" In: *IJAACS* 1.2, pp. 168–187 (cit. on p. 65).

Pantic, Maja, Alex Pentland, Anton Nijholt, and Thomas S. Huang (2007). "Human Computing and Machine Understanding of Human Behavior: A Survey." In: *Artifical Intelligence for Human Computing, ICMI 2006 and IJCAI 2007 International Workshops, Banff, Canada, November 3, 2006, Hyderabad, India, January 6, 2007, Revised Seleced and Invited Papers*. Ed. by Thomas S. Huang, Anton Nijholt, Maja Pantic, and Alex Pentland. Vol. 4451. Lecture Notes in Computer Science. Springer, pp. 47–71 (cit. on p. 50).

Park, Dong Huk, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach (2017). "Attentive Explanations: Justifying Decisions and Pointing to the Evidence (Extended Abstract)." In: *CoRR* abs/1711.07373. arXiv: `1711.07373` (cit. on p. 185).

Patel, Vishal M, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa (2015). "Visual domain adaptation: A survey of recent advances." In: *IEEE signal processing magazine* 32.3, pp. 53–69 (cit. on p. 136).

Patterson, Miles L (1982). "A sequential functional model of nonverbal exchange." In: *Psychological review* 89.3, p. 231 (cit. on p. 29).

Pearl, Judea (1985). "Bayesian networks." In: *Department of Statistics, UCLA* (cit. on p. 139).

Pearson, Karl (1895). "Note on regression and inheritance in the case of two parents." In: *Proceedings of the Royal Society of London* 58, pp. 240–242 (cit. on p. 94).

Pease, Allan (1988). *Body Language*. London: Sheldon Press (cit. on p. 15).

Pelachaud, Catherine (2015). "Greta: an Interactive Expressive Embodied Conversational Agent." In: *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2015, Istanbul, Turkey, May 4-8, 2015*, p. 5 (cit. on p. 97).

Pentland, Alex (2007). "Social Signal Processing." In: *IEEE Signal. Proc. Mag.* 24.4, pp. 108–111 (cit. on p. 4).

Peters, Christopher, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi (2005). "Engagement Capabilities for ECAs." In: *AAMAS'05 workshop on Creating Bonds with ECAs* (cit. on p. 147).

Picard, Rosalind W. (1995). "Affective computing." In: (cit. on p. 3).

Plutchik, Robert and Henry Kellerman (2013). *Emotion, psychopathology, and psychotherapy*. Vol. 5. Academic press (cit. on p. 37).

Poggi, Isabella (2007). *Mind, hands, face and body. A goal and belief view of multimodal communication*. Weidler (cit. on pp. 44, 147).

Poggi, Isabella and Francesca D'Errico (2010). "Cognitive modelling of human social signals." In: *Proceedings of the 2nd international workshop on Social signal processing*. ACM, pp. 21–26 (cit. on p. 14).

Poignant, Johann et al. (2016). "The CAMOMILE Collaborative Annotation Platform for Multi-modal, Multi-lingual and Multi-media Documents." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.* Ed. by Nicoletta Calzolari et al. European Language Resources Association (ELRA) (cit. on pp. 102, 182).

Poole, David L. and Nevin Lianwen Zhang (2011). "Exploiting Contextual Independence In Probabilistic Inference." In: *CoRR* abs/1106.4864. arXiv: `1106.4864` (cit. on p. 140).

Posner, Jonathan, James A Russell, and Bradley S Peterson (2005). "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology." In: *Development and psychopathology* 17.03, pp. 715–734 (cit. on p. 36).

Poyatos, Fernando (1981). *Gesture inventories: fieldework methodology and problems* (cit. on p. 19).

Premack, David and Guy Woodruff (1978). "Does the chimpanzee have a theory of mind?" In: *Behavioral and brain sciences* 1.4, pp. 515–526 (cit. on p. 41).

Proakis, John G. and Dimitris G. Manolakis (1992). *Digital signal processing - principles, algorithms and applications (2. ed.)* Macmillan (cit. on p. 69).

Provine, R.R. (1996). "Laughter." In: *Amer. Sci.* 84.1, pp. 38–47 (cit. on p. 61).

Pudil, Pavel, Jana Novovicová, and Josef Kittler (1994). "Floating search methods in feature selection." In: *Pattern Recognition Letters* 15.10, pp. 1119–1125 (cit. on p. 70).

Rabiner, Lawrence R. and Biing-Hwang Juang (1993). *Fundamentals of speech recognition*. Prentice Hall signal processing series. Prentice Hall (cit. on p. 110).

Ratner, Carl (1989). "Back to Dr. Ratner's Home Page Journal of Mind and Behavior, 1989, 10, 211-230 A Social Constructionist Critique of Naturalistic Theories of Emotion." In: *Journal of Mind and Behavior* 10, pp. 211–230 (cit. on p. 136).

Reck, Corinna, Daniela Noe, Ulrich Stefenelli, Thomas Fuchs, Francesca Cenciotti, Eva Stehle, Christoph Mundt, George Downing, and Edward Z Tronick (2011). "Interactive coordination of currently depressed inpatient mothers and their infants during the postpartum period." In: *Infant Mental Health Journal* 32.5, pp. 542–562 (cit. on p. 44).

Reece, Michael M and Robert N Whitman (1962). "Expressive movements, warmth, and verbal reinforcement." In: *The Journal of Abnormal and Social Psychology* 64.3, p. 234 (cit. on p. 22).

Reeves, Byron and Clifford Nass (1996). *The media equation - how people treat computers, television, and new media like real people and places*. Cambridge University Press (cit. on p. 3).

Reis, Harry T and Susan Sprecher (2009). *Encyclopedia of Human Relationships: Vol. 1*. Sage (cit. on pp. 27, 28).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144 (cit. on p. 185).

Rich, Charles, Brett Ponsleur, Aaron Holroyd, and Candace L. Sidner (2010). "Recognizing engagement in human-robot interaction." In: *Proceedings of the 5th ACM/IEEE International Conference on Human Robot Interaction, HRI 2010, Osaka, Japan, March 2-5, 2010*. Ed. by Pamela J. Hinds, Hiroshi Ishiguro, Takayuki Kanda, and Peter H. Kahn Jr. ACM, pp. 375–382 (cit. on pp. 32, 135).

Rich, Charles and Candace L. Sidner (2012). "Using Collaborative Discourse Theory to Partially Automate Dialogue Tree Authoring." In: *Intelligent Virtual Agents - 12th International Conference, IVA 2012, Santa Cruz, CA, USA, September, 12-14, 2012. Proceedings*, pp. 327–340 (cit. on p. 158).

Richmond, Virginia P, James C McCroskey, and Steven K Payne (1991). *Nonverbal behavior in interpersonal relations*. Prentice Hall Englewood Cliffs, NJ (cit. on p. 21).

Riek, Laurel D. and Peter Robinson (2011). "Challenges and Opportunities in Building Socially Intelligent Machines [Social Sciences]." In: *IEEE Signal Process. Mag.* 28.3, pp. 146–149 (cit. on p. 136).

Rimé, Bernard (2009). "Emotion elicits the social sharing of emotion: Theory and empirical review." In: *Emotion Review* 1.1, pp. 60–85 (cit. on p. 39).

Ringeval, Fabien, Andreas Sonderegger, Jürgen S. Sauer, and Denis Lalanne (2013). "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions." In: *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013, Shanghai, China, 22-26 April, 2013*, pp. 1–8 (cit. on p. 52).

Ritschel, Hannes, Tobias Baur, and Elisabeth André (2017). "Adapting a Robot's linguistic style based on socially-aware reinforcement learning." In: *26th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2017, Lisbon, Portugal, August 28 - Sept. 1, 2017*, pp. 378–384 (cit. on p. 131).

Rizzolatti, Giacomo and Laila Craighero (2004). "The mirror-neuron system." In: *Annu. Rev. Neurosci.* 27, pp. 169–192 (cit. on p. 29).

Robinson, Laura F and Harry T Reis (1989). "The effects of interruption, gender, and status on interpersonal perceptions." In: *Journal of nonverbal behavior* 13.3, pp. 141–153 (cit. on p. 33).

Roger, Derek, Peter Bull, and Sally Smith (1988). "The development of a comprehensive system for classifying interruptions." In: *Journal of Language and Social Psychology* 7.1, pp. 27–34 (cit. on pp. 33, 157).

Rosenberg, Milton J and Carl I Hovland (1960). "Cognitive, affective, and behavioral components of attitudes." In: *Attitude organization and change: An analysis of consistency among attitude components* 3, pp. 1–14 (cit. on p. 42).

Rosenthal, Stephanie and Anind K. Dey (2010). "Towards maximizing the accuracy of human-labeled sensor data." In: *Proceedings of the 15th International Conference on Intelligent User Interfaces, IUI 2010, Hong Kong, China, February 7-10, 2010*. Ed. by Charles Rich, Qiang Yang, Marc Cavazza, and Michelle X. Zhou. ACM, pp. 259–268 (cit. on pp. 102, 128).

Rossberg-Gempton, Irene and Gary D Poole (1993). "The effect of open and closed postures on pleasant and unpleasant emotions." In: *The Arts in psychotherapy* 20.1, pp. 75–82 (cit. on p. 23).

Roy, Nicholas and Andrew McCallum (2001). "Toward optimal active learning through monte carlo estimation of error reduction." In: *ICML, Williamstown*, pp. 441–448 (cit. on p. 104).

Ruch, Willibald. and Paul Ekman (2001). "The Expressive Pattern of Laughter." In: *Emotion qualia, and consciousness*. Ed. by A. W. Kaszniak, pp. 426–443 (cit. on p. 61).

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1985). *Learning internal representations by error propagation*. Tech. rep. DTIC Document (cit. on p. 73).

Sabourin, Jennifer, Bradford W. Mott, and James C. Lester (2011). "Modeling Learner Affect with Theoretically Grounded Dynamic Bayesian Networks." In: *Affective Computing and Intelligent Interaction - 4th International Conference, ACII 2011, Memphis, TN, USA, October 9-12, 2011, Proceedings, Part I*, pp. 286–295 (cit. on p. 138).

Sacks, Harvey, Emanuel Schlegloff, and Gail Jefferson (1974). “A Simplest Systematics for the Organization of Turn-Taking.” In: *Journal of Language* 50.4, pp. 696–735 (cit. on p. 32).

Salam, Hanan and Mohamed Chetouani (2015). “A multi-level context-based modeling of engagement in Human-Robot Interaction.” In: *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015, Ljubljana, Slovenia, May 4-8, 2015*. IEEE Computer Society, pp. 1–6 (cit. on p. 135).

Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller (2017). “Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models.” In: *CoRR* abs/1708.08296. arXiv: `1708.08296` (cit. on p. 185).

Sanghvi, Jyotirmay, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W. McOwan, and Ana Paiva (2011). “Automatic analysis of affective postures and body motion to detect engagement with a game companion.” In: *Proceedings of the 6th International Conference on Human Robot Interaction, HRI 2011, Lausanne, Switzerland, March 6-9, 2011*, pp. 305–312 (cit. on p. 147).

Santos, J Reynaldo A (1999). “Cronbach’s alpha: A tool for assessing the reliability of scales.” In: *Journal of extension* 37.2, pp. 1–5 (cit. on p. 95).

Saucier, Gerard (1994). “Mini-Markers: A Brief Version of Goldberg’s Unipolar Big-Five Markers.” In: *J. Pers. Assess.* 63.3, p. 506 (cit. on p. 58).

Scheff, Thomas J and Suzanne M Retzinger (2000). “Shame as the master emotion of everyday life.” In: *Journal of Mundane Behavior* 1.3, pp. 303–324 (cit. on p. 172).

Scheflen, Albert E (1964). “The significance of posture in communication systems.” In: *Psychiatry* 27.4, pp. 316–331 (cit. on pp. 15, 32).

Scherer, Klaus A. and Marcel R. Zentner (2001). “EMOTIONAL EFFECTS OF MUSIC: PRODUCTION RULES.” In: (cit. on p. 37).

Scherer, Klaus R (2005). “What are emotions? And how can they be measured?” In: *Social science information* 44.4, pp. 695–729 (cit. on p. 38).

Scherer, S., F. Schwenker, W. N. Campbell, and G. Palm (2009). “Multimodal laughter detection in natural discourses.” In: *Proceedings of 3rd International Workshop on Human-Centered Robotic Systems (HCRS09)*. Ed. by H. Ritter, G. Sagerer, R. Dillmann, and M. Buss. Cognitive Systems Monographs. Springer, pp. 111–121 (cit. on p. 62).

Schmidt, Thomas (2004). “Transcribing and annotating spoken language with EXMARaLDA.” In: *Proceedings of the International Conference on Language Resources and Evaluation: Workshop on XML based richly annotated corpora, Lisbon 2004*. EN. Paris: ELRA, pp. 879–896 (cit. on p. 85).

Schröder, Marc (2010). "The SEMAINE API: Towards a Standards-Based Framework for Building Emotion-Oriented Systems." In: *Adv. Human-Computer Interaction* 2010, 319406:1–319406:21 (cit. on p. 97).

Schröder, Marc et al. (2012). "Building Autonomous Sensitive Artificial Listeners." In: *IEEE Trans. Affective Computing* 3.2, pp. 165–183 (cit. on p. 51).

Schuller, Björn W., Anton Batliner, et al. (2007). "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals." In: *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*. ISCA, pp. 2253–2256 (cit. on p. 111).

Schuller, Björn W., Stefan Steidl, et al. (2013). "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism." In: *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*. Ed. by Frédéric Bimbot, Christophe Cerisara, Cécile Fougeron, Guillaume Gravier, Lori Lamel, François Pellegrino, and Pascal Perrier. ISCA, pp. 148–152 (cit. on pp. 59, 110, 113).

Schuller, Björn W., Martin Wöllmer, Tobias Moosmayr, and Gerhard Rigoll (2009). "Recognition of Noisy Speech: A Comparative Survey of Robust Model Architecture and Feature Enhancement." In: *EURASIP J. Audio, Speech and Music Processing* 2009 (cit. on p. 69).

Schwab, Frank (2000). "Affektchoreographien. Eine evolutionspsychologische Analyse von Grundformen mimisch-affektiver Interaktionsmuster." PhD thesis. Dissertation am FB 5.3 Empirische Humanwissenschaften der Universität des Saarlandes (cit. on p. 41).

Senju, Atsushi and Gergely Csibra (2008). "Gaze following in human infants depends on communicative signals." In: *Current Biology* 18.9, pp. 668–671 (cit. on p. 27).

Serrano, Marcos, Laurence Nigay, Jean-Yves Lionel Lawson, Andrew Ramsay, Roderick Murray-Smith, and Sebastian Denef (2008). "The openinterface framework: a tool for multimodal interaction." In: *Extended Abstracts Proceedings of the 2008 Conference on Human Factors in Computing Systems, CHI 2008, Florence, Italy, April 5-10, 2008*, pp. 3501–3506 (cit. on p. 77).

Settles, Burr (2010). "Active learning literature survey." In: 52.55–66, 11 pages (cit. on pp. 71, 99, 104, 105).

Settles, Burr (2012). *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers (cit. on p. 101).

Sheth, Bhavin R, James Liu, Olayemi Olagbaju, Larry Varghese, Rosleen Mansour, Stacy Reddoch, Deborah A Pearson, and Katherine A Loveland (2011). "Detecting social and non-social changes in natural scenes: Performance of children with and without autism spectrum disorders and typical adults." In: *J. Autism Dev. Disord.* 41.4, pp. 434–446 (cit. on p. 18).

Sidnell, Jack (2007). "Comparative studies in conversation analysis." In: *Annu. Rev. Anthropol.* 36, pp. 229–244 (cit. on p. 32).

Sidner, Candace L., Cory D. Kidd, Christopher Lee, and Neal Lesh (2004). "Where to look: a study of human-robot engagement." In: *Proceedings of the 9th International Conference on Intelligent User Interfaces, IUI 2004, Funchal, Madeira, Portugal, January 13-16, 2004*, pp. 78–84 (cit. on p. 44).

Sinha, Priyabrata (2009). *Speech processing in embedded systems*. Springer Science & Business Media (cit. on p. 67).

Smith, Craig A and Richard S Lazarus (1990). "Emotion and adaptation." In: (cit. on p. 37).

Stikic, Maja, Kristof Van Laerhoven, and Bernt Schiele (2008). "Exploring semi-supervised and active learning for activity recognition." In: *12th IEEE International Symposium on Wearable Computers (ISWC 2008), September 28 - October 1, 2008, Pittsburgh, PA, USA*, pp. 81–88 (cit. on p. 102).

Sun, Chen, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta (2017). "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era." In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 843–852 (cit. on p. 50).

Sun, Xiaofan, Jeroen Lichtenauer, Michel François Valstar, Anton Nijholt, and Maja Pantic (2011). "A Multimodal Database for Mimicry Analysis." In: *Affective Computing and Intelligent Interaction - 4th International Conference, ACII 2011, Memphis, TN, USA, October 9-12, 2011, Proceedings, Part I*, pp. 367–376 (cit. on p. 51).

Tamir, Maya (2011). "The maturing field of emotion regulation." In: *Emotion Review* 3.1, pp. 3–7 (cit. on p. 40).

Tannen, Deborah (1994). *Gender and discourse*. Oxford University Press (cit. on p. 33).

Ter Maat, Mark and Dirk Heylen (2011). "Flipper: An Information State Component for Spoken Dialogue Systems." In: *Intelligent Virtual Agents - 11th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings*, pp. 470–472 (cit. on p. 158).

Ter Maat, Mark, Khiet P. Truong, and Dirk Heylen (2010). "How Turn-Taking Strategies Influence Users' Impressions of an Agent." In: *Intelligent Virtual Agents, 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings*, pp. 441–453 (cit. on p. 34).

Thomas, Andrew P and Peter Bull (1981). "The role of pre-speech posture change in dyadic interaction." In: *British Journal of Social Psychology* 20.2, pp. 105–111 (cit. on p. 32).

Timmer, Sjoerd T., John-Jules Ch. Meyer, Henry Prakken, Silja Renooij, and Bart Verheij (2017). "A two-phase method for extracting explanatory arguments from Bayesian networks." In: *Int. J. Approx. Reasoning* 80, pp. 475–494 (cit. on p. 132).

Tomkins, Silvan S. (1984). "Affect theory." In: *Approaches to emotion* 163, p. 195 (cit. on p. 40).

Tong, Simon and Daphne Koller (2001). "Support Vector Machine Active Learning with Applications to Text Classification." In: *Journal of Machine Learning Research* 2, pp. 45–66 (cit. on p. 102).

Traum, David R., David DeVault, Jina Lee, Zhiyang Wang, and Stacy Marsella (2012). "Incremental Dialogue Understanding and Feedback for Multiparty, Multimodal Conversation." In: *Intelligent Virtual Agents - 12th International Conference, IVA 2012, Santa Cruz, CA, USA, September, 12-14, 2012. Proceedings*. Ed. by Yukiko I. Nakano, Michael Neff, Ana Paiva, and Marilyn A. Walker. Vol. 7502. Lecture Notes in Computer Science. Springer, pp. 275–288 (cit. on p. 4).

Trigeorgis, George, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn W. Schuller, and Stefanos Zafeiriou (2016). "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network." In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pp. 5200–5204 (cit. on p. 70).

Urbain, Jérome, Radoslaw Niewiadomski, Elisabetta Bevacqua, Thierry Dutoit, Alexis Moinet, Catherine Pelachaud, Benjamin Picart, Joëlle Tilmanne, and Johannes Wagner (2010). "AVLaughterCycle." In: *J. Multimodal User Interfaces* 4.1, pp. 47–58 (cit. on pp. 62, 111, 117).

Valente, Fabio, Samuel Kim, and Petr Motlı cek (2012). "Annotation and Recognition of Personality Traits in Spoken Conversations from the AMI Meetings Corpus." In: *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*. ISCA, pp. 1183–1186 (cit. on p. 52).

Valstar, Michel F., Tobias Baur, et al. (2016). "Ask Alice: an artificial retrieval of information agent." In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016*, pp. 419–420 (cit. on pp. 53, 145).

Valstar, Michel François, Brais Martinez, Xavier Binefa, and Maja Pantic (2010). "Facial point detection using boosted regression and graph models." In: *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pp. 2729–2736 (cit. on pp. 147, 153).

Van Baaren, Rick, Loes Janssen, Tanya L Chartrand, and Ap Dijksterhuis (2009). "Where is the love? The social aspects of mimicry."

In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1528, pp. 2381–2389 (cit. on pp. 29, 30).

Van der Stigchel, Stefan, Martijn Meeter, and Jan Theeuwes (2006). "Eye movement trajectories and what they tell us." In: *Neuroscience & Biobehavioral Reviews* 30.5, pp. 666–679 (cit. on p. 27).

Varni, Giovanna, Marie Avril, Adem Usta, and Mohamed Chetouani (2015). "SyncPy: a Unified Open-source Analytic Library for Synchrony." In: *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence, INTERPERSONAL@ICMI 2015, Seattle, Washington, USA, November 13, 2015*, pp. 41–47 (cit. on p. 135).

Verheij, Bart (2014). "To catch a thief with and without numbers: arguments, scenarios and probabilities in evidential reasoning." In: *Law, Probability and Risk* 13.3-4, pp. 307–325 (cit. on p. 132).

Vinciarelli, Alessandro, Maja Pantic, and Hervé Bourlard (2009). "Social signal processing: Survey of an emerging domain." In: *Image Vision Comput.* 27.12, pp. 1743–1759 (cit. on p. 4).

Vinciarelli, Alessandro, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schröder (2012). "Bridging the Gap between Social Animal and Unsocial Machine: A Survey of Social Signal Processing." In: *IEEE Trans. Affective Computing* 3.1, pp. 69–87 (cit. on p. 50).

Vlek, Charlotte S., Henry Prakken, Silja Renooij, and Bart Verheij (2015). "Constructing and understanding Bayesian networks for legal evidence with scenario schemes." In: *Proceedings of the 15th International Conference on Artificial Intelligence and Law, ICAIL 2015, San Diego, CA, USA, June 8-12, 2015*, pp. 128–137 (cit. on p. 132).

Vogt, Thurid and Elisabeth André (2005). "Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition." In: *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, ICME 2005, July 6-9, 2005, Amsterdam, The Netherlands*, pp. 474–477 (cit. on p. 111).

Vogt, Thurid, Elisabeth André, and Nikolaus Bee (2008). "EmoVoice - A Framework for Online Recognition of Emotions from Voice." In: *Perception in Multimodal Dialogue Systems, 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, PIT 2008, Kloster Irsee, Germany, June 16-18, 2008, Proceedings*. Ed. by Elisabeth André, Laila Dybkjær, Wolfgang Minker, Heiko Neumann, Roberto Pieraccini, and Michael Weber. Vol. 5078. Lecture Notes in Computer Science. Springer, pp. 188–199 (cit. on p. 4).

Wagner, Johannes, Elisabeth André, Michael Kugler, and Daniel Leberle (2010). "SSI/ModelUI - A Tool for the Acquisition and Annotation of Human Generated Signals." In: *DAGA 2010*. TU Berlin. Berlin, Germany (cit. on p. 103).

Wagner, Johannes, Elisabeth André, Florian Lingenfelser, and Jonghwa Kim (2011). "Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data." In: *IEEE Trans. Affective Computing* 2.4, pp. 206–218 (cit. on p. 117).

Wagner, Johannes, Tobias Baur, Yue Zhang, Michel F. Valstar, Björn W. Schuller, and Elisabeth André (2018). "Applying Cooperative Machine Learning to Speed Up the Annotation of Social Signals in Large Multi-modal Corpora." In: *CoRR* abs/1802.02565. arXiv: `1802.02565` (cit. on pp. 83, 99).

Wagner, Johannes, Jonghwa Kim, and Elisabeth André (2005). "From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification." In: *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, ICME 2005, July 6-9, 2005, Amsterdam, The Netherlands*. IEEE Computer Society, pp. 940–943 (cit. on p. 4).

Wagner, Johannes, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth André (2013). "The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time." In: *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*. Ed. by Alejandro Jaimes, Nicu Sebe, Nozha Boujemaa, Daniel Gatica-Perez, David A. Shamma, Marcel Worring, and Roger Zimmermann. ACM, pp. 831–834 (cit. on pp. 4, 76, 117).

Wahlster, Wolfgang (1998). *User and discourse models for multimodal communication*. Morgan Kaufmann, San Francisco (cit. on p. 20).

Wallbott, Harald G (1988). "In and out of context: Influences of facial expression and context information on emotion attributions." In: *British Journal of Social Psychology* 27.4, pp. 357–369 (cit. on p. 136).

Wallbott, H.G. (1998). "Bodily expression of emotion." In: *European Journal of Social Psychology*, pp. 879–896 (cit. on p. 24).

Wang, Meng and Xian-Sheng Hua (2011). "Active learning in multimedia annotation and retrieval: A survey." In: *ACM TIST* 2.2, 10:1–10:21 (cit. on p. 102).

Weninger, Felix, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R. Hershey, and Björn W. Schuller (2015). "Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR." In: *Latent Variable Analysis and Signal Separation - 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings*, pp. 91–99 (cit. on p. 5).

White, Sheida (1989). "Backchannels across cultures: A study of Americans and Japanese." In: *Language in society* 18.01, pp. 59–76 (cit. on p. 30).

Willmott, Cort J (1981). "On the validation of models." In: *Physical geography* 2.2, pp. 184–194 (cit. on p. 94).

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alexander Klassmann, and Han Sloetjes (2006). “ELAN: a Professional Framework for Multimodality Research.” In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006.* Ed. by Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias. European Language Resources Association (ELRA), pp. 1556–1559 (cit. on p. 85).

Wlodarczak, Marcin, Hendrik Buschmeier, Zofia Malisz, Stefan Kopp, and Petra Wagner (2012). “Listener head gestures and verbal feedback expressions in a distraction task.” In: *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog, INTERSPEECH2012 Satellite Workshop* (cit. on p. 31).

Wöllmer, Martin, Marc Al-Hames, Florian Eyben, Björn W. Schuller, and Gerhard Rigoll (2009). “A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams.” In: *Neurocomputing* 73.1-3, pp. 366–380 (cit. on p. 134).

Wöllmer, Martin, Angeliki Metallinou, Florian Eyben, Björn W. Schuller, and Shrikanth S. Narayanan (2010). “Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling.” In: *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pp. 2362–2365 (cit. on pp. 134, 137).

Wöllmer, Martin, Björn W. Schuller, Florian Eyben, and Gerhard Rigoll (2010). “Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening.” In: *J. Sel. Topics Signal Processing* 4.5, pp. 867–881 (cit. on p. 138).

Wronka, Eligiusz and Wioleta Walentowska (2011). “Attention modulates emotional expression processing.” In: *Psychophysiology* 48.8, pp. 1047–1056 (cit. on p. 35).

Wu, Huijun, Chen Wang, Jie Yin, Kai Lu, and Liming Zhu (2018). “Sharing Deep Neural Network Models with Interpretation.” In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*. Ed. by Pierre-Antoine Champin, Fabien L. Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis. ACM, pp. 177–186 (cit. on p. 185).

Yap, Ghim-Eng, Ah-Hwee Tan, and HweeHwa Pang (2008). “Explaining inferences in Bayesian networks.” In: *Applied Intelligence* 29.3, pp. 263–278 (cit. on p. 140).

Yngve, Victor H (1970). “On getting a word in edgewise.” In: *Chicago Linguistics Society, 6th Meeting*, pp. 567–578 (cit. on p. 30).

Zeng, Zhihong, Jilin Tu, B. M. Pianfetti, and T. S. Huang (2008). “Audio-Visual Affective Expression Recognition Through Multistream Fused HMM.” In: *Multimedia* 10.4, pp. 570–577 (cit. on p. 51).

Zhang, Nevin Lianwen and David L. Poole (1999). "On the Role of Context-Specific Independence in Probabilistic Inference." In: *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, IJCAI 99, Stockholm, Sweden, July 31 - August 6, 1999. 2 Volumes, 1450 pages*, pp. 1288–1293 (cit. on p. 140).

Zhang, Yue, Eduardo Coutinho, Björn W. Schuller, Zixing Zhang, and Michael Adam (2015). "On rater reliability and agreement based dynamic active learning." In: *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015, Xi'an, China, September 21-24, 2015*. IEEE Computer Society, pp. 70–76 (cit. on p. 102).

Zhang, Yue, Eduardo Coutinho, Zixing Zhang, Caijiao Quan, and Björn Schuller (2015a). "Agreement-based Dynamic Active Learning with Least and Medium Certainty Query Strategy." In: *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015). JMLR W&CP volume 37*. Ed. by A Krishnamurthy, A Ramdas, N Balcan, and A Singh. Lille, France. Lille, France, pp. 1–5 (cit. on pp. 99, 182).

Zhang, Yue, Eduardo Coutinho, Zixing Zhang, Caijiao Quan, and Björn W. Schuller (2015b). "Dynamic Active Learning Based on Agreement and Applied to Emotion Recognition in Spoken Interactions." In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015*. Ed. by Zhengyou Zhang, Phil Cohen, Dan Bohus, Radu Horaud, and Helen Meng. ACM, pp. 275–278 (cit. on p. 102).

Zhang, Zixing, Eduardo Coutinho, Jun Deng, and Björn W. Schuller (2015). "Cooperative Learning and its Application to Emotion Recognition from Speech." In: *IEEE/ACM Trans. Audio, Speech & Language Processing* 23.1, pp. 115–126 (cit. on pp. 99, 100, 105).

Zhu, Xiaojin (2005). *Semi-Supervised Learning Literature Survey*. Tech. rep. Computer Sciences, University of Wisconsin-Madison (cit. on p. 99).

Zimmer, Dieter E (1988). *Die Vernunft der Gefühle: Ursprung, Natur und Sinn der menschlichen Emotion*. Piper (cit. on p. 34).

Zimmermann, Heinz (1996). *Speaking, listening, understanding*. Steiner-Books (cit. on pp. 30, 136).