1st AI-Debate Workshop

Workshop

establishing An InterDisciplinary pErspective on speech-BAsed TEchnology

Magdeburg, September, 27 2021

Editors

Astrid Carolus

Carolin Wienrich

Ingo Siegert

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

Imprint

DOI: 10.25673/38471

Publisher: Otto von Guericke University Magdeburg

Contents

Foreword

Speech-based technology has evolved dramatically over the last few years with ongoing progress. Today's speech-based technology has become some kind of companion system. From a psychological perspective, the ability to speak has been a unique human feature until recently. With speech-based technology adopting these principles, human-technology interaction resembles human-human interaction to an increasing extent. Consequently, besides the well-established scientific perspectives in the field (data and engineering science, psychology), disciplines, which have established a treasure of knowledge on humans and their "functional principles" are becoming valuable sources of knowledge (e.g. social sciences, humanities). The present workshop wants to (1) connect researchers from various scientific fields to (2) collect theoretical and methodological approaches potentially relevant for the analysis and the understanding of speech-based technology and to (3) analyze possible conceptual and methodological overlaps. To account for the limited body of literature in this area, the workshop focuses on (4) a written synopsis of the different scientific perspectives on speech-based technology.

Magdeburg/Würzburg
September 2021 Astrid Carolus, Carolin Wienrich & Ingo Siegert

Program

Time		Working Mode	Aim
09:00 – 10:00	**Opening & Introduction of participants**	Plenum	
10:00 – 11:00	**1st round: presentations: What we know and how** Presentation by each participant, answering question 1+2	Plenum	Collecting dots
11:00 – 11:30	Breakout Session // Coffee Break		Connecting researchers
11:30 – 13:00	**2nd round: presentations: What I want to learn** Presentation by each participant, answering question 3 + 4	Plenum	Collecting dots
13:00 – 14:00	Lunch Break		
14:00 – 15:30	**Teamwork**	Small groups	Organizing dots
15:30 - 15:45	Breakout Session: // Coffee Break		Connecting researchers
15:45 – 16:30	**Teamwork**: presentations	Plenum	Organizing dots
16:30 – 17:30	**Conclusion, Wrap up, Next Steps**	Plenum	

Introduction to the Workshop

Astrid Carolus[1], **Carolin Wienrich**[2], **and Ingo Siegert**[3]

[1]Julius-Maximilians-University, Wuerzburg, astrid.carolus@uni-wuerzburg.de
[2]Julius-Maximilians-University, Wuerzburg, carolin.wienrich@uni-wuerzburg.de
[2]Otto-von-Guericke-University, Magdeburg, ingo.siegert@ovgu.de

The Workshop on **Establishing an interdisciplinary perspective on speech-based technology** has the aim to collect and connect theoretical and methodological approaches from multiple disciplines to widen the perspective on speech-based technology.

Motivation and Goals of the Workshop

Speech-based technology has evolved dramatically over the last few years by technical improvements from data science/ engineering science, especially in far-field acoustic speech enhancement, together with a vast performance increase in automatic speech recognition and speech understanding, on the one side. Particularly noteworthy is the field of Affective Computing – aiming to develop technical systems able to receive, understand, and (adequately) react to affective signals [2]. As well as the recent improvements in artificial intelligence (i.e., Deep, Ensemble, Active Learning, and Fusion) – allowing to improve and adapt the system constantly and directly to a user [6, 3]. Furthermore, ubiquitous technological advances provide the technology always at the user's location [4]. Thus, technological developments have changed the human-technique interaction fundamentally, resulting in vital consequences for the human users, on the other side. Speech-based technology steps over from pure command receivers to a kind of companion system that incorporates human needs, desires, and abilities [5]. Therefore, interdisciplinary research activities need to be intensified integrating the various scientific perspectives beyond a mere technological focus.

With its focus on the individual, the psychological perspective indicates the increasing similarities between the operation of speech-based technology and human-human interaction. The human user experiences some kind of conversation with the device or with an application, which seems to listen and answer. Conducting a conversation, however, has been a uniquely human ability until recently [1]. Adopting these allegedly human qualities results in the usage of technology to fulfill criteria, which have been exclusive for human-human interactions [7]. Consequently, from the human user's perspective, the line between using a technological device or application and interacting with an (at least) humanlike counterpart has begun to blur. Consequently, the "human" in human-computer interaction has become increasingly relevant, enhancing the relevance of others than technology-oriented disciplines (e.g., data science, engineering science, media science, HCI). If "computers" transfer into humanlike counterparts, disciplines, which have established a treasure of knowledge on humans and their "functional principles", will be valuable sources of knowledge. They already know how humans think, feel and behave. They know about the way they learn, how they live together, communicate, collaborate. Social sciences (e.g., sociology, educational and political science) or humanities (e.g., psychology, philosophy, history, art) , and medical and health sciences could draw on a large body of knowledge, theoretical concepts or methodological approaches, which research on speech-based technology can benefit from. However, literature reviews reveal a narrow field of contributing disciplines, with most studies originating from standard natural or technical sciences. In contrast to other rather new digital technologies such as social robots or virtual assistants, research on speech-based technology is only at the beginning of the process to become an established area of research- although, for example, voice assistants can already be found in millions of households.

The first **workshop, "AI Debate", aims** to widen the perspective on human-technology interaction by bringing together researchers from different scientific areas rarely represented in the field. The contribution focuses on three main aims:

1. connecting researchers from different backgrounds
2. collecting theoretical and methodological approaches from the range of scientific disciplines
3. analyzing possible conceptual and methodological overlaps and cross connections

The present contribution summarizes the input comments given by researchers of different areas participating in the workshop. The comments answer five questions.

1. What do I know already? By answering this question, researchers report theories and concepts on speech-based technology from their perspective.
2. How do I study the phenomenon? Answers to this question are current methodological approaches to studying speech-based technology.
3. What would I like to know? Here, researchers formulate open questions and future directions to learn about speech-based technology from other research areas.
4. What do I want to learn from different disciplines? Researchers reify specific theoretical or methodological approaches from other disciplines to gain deeper insights into speech-based technology by answering this question.
5. What do I want to teach other disciplines? These answers are specific theoretical or methodological approaches from their field they wish to share to gain deeper insights into speech-based technology.

The Organizers

Astrid Carolus is an assistant professor for media psychology at the University of Würzburg. In her research, she focuses on humans interacting with digital technology and with media content. From a psychological perspective, she analyzes the underlying motives of media usage (as well as the refusal of usage) and its effects. Her research foci are on human users interacting with digital devices and AI-enabled technology, which she conceptualizes as psychologically relevant entities; on health-related effects of social media communication; and on teachers' digital literacy.

Carolin Wienrich is a professor for Human-Technique-Systems at the University of Würzburg and co-leader of the XR HUB Würzburg. Her research interests focus on interaction paradigms between humans and digital entities and change experiences during and after digital interventions. Her team explores antecedents, potentials, and risks of digital interactions and experiences since digital entities and digital interventions accompany humans in many contexts. Participative and human-centered research, theoretical concepts, and multi-methods stemming from psychology and computer science define her qualification in human-computer interaction.

Ingo Siegert is assistant professor for Mobile Dialog Systems at the Otto von Guericke University Magdeburg. His research interests are on signal-based analyses and interdisciplinary investigations of human-computer interaction in terms of addressee detection, perceived charisma as well as the utilization of further interaction patterns, such as filled pauses or discourse particles within the scope of voice assistants. Ingo Siegert has published 100+ peer reviewed papers and articles on several conferences and various journals and is co-organizer of various workshops and conferences.

References

[1] S. Pinker. *The language instinct. How the Mind Creates Language.* William Morrow & Co, USA., 1994.

[2] B. Schuller et al. "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge". In: *Speech Commun* 53 (9-10 Nov. 2011), pp. 1062–1087.

[3] M. Glodek et al. "Fusion paradigms in cognitive technical systems for human–computer interaction". In: *Neurocomputing* 161 (2015), pp. 17–37.

[4] S. Biundo and A. Wendemuth. "Companion-Technology for Cognitive Technical Systems". In: *KI - Künstliche Intelligenz* 30.1 (Feb. 2016), pp. 71–75.

[5] T. Gossen et al. "Modeling aspects in human-computer interaction - adaptivity, user characteristics and evaluation". In: *Companion technology: a paradigm shift in human-technology interaction.* Cham: Springer International Publishing, 2017, pp. 57–78.

[6] F. Schwenker et al. "Multimodal affect recognition in the context of human-computer interaction for companion-systems". In: *Companion technology: a paradigm shift in human-technology interaction.* Cham: Springer International Publishing, 2017, pp. 378–408.

[7] A. Carolus et al. "Impertinent mobiles - Effects of politeness and impoliteness in human-smartphone interaction". In: *Computers in Human Behavior* 93 (2019), pp. 290–300.

Author Contributions

Towards Speech-based Interactive Post hoc Explanations in Explainable AI

STEFAN HILLMANN, TU Berlin, Germany
SEBASTIAN MÖLLER, TU Berlin, and German Research Center for Artificial Intelligence (DFKI), Germany
THILO MICHAEL, TU Berlin, Germany

CCS Concepts: • **Human-centered computing → Natural language interfaces**.

Additional Key Words and Phrases: XAI, Explainable AI, post hoc explanations, spoken dialog, argumentation

1 INTRODUCTION

AI-based systems offer solutions for information extraction (e.g., finding information), information transformation (e.g., machine translation), classification (e.g., classifying news as fake or true), or decision support (e.g., providing diagnoses and treatment proposals for medical doctors) in many real-world applications. The solutions are based on machine leaning (ML) models and are commonly offered to a large and diverse group of users, some of them experts, many others naïve users from a large population. Nowadays and in particular deep neural network architectures are black-boxes for users and even developers [1, 4, 9] (also cp. [6]). A major goal of Explainable Artificial Intelligence (XAI) is making complex decision-making systems more trustworthy and accountable [7, p. 2]. That is why XAI seeks to ensure transparency, interpretability, and explainability [9].

Common to most users is that they are not able to understand the functioning of the AI-based systems, i.e., those are perceived as black-boxes. Humans are confronted with the results, but they cannot comprehend what information was used by the system for reaching this result (interpretability), and in which way this information was processed and weighted (transparency). The underlying reason is that an explicit functional description of the system is missing or even not possible in most Machine-Learning-(ML)-based AI systems – the function is trained by adjusting the internal parameters, and sometimes also the architecture is learned from a basic set of standard architectures. However, natural language and speech-based explanations allow better explainability [1, p. 11] due to interactive post hoc explanations in from of an informed dialog [7, p. 2]. Additionally, AI is also addressed by regulations, e.g., of the EU [2, 3], and thus becomes even more relevant for industry and research. Here, not at least recognition of bias in AI systems' decision plays an important role.

2 WHAT DO WE KNOW ALREADY?

XAI has attracted much attention to increase trust in AI-based systems, and finally to enable acceptance. Here, our perspective on trust follows Ribeiro's et al. [8] approach. They differentiate two different aspects of trusting in an (AI) model: "(1) trusting a prediction" (or decision), and "(2) trusting a model, i.e. whether the user trusts a model to behave in reasonable ways if deployed". Trusting in a prediction is the crucial aspect for a user affected by the decisions of an AI system. Trust means, the user trusts to a sufficient degree in the (factual) correctness of the model's decision [8]. Especially when interacting with an AI system in a human-like manner (e.g., by natural language), trust is strongly related to the user's perceived competence of an AI system. Systems able to provide comprehensible explanations will potentially show a higher competence than systems just providing decisions or predictions, without any explanation.

Authors' addresses: Stefan Hillmann, TU Berlin, Berlin, Germany, stefan.hillmann@tu-berlin.de; Sebastian Möller, TU Berlin, and German Research Center for Artificial Intelligence (DFKI), Berlin, Germany; Thilo Michael, TU Berlin, Berlin, Germany.

1st AI-Debate Workshop © OvGU, Editors; DOI: 10.25673/38473

XAI frequently focuses on what information is encoded in a model's intermediate representation. One way to do this inspection is through different kinds of probing tasks that enable to study the behavior of the AI model towards different input. Another way is to use saliency methods that show activation of model features when providing a certain output. The saliency maps can be applied to the input space, thus highlighting which input information has served to what degree in providing the desired output.

Evaluations in the context of textual explanations show that automatically generated explanations can increase human understanding, trust and confidence in the AI system for certain tasks [5]. However, whereas graphical highlighting might be helpful for experts, it comes to its limits when more complex relationships are to be analyzed. Such complex relationships could better be presented via natural language, which is our common way to express relationships between entities, and to pronounce judgments. Natural explanations are contrastive, selective and social [7] and argumentation-based dialogs are the most suitable for this purpose [10, p. 14]. We thus advocate for explanations about the behavior of AI-based systems which are generated in the form of natural language, and preferably using speech.

Let's take the example of a decision support system for medical doctors, which provides diagnosis support for kidney transplant patients. The doctor would have to find and justify his own diagnosis; knowing the reasons why the system proposes a certain diagnosis (e.g., the physiological parameters, comparable cases, risk factors, etc.) would enable him to judge his own and the system's diagnosis, and to better consider all relevant information for the decision. Another example is an AI-based system that identifies fake news on a public news channel: The system might provide reasons on why it judges certain news to be fake, e.g., by citing contradicting trustworthy information, highlighting the source of the fake news, or others.

A speech-based system could provide such explanations in a rather unobtrusive way. It could be triggered by the user who would like to know more about the basis for the decision. The user could ask a rather general question ("What makes you confident that this information is fake?") or a more specific one ("Did the author of this news article already spread fake news?" or "Which political party does he belong to?").

3 HOW DO WE STUDY THE PHENOMENON?

Research on this topic requires empirical analysis with a range of systems showing different degrees of performance, and offering different ways of providing explanations. The explanations could be given via text or speech; they could be given one-shot, as a user-driven question-answer dialog, or via a more system-initiative conversation. Explanations could include external sources of information which could be distinguished via voice characteristics, and use prosodic features, e.g., to gradually express confidence . Ultimately, speech and natural language could be combined with other modalities to enable multimodal explanations, adapted towards the user's needs and usage situation.

With such systems, comparative empirical studies should be carried out, comparing different system versions with different ways of generating explanations. The empirical studies should be performed with a representative group of users of the target application (if possible), as the users' needs for explanations might duffer according to their background knowledge.

4 WHAT WE WOULD LIKE TO KNOW?

The aim of this research is to enable AI systems which provide helpful and comprehensible explanations to the users. A proper form and of post hoc explanations could help to increase trust in the system, and finally its acceptance. Speech-based systems are particularly adequate for this purpose, as speech characteristics could be dedicatedly manipulated to increase explainability.

1st AI-Debate Workshop © OvGU, Editors; DOI: 10.25673/38473

5 WHAT DO I WANT TO LEARN FROM DIFFERENT DISCIPLINES?

For appropriate generation (form and degree of detail) of explanations we count input from cognitive psychology and psychology of explanations as well as argumentation (as part of NLP). Social sciences are important concerning acceptance beyond pure performance of AI-systems (cp. [1, 7]). AI and XAI are needed for extraction of information provided in conversational explanations.

6 WHAT DO WE WANT TO TEACH OTHER DISCIPLINES?

Knowledge about how speech and language technology can be used to provide (and evaluate) explanations in AI will open a powerful tool to the (X)AI research community. We expect that knowledge on speech technology, dialog design, argumentation and natural language generation will enable better explanations for different types of applications.

REFERENCES

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58 (June 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[2] European Commission. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council (GDPR).

[3] European Commission. 2021. Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.

[4] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable Artificial Intelligence. *Science Robotics* 4, 37 (Dec. 2019), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[5] Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-Grounded Evaluations of Explanation Methods for Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5194–5204. https://doi.org/10.18653/v1/D19-1523

[6] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[7] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 279–288. https://doi.org/10.1145/3287560.3287574

[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco California USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[9] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. 2020. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access* 8 (2020), 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199

[10] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and Explainable Artificial Intelligence: A Survey. *The Knowledge Engineering Review* 36 (2021), e5. https://doi.org/10.1017/S0269888921000011

Impulse: How do we arrive at a decent spoken dialogue system?

ALJOSCHA BURCHARDT, JAN NEHRING, DFKI Speech and Language Technology Lab, Germany

Additional Key Words and Phrases: Spoken dialogue systems, interdisciplinarity

1 WHAT DO WE KNOW ALREADY?

Several of today's applications of speech-based technologies just use a different modality, but are not dramatically different from their written counterpart. If we take voice search, e.g., there are obviously some advantages such that it can be hands-free, is possible without a display, e.g., in extreme lightning situations etc. And there are disadvantages as it is, e.g., not possible to utilize lightweight input support via type-ahead and it is especially challenging to present the search results in spoken language. But the search process itself is not fundamentally different. But when it comes to the communicative aspect of language, speech-based systems can make a difference and what comes to mind immediately are spoken dialogue systems.

In a very strong simplification, AI (Artificial Intelligence) systems are either *modeled* in some symbolic framework or *learnt from data* using some form of machine learning (ML).[1] In terms of performance, ML has recently often outperformed the symbolic approaches while suffering from lack of control and transparency. In the realm of dialogue processing, we typically find a mix of technologies. The periphery (speech input and output) is usually data-driven as is the recognition of user intents. The dialogue itself is usually modeled using one of several existing frameworks that are also the basis of our personal assistants. That is why the dialogues are typically centered around simple tasks such as accessing an FAQ list or switching the light on where it is comparably easy to measure task completion, i.e., success of the dialogue. Some systems that have learnt simple dialogues like asking for opening hours do exists, but we are very far from the open domain dialogue capacities of a four year old child.

On the other hand chatbots have more to offer then fulfilling a certain task in a short time. We performed user studies that suggest that chatbots can generate a more playful and interesting user experience than traditional interfaces. When the chatbot designer also targets other goals beyond task completion, such as user satisfaction, fun or curiosity, we believe that chatbots can generate a fun and rich experience.

2 HOW DO WE STUDY THE PHENOMENON?

We predominantly built chatbots for industry applications. In these projects the goal of creating an interesting character for the chatbot was often subordinate to other project goals. Nevertheless, in the user studies we performed, even these rather serious chatbots have been perceived as more interesting, innovative and generally attractive than traditional user interfaces by the participants. Methodologically, we performed classic user experience evaluation using questionnaires such as AttrakDiff [1] and the User Experience Questionnaire [2], sometimes accompanied by structured interviews.

In a cooperation with the University of Arts / Berlin (UdK) we created chatbots for the mere purpose of entertainment and although we could not perform a formal user evaluation, the feedback to these chatbots was very positive.

[1] The author is aware that this is not a scientifically clean partition.

Author's address: Aljoscha Burchardt, Jan Nehring, Aljoscha.Burchardt@dfki.de, DFKI Speech and Language Technology Lab, Alt-Moabit 91c, Berlin, Germany, 10559.

3 WHAT WE WOULD LIKE TO KNOW?

Human dialogue or – more generally – conversation has been studied in a huge variety of research areas including philosophy, cognitive psychology, (psycho)linguistics, sociology etc. Our question would be how we can inform and drive the design process of future dialogue systems starting from the insights and frameworks used in these fields of study: How can we model a (good) dialogue? What brings a dialogue further? How can we make sure that the partners understand each other (grounding)? What is the final goal of a dialogue? How can we evaluate and measure the success of a dialogue?

How can chatbots generate a rich and entertaining user experience? Of which dimensions does this user experience consist? And how do we measure these dimensions? If we had such an evaluation framework, we could evaluate different chatbot implementations to find out which one provides the best user experience beyond task success.

The most central question, however, would be how we can translate all these insights from different areas into either formal models or data in order to incorporate them in AI systems as sketched above. This will require a high degree of interdisciplinary cooperation.

4 WHAT DO WE WANT TO LEARN FROM DIFFERENT DISCIPLINES?

We think the different disciplines would first and foremost need to find a common language. Not only on the meta level talking about the phenomenon and their different perspectives and research questions, but also when describing the phenomenon of study itself, e.g., when describing what kind of knowledge a system (dialogue partner) would need about the world, the current situation or the (human) dialogue partner.

5 WHAT DO WE WANT TO TEACH OTHER DISCIPLINES?

We would try to make people see the world through "digital lenses": Everything a weak AI system can do has to be taught to it before in some way. For a translation system, this is comparably easy, we feed it with (human translated) texts. For a summarization system, it is already much more complicated. We need to feed it with texts and their summaries. There is much more variety (and conversely, sparseness) in this data, but we can imagine how it would work. What data and knowledge structures do we need to do for a dialogue system that goes beyond the current state of the art?

REFERENCES

[1] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & Computer 2003: Interaktion in Bewegung*, G. Szwillus and J. Ziegler (Eds.). B. G. Teubner, Stuttgart, 187–196.

[2] Bettina Laugwitz, Martin Schrepp, and Theo Held. 2006. Konstruktion eines Fragebogens zur Messung der User Experience von Softwareprodukten. In *Mensch und Computer 2006: Mensch und Computer im Strukturwandel*, Andreas M. Heinecke and Hansjürgen Paul (Eds.). Oldenbourg Verlag, München, 125–134.

Engagement Recognition Using Audio Channel Only

DENIS DRESVYANSKIY*, Institute for Communications Engineering, Ulm University, Germany and ITMO University, Russia
INGO SIEGERT*, Mobile Dialog Systems, Institute for Information Technology and Communications, Germany
ALEXEI KARPOV, SPIIRAS, SPC RAS, Russia and ITMO University, Russia
WOLFGANG MINKER, Institute for Communications Engineering, Ulm University, Germany

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**; • **Computing methodologies → Artificial intelligence**.

Additional Key Words and Phrases: human-computer interaction, paralinguistics, engagement recognition, audio processing

1 INTRODUCTION

Utilizing dialogue assistants endowed with weak artificial intelligence has become a common technology, which is widespread across many industrial spheres - from operating robots using voice to speaking with an intelligent bot by telephone. However, such systems are still far from being essentially intelligent systems, since they cannot fully mimicry or replace humans during human-computer interaction (HCI). Nowadays, paralinguistic analyses is becoming one of the most important parts of HCI, because current requirements to such systems have been increased due to sharped improvement of speech-recognition systems: now, the HCI system should not only recognize, ***what*** the user is talking about, but also ***how*** he/she is talking, and ***which intention/state*** does he/she have now. Those include analyzing and evaluating such high-level features of dialogue as stress, emotions, engagement, and many others.

Although there have been a lot of studies in paralinguistics devoted to recognizing high-level features (such as emotions[1] and stress[17, 25]) using audio cues, there are still almost no insights on how it could work for engagement.

2 WHAT DO WE KNOW ALREADY?

Engagement is a complex phenomenon, which is still not strictly defined in the scientific community. The most common definitions in the context of HCI are stated in the following. Sidner et al. in [24] stated engagement as a "process by which two (or more) participants establish, maintain and end their perceived connection", where "connection" can be expressed in various ways. Poggi [21] characterized engagement by "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". However, the engagement was considered also in terms of qualities of interfaces [22], user experience [20], in the context of social media [12], and many other areas.

From a theoretic point of view, researchers divide the engagement concept into cognitive, emotional, and behavioral components [7]. The cognitive part is mostly expressed by a person's attention to the interlocutor or task to do, while emotional (affective) engagement encompasses the person's emotions and attitudes, which are reflected by the enjoyment of the particular action. Apart from "brain"-related components of engagement, some researchers further point out a behavioral construct of engagement[5, 16], which is strictly conveyed by actions, giving a possibility to measure engagement more objectively.

[*] Both authors contributed equally to this research.

1st AI-Debate Workshop © OvGU, Editors; DOI: 10.25673/38475

Whichever point researchers will lean ultimately, engagement recognition using audio-only is becoming a hot topic in HCI, especially in dialogue systems and human-robot interactions. The key point is that people are able to understand, whether an interlocutor is engaged in the conversation or not, yet it is still difficult for all kinds of HCI systems.

3 HOW DO WE STUDY THE PHENOMENON?

Depending on the usage, the present engagement recognition systems are based on either facial or multi-modal features. In e-learning, researchers [18, 26] use mostly facial features, since a participant is usually silent during lecturer's monologue. However, when the participant is involved in the conversation with other persons (or robots), utilization of facial features is not enough, since a complexity of signals interpretation increases: there are more visual occlusions due to high movements amount of other participants and visual activity of the target participant. However, at the same time, we have a richness of the acoustic signals - speech itself and social signals expressed via laughter, fillers, backchannels and many others. In that case, researchers deploy multi-modal systems, which fuse various cues to do a final prediction. Those include postures and gestures [6, 9], gaze activity [11], visual focus of attention [23], and audio features[14, 15]. It should be noted, however, that audio features in this case mean high-level features such as laughter, backchannels, and turn-taking and play just a complementary role for the generation of the final decision.

According to aforementioned use-cases, there are several databases to study engagement: (1) devoted to e-learning [10, 19], (2) acquired during human-robot interaction [3, 13] and (3) represent human-human dyadic conversations [4]. To the best of our knowledge, neither was exploited for engagement recognition using audio-only features.

4 WHAT WE WOULD LIKE TO KNOW?

One of the key problems in all presented databases and in the engagement recognition domain overall is inconsistency in label scale - it differs from paper to paper, including 2-point scale (disengaged, engaged) [15], 4-point scale (very low, low, high, very high engagement) [10, 14] and 5-point scale (disengaged, low engaged, neutral, engaged, highly engaged)[4, 6]. Sometimes researchers use even more fine-grained scales with more classes [2, 9], although it is rare. Utilization of the 5-level scale looks the most attractive since we can neatly adjust the system response to the user. However, it is not clear, whether human annotators are able to distinctly separate the engagement on 5 states. To prove it, a comprehensive perception study is needed. First of all, there is an important question to be answered: ***Is it theoretically possible to set apart engagement on 5 levels***?

The second key problem lies in the domination of video-based systems in engagement recognition. While researchers are actively implementing multi-modal and video systems to capture user engagement in domains related to e-learning and offline conversations, dialogue related technical systems, such as voice assistants endure a lack of analysis of such characteristics just because they are limited to speech analyses. Audio is mostly exploited as complementary information in multi-modal systems for final decision making, and therefore is not used as a standalone signal for engagement recognition. However, there are many use-cases, when a researcher has only audio, yet should predict the user's state such as engagement. There are almost no studies on audio features in this direction. Thus, it comes to the second important question: ***Can we build a system, which is able to recognize interlocutor's engagement using audio-only (speech and linguistic) cues***?

5 WHAT DO WE WANT TO LEARN FROM DIFFERENT DISCIPLINES?

Trying to answer the second question, we need to define suitable audio features, which are able to characterize engagement. Since people are capable to understand the engagement of interlocutors, it can be assumed that there exist a set of features able to effectively express engagement through audio cues. Finding an "optimal" (in terms of the efficiency of the engagement recognition system) set of audio features, which will highly correlate with user engagement state, will allow training a machine learning model to automatize the process of engagement identification. However, just iterating over all available features is not efficient and hardly implementable, as to characterize acoustics a large number of different features can be used [8]. Thus, we need strong theory-based evidence, which can advise the direction of feature search.

The problem of engagement scales should be also solved starting from the theoretical background. Despite empirical studies on different scales and fact that 4-point and 5-point scales can be turned into 2-point or 3-point scales, the choice of scale should be firstly theoretically supported with psychological researches.

6 WHAT DO WE WANT TO TEACH OTHER DISCIPLINES?

The machine learning (ML) approach is widely known among many research areas. In truth, ML and deep learning (DL) have permeated more or less into all research disciplines. Today it is difficult to imagine the processing and analysis of some data acquired during any work without ML. On the other hand, DL is a fast developing domain of ML, earning a new "breath" with developing special frameworks capable to run DL scripts on GPUs. However, the task of an ML engineer is not only to develop algorithms, but also to work with data during the whole development loop: acquisition, analysis, pre-process, and post-process. All these stages require specific skills to be applied and failure in any of them can be cause of turning into "wrong" data, leading to the incorrect decision provided by the chosen algorithm.

In the case of engagement recognition, the implementation of an recognition system will be based on theoretically (and experimentally) proved valuable features. There are numerous techniques we can utilize for data analysis - processing the raw audios with 1D or 2D convolutional neural networks (CNN), engineering new features from raw signal to diverse the set of available features, using linear and non-linear transformations to reduce the dimension of selected features, evaluating the significance of the obtained features and many others.

7 CONCLUSION

It has been a long path to teach computer systems understanding of human speech. Today, such systems are able to some extend. However, still far from maintain the conversation in the way humans do: dialogue systems do not take into account paralinguistics, were we see engagement of the interlocutor (user) as one important aspect. Although the problem of engagement recognition is under consideration in the research community nowadays, the limitation to audio-only cues is still challenging. To eliminate this drawback, we the developments should be guarded by theoretical groundings to be able to define appropriate features and develop an automated engagement recognition system, which will be able to reliably predict user engagement and act appropriately when the user is about to fall into disengagement during a conversation with a speech-based technical system.

ACKNOWLEDGMENTS

The research was supported by the German Federal Ministry of Education and Research project “RobotKoop: Cooperative Interaction Strategies and Goal Negotiations with Learning Autonomous Robots".

REFERENCES

[1] Mehmet Berkehan Akçay and Kaya Oğuz. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication* 116 (2020), 56–76.

[2] Roman Bednarik, Shahram Eivazi, and Michal Hradis. 2012. Gaze and conversational engagement in multiparty video conversation: an annotation scheme and classification of high and low levels of engagement. In *Proc of the 4th Gaze-In'12*. 1–6.

[3] Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. 2017. UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 464–472.

[4] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 350–359.

[5] Hamish Coates. 2007. A model of online and general campus-based student engagement. *Assessment & Evaluation in Higher Education* 32, 2 (2007), 121–141.

[6] Soumia Dermouche and Catherine Pelachaud. 2018. From analysis to modeling of engagement as sequences of multimodal behaviors. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[7] Kevin Doherty and Gavin Doherty. 2018. Engagement in HCI: conception, theory and measurement. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–39.

[8] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. of the ACM MM-2010*.

[9] Joseph F Grafsgaard, Joseph B Wiggins, Alexandria Katarina Vail, Kristy Elizabeth Boyer, Eric N Wiebe, and James C Lester. 2014. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In *Proceedings of the 16th International Conference on Multimodal Interaction*. 42–49.

[10] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. 2016. Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885* (2016).

[11] Ryo Ishii, Yukiko I Nakano, and Toyoaki Nishida. 2013. Gaze awareness in conversational agents: Estimating a user's conversational engagement from eye gaze. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 2 (2013), 1–25.

[12] Alejandro Jaimes, Mounia Lalmas, and Yana Volkovich. 2011. First international workshop on social media engagement (SoME 2011). In *ACM SIGIR Forum*, Vol. 45. ACM New York, NY, USA, 56–62.

[13] Dinesh Babu Jayagopi, Samira Sheiki, David Klotz, Johannes Wienke, Jean-Marc Odobez, Sebastien Wrede, Vasil Khalidov, Laurent Nyugen, Britta Wrede, and Daniel Gatica-Perez. 2013. The vernissage corpus: A conversational human-robot-interaction dataset. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 149–150.

[14] Jaebok Kim, Khiet P Truong, Vicky Charisi, Cristina Zaga, Vanessa Evers, and Mohamed Chetouani. 2016. Multimodal detection of engagement in groups of children using rank learning. In *International Workshop on Human Behavior Understanding*. Springer, 35–48.

[15] Divesh Lala, Koji Inoue, Pierrick Milhorat, and Tatsuya Kawahara. 2017. Detection of social signals for recognizing engagement in human-robot interaction. *arXiv preprint arXiv:1709.10257* (2017).

[16] Xuezhao Lan, Claire Cameron Ponitz, Kevin F Miller, Su Li, Kai Cortina, Michelle Perry, and Ge Fang. 2009. Keeping their attention: Classroom practices associated with behavioral engagement in first grade mathematics classes in China and the United States. *Early Childhood Research Quarterly* 24, 2 (2009), 198–211.

[17] Iulia Lefter, Gertjan J Burghouts, and Léon JM Rothkrantz. 2015. Recognizing stress using semantics and modulation of speech and gestures. *IEEE Transactions on Affective Computing* 7, 2 (2015), 162–175.

[18] Jiacheng Liao, Yan Liang, and Jiahui Pan. 2021. Deep facial spatiotemporal network for engagement prediction in online learning. *Applied Intelligence* (2021), 1–13.

[19] Love Mehta, Aamir Mustafa, et al. 2018. Prediction and localization of student engagement in the wild. In *Digital Image Computing: Techniques and Applications (DICTA), 2018 International Conference on, IEEE.*

[20] Heather L O'Brien and Elaine G Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology* 59, 6 (2008), 938–955.

[21] Isabella Poggi. 2007. *Mind, hands, face and body: a goal and belief view of multimodal communication*. Weidler.

[22] W Quesenbery. 2003. Dimensions of usability. Content and complexity: Information design in technical communication.

[23] Hanan Salam and Mohamed Chetouani. 2015. Engagement detection based on mutli-party cues for human robot interaction. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 341–347.

[24] Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence* 166, 1-2 (2005), 140–164.

[25] Mariette Soury and Laurence Devillers. 2013. Stress detection from audio on multiple window analysis size in a public speaking task. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 529–533.

[26] Woo-Han Yun, Dongjin Lee, Chankyu Park, Jaehong Kim, and Junmo Kim. 2018. Automatic recognition of children engagement from facial video using convolutional neural networks. *IEEE Transactions on Affective Computing* 11, 4 (2018), 696–707.

Acceptance by Design: Voice Assistants

MARIA RAUSCHENBERGER, Faculty of Technology, University of Applied Sciences Emden/Leer, Germany

CCS Concepts: • **Human-centered computing → Interaction design**; *Human computer interaction (HCI)*; *Ubiquitous and mobile computing*.

Additional Key Words and Phrases: Design process, Speech-based Technology, Technology Adoption, Voice User Interface, VUI, Voice Assistant, User Experience, UX, Human-Centered Design Process, Context of Use

1 INTRODUCTION

Voice Assistants (VAs), Speech-based Technology (SBT), and Voice User Interfaces (VUI) have gained popularity in recent years; however, privacy concerns stop people from using them. Although the terms are used for slightly different purposes, we mainly use "VAs" for simplicity. To increase the adoption rate and address users' needs, we propose reducing barriers with *Acceptance by Design.* By that, we mean using, *e.g.,* VA designs that implicitly show if they collect data or how much data is collected. We can do that by first exploring the user requirements and concerns with methods from the Human-Centered Design Process [4] to provide guidelines for the future. This is a complex challenge that requires a deep understanding from different disciplines to design VAs for users skeptical about them. The HCD offers a holistic approach for capturing user requirements or discovering new dimensions of the UX that influence the user's perception or awareness of the interactive system. We must keep in mind that users have different needs (*e.g.,* children *vs.* parents *vs.* people with disabilities) that could result in different design solutions. Research has already started to explore with questionnaires [9, 10] and interviews the risks and opportunities of VAs (*e.g.,* cultures [6, 22], use cases [1, 3], existing users, or potential users [1, 3, 22]).

Furthermore, these guidelines might help others to design future interactive systems for different use cases, such as augmented reality glasses for university students in labs.

2 WHAT DO I KNOW ALREADY?

Voice assistants (VAs) have become ubiquitous in recent years. With the rise of products such as Alexa (Amazon) and Siri (Apple), VA use is expected to increase even further [23]. Not only do VAs have many advantages, but users already use speech commands while driving or changing music in a car. Future use cases might include commands while doing something else, such as cooking. Other opportunities depend on the user's abilities, such as for children who cannot write yet or people with disabilities, *e.g.,* visual impairments. While parents often have privacy concerns, children tend to be less fearful using new systems and therefore take advantage of a voice interface. People with disabilities use VAs intensively, not only for goal-orientated tasks, but also for pleasure, *e.g.,* games [8].

In general, users from Germany have shown high VA usage, but they also have concerns about privacy due to the collection of data [3, 7, 22]. This fact comes as no surprise, since VAs are placed in private spaces that are highly sensitive. Information and conversations originally did not leave spaces such as living rooms, bathrooms, or bedrooms. With VA microphones that are triggered by commands such as "*Alexa*," users feel like there is a fly on the wall that is always listening.

1st AI-Debate Workshop © OvGU, Editors; DOI: 10.25673/38476

3 HOW DO I STUDY THE PHENOMENON?

Innovations trigger concerns and therefore require early adopters to explore the use cases and lower the barriers for using new technologies such as VAs. But more importantly, how we design such systems (*e.g., VAs, SBTs, VUIs*) with the potential to invade private spaces could have a great impact on the adoption rate of users and lower the barrier for using such systems. Therefore, we propose designing VAs with the *Human-Centered Design Process* (*HCD*) [4] to address users' requirements and concerns. However, it is not enough to just design with the HCD to overcome users' concerns and have them use the benefits of speech-based technology for their daily routine and different use cases independent of their ability. We also need to understand the triggers and use cases of users' concerns in order to derive design guidelines or patterns to prompt *Acceptance by Design.*

We have started to explore user risks and potential with technology-based users, as this target group shows a range of different interactions as well as privacy concerns [7].

When that is finished, we can derive user requirements not only for the usage, but also for the design of the Voice User Interfaces. To the best of our knowledge, there are no guidelines like those for Graphical Interfaces, such as the "Nielsen Heuristic" or "Shneiderman's 8 golden rules" [2]. The same is true for the design of the system and the evaluation of VAs.

Proposals for measuring the VA user experience exist, but have not yet been evaluated [5, 10, 10].

Hence, we need to explore the contexts of use and derive guidelines to design VAs that are *accepted by design.* We must also develop tools to evaluate our VA in order to improve the adoption rate through the affordance of the design. An example could be using ambient light to encode whether or not the microphone is listening [11]. In addition, we can design prototypes with different users, design concepts, and use cases and employ qualitative (*e.g.,* interviews) and quantitative evaluation (*e.g.,* questionnaires for VUIs [9]) to determine which designs are accepted by users or raise concerns. I also developed systems for user groups that need more protection, such as children [16, 19, 20]. In such cases, not only the user (a child) needs to be considered, but also the legal guardians (*e.g.,* parents) need to be included in the design process. Otherwise, the new system may not be successful if parents do not allow their children to use it.

In other research projects, I focused on how to handle small data in machine learning (prediction and clustering) [14, 15, 24]. This is the first step towards valid and solid results that can be used to make design choices. Next, we could explore ways to use small data analysis to improve speech recognition and response quality.

4 WHAT WOULD I LIKE TO KNOW?

When the aim is to design a system that focuses on the adoption of a new technology or its acceptance through design, the questions asked can include: "*How can one design VAs so that they provide privacy or transparency through their design?*" and "*Can we ensure privacy with our design choices, as is done for light encoding [11]?*"

Attributes could focus on how much data is collected or if the microphone is "*listening*". The main difference between public cameras and technology in a user's home, such as VAs, is the invasion of personal space. Therefore, figuring out how to design a VA system that is accepted and, for this example, focuses on privacy concerns, is also important for any other new technology that captures data from individuals (such as cameras, augmented reality technology, or data glasses).

The main problem seems to be that users do not know if their private conversations are being captured, leading them to ask questions such as: "*How much is captured?*" and "*What is happening with this captured data*?" The lack of transparency around sensitive data handling makes users uncomfortable. At the same time, the data is needed to improve speech recognition and response quality [13].

Hence, there are three possible solutions: propose new ways to increase speech recognition or response quality, collect less data, or design the system to increase transparency. This is a complex and multidisciplinary challenge that requires expertise from different backgrounds, *e.g,* machine learning for small data prediction, design knowledge, and user studies. When these approaches are explored and user needs are understood, the adoption rate for VAs might increase and privacy concerns might decrease.

5 WHAT DO I WANT TO LEARN FROM DIFFERENT DISCIPLINES?

Exploring and designing VAs can benefit from both the well-known methods, artifacts, and evaluation approaches in the domain of the HCD [2, 4] as well as the new approaches of small data analysis in machine learning (ML) [14, 15, 24]. The goal should be to increase the acceptance of the new technology by its design. We must think about further ways to adapt existing methods and develop new methods or artifacts from other disciplines which might not yet be known.

For example, we can use existing methods to collect user requirements and cluster them with the context analysis [12]. The advantage of the context analysis is the structured way in which assumptions are backed up or put into context by users' statements. Additionally, we should consider using the existing UEQplus Framework with newly developed VUI scales to measure VAs [9, 21].

Another way might be to avoid collecting big data with microphones and instead use small data for the machine learning analysis to increase the quality of speech recognition and responses, as is necessary in other domains due to the lack of data [14, 15, 24]. From previous research, we can learn how to use small data for prediction in the VA domain.

To sum up, I would like to know about methods from other disciplines that could be integrated into the design approach of VAs to create better systems and to cluster the requirements of good systems.

6 WHAT DO I WANT TO TEACH OTHER DISCIPLINES?

I want to teach other disciplines how to include user contexts and requirements to design good systems. We are already exploring these areas to better understand VA users' needs. Technology-based users from Germany have access to VAs, but do not use them because of privacy concerns. Hence, users' needs and concerns should be considered to design better systems, thereby reducing concern and increasing the usage rate.

As a result, evaluation tools such as the UEQplus [21] will include VA scale dimensions [9] to support the design process of VAs. This is just one example of how HCD and its methods can help to lower the barriers of new technologies such as VAs and increase their adoption rates.

I also want to raise awareness about how to handle small data analysis in order to prompt other disciplines to use even small data in machine learning [14, 15, 24]. This is because collecting less data might help users to trust more in the system. More controlled data and less data overall could mean, for example, that the data is collected in an initiation phase to improve speech recognition (e.g., due to a dialect). For this example, knowledge from linguists might be helpful to improve the machine learning algorithm, as we have already done in other projects [17, 18].

To sump up, I want to teach other disciplines what can be done already to design better systems and lower the concerns with the design.

REFERENCES

[1] Maresa Biermann, Evelyn Schweiger, and Martin Jentsch. 2019. Talking to Stupid?!? Improving Voice User Interfaces. https://doi.org/10.18420/muc2019-up-0253

[2] Henning Brau and Florian Sarodnick. 2006. *Methoden der Usability Evaluation (Methods of Usability Evaluation)* (2 ed.). Verlag Hans Huber, Bern. 251 pages. http://d-nb.info/1003981860http://www.amazon.com/Methoden-Usability-Evaluation-Henning-Brau/dp/3456842007

[3] BVDW e.V. 2017. Digital Trends Umfrage zu digitalen Sprachassistenten. Bundesverband Digitale Wirtschaft (BVDW) e.V. [Digital Trends Survey on digital language assistants. Federal Association of Digital Economy]. https://www.bvdw.org/themen/publikationen/detail/artikel/digital-trends-umfrage-zu-digitalen-sprachassistenten/. https://www.bvdw.org/themen/publikationen/detail/artikel/digital-trends-umfrage-zu-digitalen-sprachassistenten/

[4] DIN Deutsches Institut für Normung e. V. 2020. *DIN EN ISO 9241-210:2020-03, Ergonomie der Mensch-System-Interaktion - Teil 210: Menschzentrierte Gestaltung interaktiver Systeme; Deutsche Fassung.* Technical Report. Beuth Verlag GmbH. https://doi.org/10.31030/3104744

[5] Kate Hone. 2014. Usability measurement for speech systems : SASSI revisited. "https://www.semanticscholar.org/paper/Usability-measurement-for-speech-systems-%3A-SASSI-Hone/5db24db011fd5b867b95ac29b0b1085dc552eef9"

[6] Bret Kinsella and Ava Mutchler. 2018. Voice Assistant Consumer Adoption Report. https://voicebot.ai/wp-content/uploads/2019/01/voice-assistant-consumer-adoption-report-2018-voicebot.pdf. https://voicebot.ai/wp-content/uploads/2019/01/voice-assistant-consumer-adoption-report-2018-voicebot.pdf

[7] Andreas Klein, Andreas Hinderks, Maria Rauschenberger, and Jörg Thomaschewski. 2020. Exploring Voice Assistant Risks and Potential with Technology-based Users. , 147-154 pages. https://doi.org/10.5220/0010150101470154

[8] Andreas M. Klein. 2021. Toward a User Experience Tool Selector for Voice User Interfaces [DC]. *W4A'21, April 19-20, 2021, Ljubljana, Slovenia* (2021), 2–3. https://doi.org/10.2196/18431.4

[9] Andreas M. Klein, Andreas Hinderks, Martin Schrepp, and Jörg Thomaschewski. 2020. Construction of UEQ+ scales for voice quality. In *Proceedings of the Conference on Mensch und Computer*. ACM, New York, NY, USA, 1–5. https://doi.org/10.1145/3404983.3410003

[10] Andreas M. Klein, Andreas Hinderks, Martin Schrepp, and Jörg Thomaschewski. 2020. Measuring User Experience Quality of Voice Assistants. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*. 1–4. https://doi.org/10.23919/CISTI49556.2020.9140966

[11] Andrii Matviienko, Maria Rauschenberger, Vanessa Cobus, Janko Timmermann, Heiko Müller, Jutta Fortmann, Andreas Löcken, Christoph Trappe, Wilko Heuten, and Susanne Boll. 2015. Deriving design guidelines for ambient light systems. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*, Vol. 30-Novembe. ACM, New York, NY, USA, 267–277. https://doi.org/10.1145/2836041.2836069

[12] Philipp Mayring. 2000. Qualitative Content Analysis. http://www.qualitative-research.net/index.php/fqs/article/view/1089/2385

[13] Martin Porcheron, Joel E. Fischer, and Sarah Sharples. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 207–219. https://doi.org/10.1145/2998181.2998298

[14] Maria Rauschenberger and Ricardo Baeza-Yates. 2020. How to Handle Health-Related Small Imbalanced Data in Machine Learning? *i-com* 19, 3 (2020), 215–226. https://doi.org/10.1515/icom-2020-0018

[15] Maria Rauschenberger and Ricardo Baeza-Yates. 2020. Recommendations to Handle Health-related Small Imbalanced Data in Machine Learning. In *Mensch und Computer 2020 - Workshopband (Human and Computer 2020 - Workshop proceedings)*, Bernhard Hansen, Christian AND Nürnberger, Andreas AND Preim (Ed.). Gesellschaft für Informatik e.V., Bonn, 1–7. https://doi.org/10.18420/muc2020-ws111-333

[16] Maria Rauschenberger, Ricardo Baeza-Yates, and Luz Rello. 2020. Screening Risk of Dyslexia through a Web-Game using Language-Independent Content and Machine Learning. In *W4a'2020*. ACM Press, Taipei, 1–12. https://doi.org/10.1145/3371300.3383342

[17] Maria Rauschenberger, Silke Füchsel, Luz Rello, Clara Bayarri, Jörg Thomaschewski, F Silke, and Luz Rello. 2015. Exercises for German-Speaking Children with Dyslexia. *Human-Computer Interaction–INTERACT 2015* 9296 (2015), 445–452. https://doi.org/10.1007/978-3-319-22701-6

[18] Maria Rauschenberger, Silke Füchsel, Luz Rello, and Jörg Thomaschewski. 2016. A Language Resource of German Errors Written by Children with Dyslexia.. In *The International Conference on Language Resources and Evaluation — LREC 2016*. European Language Resources Association (ELRA), Portorož, Slovenia, 83–87. http://www.lrec-conf.org/proceedings/lrec2016/summaries/136.htmlhttps://repositori.upf.edu/handle/10230/32454

[19] Maria Rauschenberger, Christian Lins, Noelle Rousselle, Sebastian Fudickar, and Andreas Hain. 2019. A Tablet Puzzle to Target Dyslexia Screening in Pre-Readers. In *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good - GOODTECHS*. Valencia, 155–159.

[20] Maria Rauschenberger, Luz Rello, and Ricardo Baeza-Yates. 2019. Technologies for Dyslexia. In *Web Accessibility Book* (2 ed.), Yeliz Yesilada and Simon Harper (Eds.). Vol. 1. Springer-Verlag London, London, 603–627. https://doi.org/10.1007/978-1-4471-7440-0

[21] Martin Schrepp and Jörg Thomaschewski. 2019. Design and Validation of a Framework for the Creation of User Experience Questionnaires. *International Journal of Interactive Multimedia and Artificial Intelligence* (2019), S. 88–95. https://doi.org/10.9781/ijimai.2019.06.006

[22] Serpil Tas, Christian Hildebrandt, and René Arnold. 2019. Voice Assistants in Germany. https://www.wik.org.

[23] Sven Tuzovic and Stefanie Paluch. 2018. *Conversational Commerce – A New Era for Service Business Development?* Springer Fachmedien Wiesbaden, Wiesbaden, 81–100. https://doi.org/10.1007/978-3-658-22426-4_4

[24] Anna Christina Weigand, Daniel Lange, and Maria Rauschenberger. 2021. How can Small Data Sets be Clustered ?. In *Mensch und Computer 2021}{Workshopband}{Workshop on User-Centered Artificial Intelligence (UCAI '21)*. Association for Computing Machinery. https://doi.org/10.18420/muc2021-mci-ws02-284

Engineering a new scholarly ecosystem: security and privacy in speech communication

ANDREAS NAUTSCH, EURECOM, France, now with vitas.ai, Germany

1 BRIEF OVERVIEW; ON EFFORTS PREPARING A SIG COMING INTO PLACE

We can neither avoid that digital societies are becoming more complex rapidly, nor postpone the social contract further that *technology is to aid society.* Such is the responsibility of industry and academia, alike [to aid society]. We know that we need to come together from the different disciplines: but how do we do it? This position paper describes personal experience summarising a time period of about three years.[1] During this time, independent activities led to the formation of the Special Interest Group (SIG) on *Security and Privacy in Speech Communication* (SPSC) within the International Speech Communication Association (ISCA).[2]

Prior, in Interspeech proceedings (formerly named Eurospeech), ISCA research discussed security issues for voice biometrics[3], an early work on security appeared in 1999 [13]—playback and synthesis of spoken digits to subvert voice biometrics—; whereas early works on privacy appeared in 2005 [7] and 2007 [28]—concerning speech recognition from sensor level to segmentation of multi-person and situated spontaneous speech.[4] In subsequent Interspeech proceedings, speeding-up search on large-scale data, binarisation concepts were investigated: cryptographic approaches (not operating on floating point data) are enabled, e.g., by masking speech-derived signals [29] and by masking parameters of probabilistic models [2] (that could generate speech signals; not only recognise biometric identity).

In the past decade, several efforts have been undertaken to foster research on security and privacy in the setting of speech and language technology. Not all of these efforts have been successful. Of the successful initiatives within ISCA and the IEEE Signal Processing Society, which are bridging across fields within and beyond speech technology, one can mention the biannual anti-spoofing challenges coined *ASVspoof* (security of voice biometrics) [6] and the line of PhD students studying privacy-preserving voice biometrics and speech processing [11, 19, 20]. Facing GDPR adoption (in 2016), Interspeech 2015 featured the special event *Privacy Issues in Speech Data Collection and Usage*[5]; shortly after, research on patient privacy started intersecting with secure computation and cryptography leading to the 2018 papers [3, 23]. International projects followed shortly after, such as the H2020 COMPRISE (2018 start), the SECURE research project at Aalborg University (2018 start), and the JST-ANR VoicePersonae (2019 start).

The ISCA SIG-SPSC formed in 2019 at Interspeech, which featured also a special session on *Privacy in Speech and Audio Interfaces* among which's co-organisers was the inaugural chair of SPSC. The SIG is formed not only by people from the above efforts, who met at the (early) 2019 conference IEEE ICASSP. One of the igniting drives for this was created by bringing interdisciplinary experts together through the writing of [15, 16], i.e., co-authors coming from voice biometrics, study of the Law, speech and language technology, secure computation and cryptography, and border control biometrics. To foster its interdisciplinary approach in nurturing multidisciplinary skills of emerging scholars, the SIG was proposed to deliberately become a joint body with other SIGs of ISCA and beyond.

[1] This period spans wrapping up my dissertation to solidifying post-doctoral position abroad, and deciding to stop applying for faculty positions, eventually.

[2] See: www.spsc-sig.org and www.isca-speech.org—ISCA papers are freely online available at www.isca-speech.org/archive, covering decades of research on speech technology and all its technological facets.

[3] Typically, *voice biometrics* is referred to *speaker verification* and *speaker recognition.*

[4] The film industry addressed security and privacy issues regarding 'speech' way earlier; it should not be necessary to mention the Stasimuseum (www.stasimuseum.de) educating about the GDR Ministry for State Security (Stasi) 'listening-in'.

[5] During the panel, a senior researcher mentioned that she found her voice in a movie (taken from a speech synthesis database); being asked is better.

1st AI-Debate Workshop © OvGU, Editors; DOI: 10.25673/38477

2 WHAT DO I KNOW ALREADY?

My main background is in voice biometrics: are two audio files from the same speaker? Comparison results (scores) are thresholded to make yes/no decisions. Yet, antithetical methodological and diverging goal sets are at play:

- The philosophical debate in statistics on quantifying epistemic uncertainty and aleatory uncertainty [18] results in antithetical perspectives on what performance is. In one, error trade-offs are reported and technology integrators are externalised (e.g., [8]). In the other, the perspective on error trade-offs are the basis to cost and information models (e.g., [5, 21]). The latter framework is also used by forensic practitioners to validate how well they prepare evidence when reporting to a judge/jury (regardless of which beliefs a judge/jury might have) [27].
- In ASVspoof (automatic speaker verification anti-spoofing), fake audio detection is investigated for strengthening voice biometrics. There, tandem systems are composed (biometrics with anti-spoofing) [24]. Going beyond related standards [10], tandem performance assessment interlinks subsystem contribution [12].
- Privacy solutions rooted in cryptography and secure multiparty computation have different drawbacks in computational time and time taken to exchange data between servers in IT infrastructures [26]. This contours the real-time demands posed to modern speech technology. Quantisation of evidence representation is a consequence: down to one decision threshold only is facilitated to maintain usability.
- In the VoicePrivacy challenge [25], the privacy-preserving task is to modify/sanitise speech data from biometric features. From a forensic method validation perspective [17], however, when transferring Shannon's original concepts of 'perfect secrecy' [22], core concepts enabling modern cryptography might not hold for statistics assumed inherently known are what speech experts research on since decades.
- Mindsets outline meaning of words, and through this societal impact of technology. 'Quality' is highly contextual. In biometrics standards [9], quality is effectively viewed as functionality of some factory piece, and the conformity of incoming material. On the contrary, when systems are to compensate environmental changes to retain performance, one cannot avoid taking a holistic approach for making design participatory and anticipatory.

To bring experts from different fields together to foster multidisciplinary skill development meets hesitation: while neat on paper, this is entirely antithetical to the academic economy and the core principles of its currencies (paper citations and research grants for very discrete work, not holistic theory). A new scholarly ecosystem is needed that is not only capable, due its experts talking with one another (not only to), but moreover: one that is productive. The pandemic revealed limiting social dynamics *incentivised*. Beneficial social dynamics in counterplay led to slowly building the SPSC community with its first main event in November 2021.[6]

3 HOW DO I STUDY THE PHENOMENON?

One needs to diverge from the norm, and seek to constantly improve. (Seasonally show-casing yet another 5% improvement while tempting does not cause shifts.) Getting active early on in more than two research communities enables personal growth, only through which collectives can grow, then societies. Settings are necessary that allow for constant dialog across disciplines—each expert has core interests placed differently through their individual experience traces.

One can but bring people at a table. For example, the term *biometric data* was put in late into the GDPR which had technological far reaching impacts—when is speech none biometric, if it is without voice or if it is not influenced by our habits how we speak and like to talk about? The European Data Protection Supervisor (EDPS) published several TechDispatches since July 2019 (public EDPS opinions for technologists, and society at large), and its first one addresses

[6] https://spsc-symposium2021.de/

Smart Speakers and Virtual Assistants[7]. In early 2020, Thomas Zerdick (head of unit *Technology and Privacy* at the EDPS) gave the keynote talk at our concept workshop *Privacy: Speech meets Legal Experts.*[8] In March 2021, the European Data Protection Board (EDPB) put their draft *Guidelines 02/2021 on Virtual Voice Assistants* to public consultation; we commented on this as SIG-SPSC based on interdisciplinary expert discussions. In an upcoming event at the Lorentz Center *Speech as Personal Identifiable Information*, we seek to further bridge between communities, namely, usability, speech and language technology, IT-security, policy and governance, and anthropology.

4 WHAT I WOULD LIKE TO KNOW?

How can we come together and make a difference? As a first step, we might inquire information and knowledge (analysis), yet, what we might be seeking actually is understanding and wisdom (synthesis). How can we nourish one another through mutual care that is productive and efficient for lifelong learners?

5 WHAT DO I WANT TO LEARN FROM DIFFERENT DISCIPLINES?

Which methodologies are there to learn from? I want to develop more capacities to better understand human communication, how technology can provide aid, so we can enable progress in societies—not to enforce them to run in circles through making people fit to machine operation. This can be achieved through solving specific problems; these problems can be hypothetical: how to figure out counterfactuals otherwise? For enriching systemic thinking: why and what for attack disciplines problems in their way?

6 WHAT DO I WANT TO TEACH OTHER DISCIPLINES?

We need to substitute teaching with play to stimulate learning through curiosity. Systemic thinking minds are punished by and not rewarded by the present day education megastructure. Compare systems thinking pioneer Russel Ackoff [1]:

- *Creativity is actively suppressed and in most schools conformity—which is anathema to creativity—is valued instead.*
- *Problems do not "belong" to any discipline. [...] The distinction made between science and the humanities (which include the arts) probably does the most harm. [...] They can be viewed and discussed separately, but they cannot be separated.*
- *We should seek wisdom more than anything else—the ability to make value judgments, to know the consequences of our actions, and to learn from our mistakes.*

More crisp in the words of the film critic Wolfgang M. Schmitt: *We only watch, but we do not see.*

A new scholarly ecosystem is in demand. One where young minds are treated as equals. There is no luxury left to afford in unproductive time spent in incentivised first and human rights second activities. *Security and privacy in speech communication* is—as the discrete topic it is—an onboarding; SPSC cannot become truly holistic (it is but speech).

Yet, remedy lies in Deming's Plan-Do-Study-Act cycle [4, 14]. Academic research today got stuck in the Do step, barely completing any analysis of the data that all disciplines generated (the begin of the Study step). How to Act next? To break up silos, to facilitate organising a body of knowledge holistically, a new scholarly ecosystem must emerge.

ACKNOWLEDGMENTS

This work was partially funded by the ANR/JST VoicePersonae project (grant No. JPMJCR18A6).

Thanks to who helped me on this journey, be it providing access to their networks, their collaboration, keeping up regular exchange across the fields, supporting the formation of SPSC, stepping up for service, and leadership.

[7] https://data.europa.eu/doi/10.2804/755512
[8] https://www.spsc-sig.org/2020-01-29-speech-legal-workshop

REFERENCES

[1] R. Ackoff and D. Greenberg. 2008. *Turning Learning Right Side Up: Putting Education Back on Track.* FT Press.

[2] X. Anguera and J.-F. Bonastre. 2010. A Novel Speaker Binary Key Derived from Anchor Models. In *Proc. Interspeech.*

[3] Ferdinand Brasser, Tommaso Frassetto, Korbinian Riedhammer, Ahmad-Reza Sadeghi, Thomas Schneider, and Christian Weinert. 2018. VoiceGuard: Secure and Private Speech Processing. In *Proc. Interspeech.*

[4] W. E. Deming, K. E. Cahill, and K. L. Allan. 2018. *Out of the Crisis.* The MIT Press.

[5] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds. 2000. The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective. *Elsevier Science Speech Communication* 31 (6 2000), 225–254.

[6] N. W. D. Evans, J. Yamagishi, and T. Kinnunen. 2013. Spoofing and countermeasures for speaker verification: a need for standard corpora, protocols and metrics. In *IEEE Signal Processing Society Newsletter.*

[7] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano. 2005. Applications of NAM Microphones in Speech Recognition for Privacy in Human-Machine Communication. In *Proc. Interspeech.*

[8] ISO/IEC JTC1 SC37 Biometrics. 2006. *ISO/IEC 19795-1:2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework.* International Organization for Standardization and International Electrotechnical Committee. confirmed in 2011 and in 2016.

[9] ISO/IEC JTC1 SC37 Biometrics. 2017. *ISO/IEC 2382-37:2017 Information Technology - Vocabulary - Part 37: Biometrics.* International Organization for Standardization.

[10] ISO/IEC JTC1 SC37 Biometrics. 2017. *ISO/IEC 30107-3. Information Technology - Biometric presentation attack detection - Part 3: Testing and Reporting.* International Organization for Standardization.

[11] A. Jimenez. 2019. *An Information Theoretic Approach for Privacy Preservation in Distance-based Machine Learning.* Ph.D. Dissertation. Carnegie Mellon University.

[12] Tomi Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, Md Sahidullah, J. Yamagishi, and D. A. Reynolds. 2020. Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification: Fundamentals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2195–2210.

[13] J. Lindberg and M. Blomberg. 1999. Vulnerability in Speaker Verification - A Study of Technical Impostor Techniques. In *Proc. Eurospeech.*

[14] R. Moen and C. Norman. 2009. The History of the PDCA Cycle. In *Proc. ANQ Congress.*

[15] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans. 2019. The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps towards a Common Understanding. In *Proc. Interspeech.* 3695–3699.

[16] A. Nautsch, A. Jimenez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delcrétaz, G. Chollet, N. Evans, T. Schneider, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch. 2019. Preserving Privacy in Speaker and Speech Characterisation. *Computer Speech and Language, Special issue on Speaker and language characterization and recognition: voice modeling, conversion, synthesis and ethical aspects* 58 (11 2019), 441–480.

[17] A. Nautsch, J. Patino, N. Tomashenko, J. Yamagishi, P.-G. Noé, J.-F. Bonastre, M. Todisco, and N. Evans. 2020. The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment. In *Proc. Interspeech.* ISCA, 1698–1702. https://doi.org/10.21437/Interspeech.2020-1815

[18] T. O'Hagan. 2004. Dicing with the unknown. *Significance* 1, 3 (2004), 132–133. https://doi.org/10.1111/j.1740-9713.2004.00050.x

[19] M. Pathak. 2013. *Privacy-Preserving Machine Learning for Speech Processing.* Ph.D. Dissertation. Carnegie Mellon University.

[20] J. Portêlo. 2015. *Privacy-preserving frameworks for speech mining.* Ph.D. Dissertation. Universidade de Lisboa.

[21] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, and J. Gonzalez-Rodriguez. 2018. Deconstructing Cross-entropy for Probabilistic Binary Classifiers. *Entropy* 20, 3 (3 2018), 208.

[22] C. E. Shannon. 1949. Communication Theory of Secrecy Systems. *Bell System Technical Journal* 28, 4 (10 1949), 656–715.

[23] F. Teixeira, A. Abad, and I. Trancoso. 2018. Patient Privacy in Paralinguistic Tasks. In *Proc. Interspeech.*

[24] M. Todisco, X. Wang, V. Vestman, Md. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee. 2019. ASVspoof 2019: future horizons in spoofed and fake audio detection. In *Proc. Interspeech.* 1008–1012.

[25] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. M. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco. 2020. Introducing the VoicePrivacy Initiative. In *Proc. Interspeech.*

[26] A. Treiber, A. Nautsch, J. Kolberg, T. Schneider, and C. Busch. 2019. Privacy-preserving PLDA speaker verification using outsourced secure computation. *Speech Communication* 114 (2019), 60–71. https://doi.org/10.1016/j.specom.2019.09.004

[27] S. E. Willis, L. Mc Kenna, S. Mc Dermott, A. Barrett, B. Rasmusson, et al. 2015. *ENFSI Guideline for Evaluative Reporting in Forensic Science.* European Network of Forensic Science Institutes. [Online] http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf, accessed: 2017-05-22.

[28] D. Wyatt, T. Choudhury, and J. Bilmes. 2007. Conversation Detection and Speaker Segmentation in Privacy-Sensitive Situated Speech Data. In *Proc. Interspeech.*

[29] X. J. Zhand, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen. 2008. Frequency-Domain Parameter Estimations for Binary Masked Signals. In *Proc. Interspeech.*

Old Problems and New Technology

An Evolutionary Perspective on Emotion in HCI

MICHAEL BRILL, Institute of Human-Computer-Media, University of Würzburg, Germany

CCS Concepts: • **Human-centered computing → Natural language interfaces**; • **Computing methodologies → Philosophical/theoretical foundations of artificial intelligence**; • **Social and professional topics → User characteristics**.

Additional Key Words and Phrases: speech-based technology, interdisciplinarity, social science, humanities, conceptualization, methodology

1 WHAT DO I KNOW ALREADY?

With a background in media psychology and evolutionary psychology, I am accustomed to a biologically grounded perspective that asks for the – speaking in evolutionary terms – ultimate reasons for human experience and behavior. Human beings represent the outcome of a long-lasting process of natural selection. A considerable and rather recent portion of this process occurred when humans existed in nomadic, small groups of hunter-gatherers. Compared to this phylogenetic episode, the short time that has passed since the emergence of modern communication technology resembles the blink of an eye. Within about five generations, verbal communication ceased to be an exclusive domain of interpersonal, face-to-face interaction. It became possible to communicate across space, time, and to large audiences. Just within the last generation, computers were enabled to interact with their users via natural language interfaces in everyday situations. Thus, verbal communication first became detached from the spatial and temporal co-presence of other human beings, and could then occur entirely without a human interaction partner. Because this short, recent period represents a negligible duration for evolutionary processes, the use of modern media technology is presumed to still function based on the archaic mechanisms that constitute the human mind's evolved capacity. Therefore, psychological and communication research has considered the evolved functioning of the human mind in their study of human experience and behavior during interaction with modern media technology. So far, numerous studies have used the frameworks of media equation [7, 8] and computers are social actors (CASA) [4–6] to gather evidence on the social mechanisms that are active during HCI situations.

2 HOW DO I STUDY THE PHENOMENON?

Applying an evolutionary perspective involves asking for the ultimate reasons for an organism's given behavior. To this end, researchers need to hypothesize what adaptive problem of our ancestral environment could have been addressed by this behavior, and how the organism's behavior could have increased the fitness to survive and reproduce in this environment [3]. For emotional aspects of human experience and behavior, Bischof [1] offered a biologically grounded explanation for the emergence of emotion. Bischof models emotions as an evolved mechanism that enables higher organisms to produce appropriate solutions to adaptive problems. While there is – in principle – a variety of mechanisms that can produce adaptive behavior, the model comprehends emotions as a particularly versatile and flexible mechanism that enables organisms to show, for example, complex social behavior. While such adaptive problems may have threatened the organism's capacity to survive and reproduce within the ancestral environment, remnants of these problems – or even the very same problems – can still be found in today's modern environment. It can be argued [10] that the evolved mechanism of emotion is active during media use, as well, which includes the use of advanced interactive media [2]. Bischof's model is rarely used for empirical research, but it offers an informative

1st AI-Debate Workshop © OvGU, Editors; DOI: 10.25673/38478

approach to understanding emotions. Further, it can be used to corroborate existing emotion theories. For example, the component process model of emotion [9] describes emotion as the result of different appraisal steps, with the very first step being the decision if a given stimulus has subjective relevance. Here, we can link subjective relevance to the evolved nature of the emotion system and hypothesize that adaptive problems will most likely be recognized as relevant events. Conversely, if emotions are experienced in a given situation, researchers can investigate if an adaptive problem was present, leading to respective appraisals of the situation. By relying on the established component process model as a framework for experimental studies, it is possible to use standardized emotion self-report measures, and to employ observational methods such as coding of facial actions, as well.

3 WHAT I WOULD LIKE TO KNOW?

So far, Bischof's explanation has been applied to model the use of non-interactive and interactive entertainment media [2, 10]. Since emotions can be a factor in the interaction with speech-based technology, it would be very interesting to apply the model's theoretical framework to this domain, as well. With regard to HCI research, I am interested in the design implications that result from an evolutionary perspective on emotions in the human factor in HCI. Today's HCI situations in everyday life revolve around problems of rather low relevance, such as information on weather and time, or for entertainment purposes. However, with the increasing sophistication and proliferation of speech-based, artificial intelligence technology in professional settings, it can be assumed that future HCI situations will involve tasks of high subjective relevance for the users, and of high objective relevance for their employers. Examples may include the use of 5G-enabled, AI-driven augmented-reality systems in manufacturing or maintenance work, where any error will consume valuable resources. For these settings, it may be of considerable future interest to ask how findings from interpersonal, social interaction in workplace-settings translate to equivalent HCI situations, and how evolved psychological mechanisms shape the interaction processes in HCI in high-stake situations.

4 WHAT DO I WANT TO LEARN FROM DIFFERENT DISCIPLINES?

Apart from the technical aspects of implementing sophisticated HCI work-place situations, I am very interested in the perspectives of other disciplines on emotion in HCI. This includes the respective disciplines' view on emotion in the human factor in HCI, but also the inclusion of simulated emotion in speech-based technology. I am really looking forward to the interdisciplinary exchange with other researchers about their respective experiences, and to learning how they expect future technology to change our interaction with AI.

5 WHAT DO I WANT TO TEACH OTHER DISCIPLINES?

I would like to present and explain Bischof's model for the emergence of emotion, and share this biologically grounded perspective on human behavior. As a very condensed summary, Bischof argues that emotions represent a crucial advancement in higher organisms since they disempower the rather absolute need for the fulfillment of biological drives. Emotions serve as a buffer between current needs of an organism on the one hand, and the production of behavior on the other hand. Therefore, the cognitive systems for planning and production of behavior are now subject to the more diverse and gradual influence by emotions, instead of being influenced by inflexible drives. This way, the cognitive system enjoys greater liberty in producing a wider range of possible solutions to an adaptive problem. This flexibility, together with, for example, proper display systems for an organism's internal emotional state, are an important aspect in the complex functioning of social animals. We can merge this perspective with current appraisal theories of emotion in order to derive hypotheses for specific HCI situations, and then design experimental studies for empirical testing.

REFERENCES

[1] N. Bischof. 1989. Emotionale Verwirrungen – Oder: Von den Schwierigkeiten im Umgang mit der Biologie [Emotional intemplement – or about the difficulties to deal with biology]. *Psychologische Rundschau* 40 (1989), 188–205. https://doi.org/10.5282/ubm/epub.2877

[2] M. Brill, B. P. Lange, and F. Schwab. 2018. Digital sandboxes for stone age minds: Virtual worlds as Bischofian fitness potential landscapes. In *Evolutionary psychology and digital games: Digital hunter-gatherers*, Johannes Breuer, Daniel Pietschmann, Benny Liebold, and Benjamin P. Lange (Eds.). Routledge, New York, 32–48.

[3] M. Brill and F. Schwab. 2020. Evolutionary Reasoning in Communication Scholarship. In *The Handbook of Communication Science and Biology*, Kory Floyd and René Weber (Eds.). Routledge, New York, 93–105.

[4] A. Carolus, J. F. Binder, R. Muench, C. Schmidt, F. Schneider, and S. L. Buglass. 2019. Smartphones as digital companions: Characterizing the relationship between users and their phones. *New Media & Society* 21, 4 (2019), 914–938. https://doi.org/10.1177/1461444818817074

[5] A. Carolus, R. Muench, C. Schmidt, and F. Schneider. 2019. Impertinent mobiles - Effects of politeness and impoliteness in human-smartphone interaction. *Computers in Human Behavior* 93 (2019), 290–300. https://doi.org/10.1016/j.chb.2018.12.030

[6] A. Gambino, J. Fox, and R. A. Ratan. 2020. Building a stronger CASA: Extending the computers Are social actors paradigm. *Human-Machine Communication* 1 (2020), 71–86. https://doi.org/10.30658/hmc.1.5

[7] C. Nass, B.J. Fogg, and Y. Moon. 1996. Can computers be teammates? *International Journal of Human-Computer Studies* 45, 6 (1996), 669–678. https://doi.org/10.1006/ijhc.1996.0073

[8] B. Reeves and C. Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge University Press, Cambridge, UK.

[9] K. R. Scherer. 2009. Emotions are emergent processes: they require a dynamic computational architecture. *Phil. Trans. R. Soc. B* 364 (12 2009), 3459–3474. https://doi.org/10.1098/rstb.2009.0141

[10] F. Schwab. 2010. *Lichtspiele: eine evolutionäre Medienpsychologie der Unterhaltung.* Kohlhammer, Stuttgart, Germany.

Memory performance and text-to-speech functionality

A case for evidence-driven customization?

JENS F. BINDER, Nottingham Trent University, UK
RICHARD BOWEN, Nottingham Trent University, UK

CCS Concepts: • **Human-centered computing → Natural language interfaces**; • **Computing methodologies → Philosophical/theoretical foundations of artificial intelligence**; • **Social and professional topics → User characteristics**.

Additional Key Words and Phrases: speech-based technology, interdisciplinarity, social science, humanities, conceptualization, methodology

1 WHAT DO WE KNOW ALREADY

We present a psychological perspective on memory performance and text-to-speech narration (TSN). Advances in TSN mean that by now speech can be generated at varying levels of naturalness (e.g., human-like or distorted), with different perceived demographic markers (e.g., signaling gender, age, level of education), and at different levels of familiarity (e.g., resembling own voice or modelled on the voice of known others). These advances mean that TSN is now rich in social signals, with potentially wide-ranging effects on human cognition that have rarely been considered to date. Here, we focus on assistive technologies and an approach that is intended to improve their customization.

TSN functionality is an integral part of assistive technologies, often in contexts where memory, learning and retention of information are of concern. This is, unsurprisingly, most prominent where the comprehension of textual material is to be improved, for learning disabilities [21] and visual impairment [16]. Studies have further been extended to less specific disabilities related to interaction and conversation, for example autism [4]. Assistive technologies have also, sometimes independent of TSN functionality, been used to target cognitive impairment and memory directly [6], e.g., for the purpose of rehabilitation. An area of development, therefore, where issues of memory and TSN become most closely intertwined concerns users who have ongoing and complex support needs in their daily lives due to cognitive impairment and would ideally benefit from continuous human-human interaction, such as users with dementia [13, 14, 17].

Memory has been shown to be dependent on social cues at various stages of the encoding and retrieval process, but research on memory performance to date offers little that allows for considering the specific effects of TSN. We therefore draw on psychological models that imply malleable and context-dependent memory performance, to outline further avenues for theory-informed research. Socially motivated information processing and a dynamic memory system are the two fundamentals that suggest that TSN, as it enables human-machine interaction to emulate more and more human-human interaction, can give rise to memory effects that are best documented in the extensive literature on social cognition, group dynamics and social identity.

First of all, humans show different levels of memory performance for stimuli perceived as animate and inanimate [18]. Animacy has been to shown to have an effect on language comprehension and the organization of knowledge [12]. This may be due to richer encoding of animate stimuli [10] and to increased allocation of attention in processing [8]. While the experimental work on animacy and adaptive memory has not considered speech characteristics to date, the theory implies, first of all, TSN effects that rest on the naturalness of language. This encompasses all aspects of TSN that help to create the illusion of natural speech: cadence and flow of speech and more specific prosodic aspects that affect intonation, stress, and rhythm. Such effects should be independent of the time that is needed to get used to

1st AI-Debate Workshop © OvGU, Editors; DOI: 10.25673/38479

artificial-sounding speech. Animacy effects run more deeply and should persist even when users are given time to get acquainted with artificial-sounding speech.

The effects of naturalness are not necessarily novel in the domain of interactive technologies, and are captured by design principles going back as far as the uncanny valley hypothesis [11]. What is novel is that speech generation is now at a point where we can expect characteristics of this particular feature to have measurable effects on user cognition. What is more, we can now draw on theories of memory dynamics in a social context, assuming that speech characteristics can come with any number of social cues, most notably those that suggest membership in a particular social category.

Self-relevance and social categorization in terms of ingroup and outgroup membership are drivers for selective attention and selective memory. Some of these effects are due to ingroup favoritism and refer to, for example, better memory for positive behaviors observed in ingroup members and negative behaviors observed in outgroup members [7]. There is, however, evidence for a general increase in memory performance when the focus is on information relevant to the ingroup rather than the outgroup [3]. In team settings, memory performance has been shown to depend on the team composition (in terms of, e.g., social closeness, but also gender) [1, 2]. These effects are explained by an adaptive process. Individuals adjust their encoding of information, and the organization of knowledge in memory, according to their assumptions about others' knowledge structure [9, 19, 20].

2 HOW WE STUDY THE PHENOMENON AND WHAT WE WOULD LIKE TO KNOW

Research on memory has used a range of tasks in laboratory-based experimentation, and these methods are easily adapted for studying effects of TSN. Our current study, as work in progress, focuses on the factor of naturalness of voice in TSN and its effects on short-term memory. Adapting a standard list learning procedure, participants are presented with four lists of short phrases, each with a different narrator: high and low cadence text-to speech narrators as well as high and low cadence human narrators. Narrators read out the items on these lists, which consist of short three-word phrases containing a noun and a verb i.e. 'warn of politics'. Each list is presented twice, with either a male or female voice, after which the participant is asked to verbally recall any phrases or words they can remember. For TSN phrases, easily accessible standard software is used: the Microsoft Windows Narrator App [5] and Amazon Polly [15]. Responses are scored for their accuracy to the original lists, with full phrases and single words being scored separately. Accuracy scores are then used to calculate to what extent the type of narrator predicts performance in recall.

The experimental set-up allows to test the general hypothesis that short-term recall improves to the extent that narrator's voice becomes more similar to that of a normal human voice. At the same time, the set-up inevitably leaves gaps in our understanding. Among the more immediate lingering points are issues surrounding the type of memory under investigation (e.g., long-term, short-term, episodic, semantic), the specific nature of the stimulus materials (e.g., naturalistic, standardized, rich in content, and so forth), and the definition of what counts as more or less naturalistic voice.

3 WHAT WE WANT TO LEARN FROM OTHER DISCIPLINES AND WHAT WE WANT TO TEACH THEM

The shortcomings of focused experiments, such as the comparatively slow and incremental progress in understanding, is likely to be off-set, we argue, by the promise of arriving at robust recommendations for the customization of TSN, and potentially other technologies that feature voice and speech generation.

Next to a compelling experimental design, however, the overall outcome of such research depends on the quality and relevance of the materials and stimuli used. Other disciplines can provide crucial input for this: current standards in speech generation, the adaptability of generated speech to that of specific individuals, relevant speech characteristics that lend themselves to systematic variation in experiments, and systems of customization as they can present themselves to end users are all aspects of such input. In particular, we feel the need to go beyond freely available software solutions as they are used here and as they are frequently used in psychological research. Further, next to isolated laboratory-based research, more interdisciplinary studies on implementation and evaluation of TSN technologies in the field are also much needed opportunities that will help to identify other relevant factors to take back into controlled experimentation.

An experimental approach is not unique to psychology and can be found in many disciplines with an interest in behavioral indicators. A more unique aspect of the psychological perspective promoted here consists in the theoretical underpinnings. Theories and models of memory performance have been shaped in close co-evolution with experimental methods, but they also provide frames for orientation in applied research. Put differently, recommendations for customization will be most enduring when they are backed by evidence and a compelling theoretical explanation. Recommendations could focus on the appropriate level of liveliness, on the similarity or dissimilarity with others close to the user, on auditory ingroup and outgroup markers etc. Empirical evidence can tell us, for example, which of the following notions is more likely to be the case: that memory performance is better when speech is modelled on a close other (due to familiarity, levels of attention and motivation, activation of existing memories in the system, social compliance, and so forth), or that memory is better when speech is modelled on an unknown other (due to novelty, the need to focus more for understanding, the deeper level of elaboration, conversational norms, and so forth). Our approach can help to further specify and refine such notions, thereby turning them into more robust design recommendations. This is where psychology is likely to be of wider value to other disciplines concerned with TSN and memory.

As technologies enable exchanges that increasingly resemble human-human interaction the boundary between human-machine and human-human settings will become increasingly blurred, at least in the eyes of users. Theories and findings regarding users' mental models of human interaction partners will therefore become more important. The social aspects of memory performance provide a first testing ground for these predictions.

REFERENCES

[1] J. Andersson. 2001. Net effect of memory collaboration: how is collaboration affected by factors such as friendship, gender and age? *Scand J Psychol.* 42, 4 (2001), 367–75. https://doi.org/1467-9450.00248

[2] J. Andersson and J. Ronnberg. 1997. Cued Memory Collaboration: Effects of Friendship and Type of Retrieval Cue. *European Journal of Cognitive Psychology* 9, 3 (1997), 273–287. https://doi.org/10.1080/713752558

[3] J. J. Van Bavel and W. A. Cunningham. 2012. A Social Identity Approach to Person Memory: Group Membership, Collective Identification, and Social Role Shape Attention and Memory. *Personality and Social Psychology Bulletin* 38, 12 (2012), 1566–1578. https://doi.org/0146167212455829

[4] SA Cassidy, B Stenger, L Van Dongen, K Yanagisawa, R Anderson, V Wan, S Baron-Cohen, and Cipolla R. 2016. Expressive visual text-to-speech as an assistive technology for individuals with autism spectrum conditions. *Comput Vis Image Underst* 148 (2016), 193–200.

[5] Microsoft Corporation. 2020. Complete guide to Narrator. Retrieved July 16, 2021 from https://support.microsoft.com/en-us/windows/complete-guide-to-narrator-e4397a0d-ef4f-b386-d8ae-c172f109bdb1

[6] B.-K. Dewar, M. Kopelman, N. Kapur, and B. A. Wilson. 2014. Assistive technology for memory. In *Assistive technology for cognition: A handbook for clinicians and developers*, B. O'Neill and A. Gillespie (Eds.). Routledge, New York, 31–46.

[7] J. W. Howard and M. Rothbart. 1980. Social categorization and memory for in-group and out-group behavior. *Journal of Personality and Social Psychology* 38, 2 (1980), 301–310.

[8] JK Leding. 2019. Adaptive memory: Animacy, threat, and attention in free recall. *Mem Cognit* 47, 3 (2019), 383–394. https://doi.org/10.3758/s13421-018-0873-x

[9] S. B. Marion and C. Thorley. 2016. A meta-analytic review of collaborative inhibition and postcollaborative memory: Testing the predictions of the retrieval strategy disruption hypothesis. *European Journal of Cognitive Psychology* 142, 11 (2016), 1141–1164. https://doi.org/10.1037/bul0000071

[10] M. J Meinhardt, R. Bell, A. Buchner, and J. P. Röer. 2020. Adaptive memory: Is the animacy effect on memory due to richness of encoding? *J Exp Psychol Learn Mem Cogn* 46, 3 (2020), 416–426. https://doi.org/10.1037/xlm0000733

[11] M. Mori. 1970. Bukimi no tani ["The uncanny valley"]. *Energy* 7, 4 (1970), 33–35.

[12] J. S. Nairne, J. E. VanArsdall, and M. Cogdill. 2017. Remembering the Living: Episodic Memory Is Tuned to Animacy. *Current Directions in Psychological Science* 26, 1 (2017), 22–27. https://doi.org/10.1177/0963721416667711

[13] S. Nakatani, S. Saiki, M. Nakamura, and K. Yasuda. 2018. Generating personalized virtual agent in speech dialogue system for people with dementia. In *International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*. Springer, Cham, 326–337.

[14] A. Russo, G. D'Onofrio, A. Gangemi, F. Giuliani, M. Mongiovi, F. Ricciardi, F. Greco, F. Cavallo, P. Dario, D. Sancarlo, V. Presutti, and A. Greco. 2019. Dialogue Systems and Conversational Agents for Patients with Dementia: The Human-Robot Interaction. *Rejuvenation Research* 22, 2 (2019), 109–120. https://doi.org/10.1089/rej.2018.2075

[15] Amazon Web Services. 2021. Amazon Polly. Retrieved July 16, 2021 from https://aws.amazon.com/polly/

[16] R. F. Sharma and S. G. Wasson. 2012. Speech recognition and synthesis tool: assistive technology for physically disabled persons. *International Journal of Computer Science and Telecommunications* 3, 4 (2012), 86–91.

[17] F. Sposaro, J. Danielson, and G Tyson. 2010. iWander: An Android application for dementia patients. *Annu Int Conf IEEE Eng Med Biol Soc* 2010 (2010), 3875–8. https://doi.org/10.1109/IEMBS.2010.5627669

[18] J. E. VanArsdall, J. S. Nairne, J. N. S. Pandeirada, and J. R. Blunt. 2013. Adaptive memory: animacy processing produces mnemonic advantages. *Exp Psychol.* 60, 3 (2013), 172–8. https://doi.org/10.1027/1618-3169/a000186

[19] D. M. Wegner. 1987. *Transactive memory: A contemporary analysis of the group mind. In Theories of group behavior.* Springer, New York, NY.

[20] D. M. Wegner, R. Erber, and P. Raymond. 1991. Transactive memory in close relationships. *Journal of Personality and Social Psychology* 61, 6 (1991), 923–929. https://doi.org/doi/10.1037/0022-3514.61.6.923

[21] S. G. Wood, J. H. Moxley, E. L. Tighe, and R. K. Wagner. 2018. Does use of text-to-speech and related read-aloud tools improve reading comprehension for students with reading disabilities? A meta-analysis. *Journal of Learning Disabilities* 51, 1 (2018), 73–84. https://doi.org/10.1177/0022219416688170

Towards a holistic approach and measurement of humans interacting with speech-based technology

ASTRID CAROLUS, Media Psychology, Institute of Human-Computer-Media, University of Würzburg, Germany
CAROLIN WIENRICH, Human-Technique Systems, Institute of Human-Computer-Media, University of Würzburg, Germany

CCS Concepts: • **Human-centered computing → Natural language interfaces**; • **Computing methodologies → Philosophical/theoretical foundations of artificial intelligence**; • **Social and professional topics → User characteristics**.

Additional Key Words and Phrases: speech-based technology, user experience, social science, conceptualization, methodology

1 TECHNOLOGY AS PSYCHOLOGICAL RELEVANT INTERACTION PARTNERS

For a long time, the perspective of both public as well as scientific discussions of successful digitalization was limited to technical equipment and technical competences. If the human users were considered, it would rather be in a limited way focusing only on the ability to think rationally. Humans were predominantly referred to in the light of the “Homo oeconomicus” emphasizing the rational capabilities of humans in terms of their rational thinking, decision-making and action. Underlying emotional or motivational aspects were rather disregarded. From this perspective, technology was conceptualized as a tool that humans use to aim at their specific targets and to achieve their goals. Accordingly, interactions with technology were usually evaluated according to pragmatic aspects. The usability approach reflects this focus on the efficient and effective achievement of goals thereby meeting the requirements this “rational perspective” postulates. However, research has questioned this limited perspective on humans interacting with media and computer since the 1990s. Conceptualizing "computers as social actors" (CASA) studies have shown that technological devices trigger social reactions in their human users that are similar to those in human-human interactions. With their basic assumption that “media equals real life”, Reeves and Nass (1996) introduced their concept of “media equation” postulating that people involuntarily tend to treat media entities like real people [9]. Analyzing media devices (e.g., computers, smartphones, robots) and persons appearing in media content (e.g. movie characters, news anchor), studies in this research area showed that social norms, which apply to human-human interactions and which are well-studied in social science research also apply in the context of media (see also “parasocial interaction”; [4]). Other studies revealed that social-cognitive processes can be transferred to digital interaction partners [11, 14].

2 WHAT DO I KNOW ALREADY?

This widening of perspectives on the interaction of humans and technology resulted in the conceptualization of digital technologies as psychologically relevant counterparts users do not only use but interact with. Referring to the high frequency of use in terms of recurring interactions with the device, Carolus et al. (2019) investigated a possible social relationship of the users with their smartphones, which they referred to as “digital companions” [2]. Besides the pragmatic functions the phone offers, they provided empirical evidence that the users seem to feel like they were in a social relationship with their device. Following the tradition of CASA studies, characteristics of social relationships (e.g., closeness, trust) and their outcomes (e.g., stress, coping with stress) were adapted to yield a model of human–smartphone relationships, which was empirically confirmed. As a result, and with the focus on the human users’ cognitions, emotions and actions, human-technology interaction gets closer to human-human interaction. The ongoing

1st AI-Debate Workshop © OvGU, Editors; DOI: 10.25673/38480

technological progress underlines the approximation insofar as technology itself seems to get closer to the human being. Modern technology adopts "human-like" attributes, which further strengthens the (conscious and unconscious) perception that the technological counterpart “equals real life” [3, 13].

Consequently, the former “rational perspective” on human users and their interactions with technologies are fundamentally expanded by a more holistic understanding, which involves not only the human cognitive capabilities but also the emotional and the motivational functioning as well as social needs and mechanisms. The concept of user experience (UX) does more justice to this complexity in that it more strongly "encompass[s] all emotions, perceptions, preferences, perceptions, physiological and psychological reactions, behaviors, and performances [that] occur before, during, and after use" [1, 6]. However, the majority of methodological approaches and measures in this area is limited to hedonic aspects, such as pleasure in interaction or stimulation [5, 7].

Recent contributions increasingly consider eudaimonic aspects, aiming for personal goals through technology use in terms of need fulfillment or meaning [8] or social aspects, aiming for relatedness [10, 12]. Although recent UX studies have expanded the perspective substantially, they still miss to meet the whole spectrum of requirements the early and the more recent CASA studies as well as the recent technological improvements suggest. Human-computer interaction, which increasingly resembles human-human interaction, expands its power and its sphere of influence. Just as human interactions are understood as social encounters, which affect the human on cognitive, emotional and motivational levels, human-computer interactions must also be conceptualized and analyzed in these dimensions. In this way, first studies revealing experiences that cannot be classified as either hedonic or eudaimonic but are coded as "social" experiences are regarded as first indications of this idea [6, 8]. However, a systematic reappraisal aiming for a holistic theoretical as well as methodological approach to human-technology interaction is still pending.

3 HOW DO I STUDY THE PHENOMENON?

As social scientists we conduct experimental studies in the laboratory as well as online or conduct (online) surveys. Most studies are limited to only one measurement point. Long-term studies are just as rare as indirect survey methods (e.g., implicit testing, physiological measures) or field observations. A typical approach is to manipulate the appearance of the technological counterpart, for example, to analyze the participants’ reaction to it. In an online experiment we asked for the capacity of smart speakers to elicit empathy and showed that participants watching a smart speaker being treated rudely results in significantly higher ratings of empathy with the device [3]. In another study, different human-robot interactions in a work-context were simulated in virtual reality. The results revealed strong effects of the robot design on need fulfillment and gender [12].

4 WHAT I WOULD LIKE TO KNOW?

With an emphasis on theory and methodological implications we would like to learn more about how researchers from different disciplines tackle the challenges in the area of speech-based systems. We would like to provide our holistic perspective (1) on HCI in general and (2) humans interacting with speech-based technology in particular to researchers from other disciplines to get to know their point of view.

5 WHAT DO I WANT TO LEARN FROM DIFFERENT DISCIPLINES?

As research on speech-based technology is only at the beginning of the process to become an established area of research we recognize a momentum to address the research subject in its broadness to develop research programs, which combine the expertise from the variety of disciplines involved in the development, analysis or understanding of speech-based technology – but also the disciplines being rather indirectly affected (e.g. pedagogy, gerontology). In this sense, we want to derive and develop research concepts and approaches that are truly interdisciplinary and, long-term oriented and sustainable. One very first idea could be the collection and analysis of the theoretical foundations (and methodological approaches) in the field to provide a first overview of the status quo and to lay the foundation for future research in the field.

6 WHAT DO WE WANT TO TEACH OTHER DISCIPLINES?

We want to broaden the perspective on user experience to meet the whole spectrum of requirements the early and the more recent CASA studies as well as the recent technological improvements suggest.

REFERENCES

[1] Javier A Bargas-Avila and Kasper Hornbæk. 2011. Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2689–2698.

[2] Astrid Carolus, Jens F Binder, Ricardo Muench, Catharina Schmidt, Florian Schneider, and Sarah L Buglass. 2019. Smartphones as digital companions: Characterizing the relationship between users and their phones. *New Media & Society* 21, 4 (2019), 914–938.

[3] Astrid Carolus, Carolin Wienrich, Anna Toerke, Tobias Friedel, Christian Schwietering, et al. 2021. 'Alexa, I feel for you!'-Observers' Empathetic Reactions towards a Conversational Agent. *Frontiers in Computer Science* 3 (2021), 46.

[4] David C Giles. 2002. Parasocial interaction: A review of the literature and a model for future research. *Media psychology* 4, 3 (2002), 279–305.

[5] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & computer 2003*. Springer, 187–196.

[6] Marc Hassenzahl, Sarah Diefenbach, and Anja Göritz. 2010. Needs, affect, and interactive products–Facets of user experience. *Interacting with computers* 22, 5 (2010), 353–362.

[7] Mare Hassenzahl, Axel Platz, Michael Burmester, and Katrin Lehner. 2000. Hedonic and ergonomic quality aspects determine a software's appeal. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 201–208.

[8] Elisa D Mekler and Kasper Hornbæk. 2016. Momentary pleasure or lasting meaning? Distinguishing eudaimonic and hedonic user experiences. In *Proceedings of the 2016 chi conference on human factors in computing systems*. 4509–4520.

[9] Byron Reeves and Clifford Nass. 1996. *The media equation: How people treat computers, television, and new media like real people.* Cambridge university press Cambridge, United Kingdom.

[10] Carolin Wienrich and Johanna Gramlich. 2020. AppRaiseVR–An Evaluation Framework for Immersive Experiences. *i-com* 19, 2 (2020), 103–121.

[11] Carolin Wienrich, Richard Gross, Felix Kretschmer, and Gisela Müller-Plath. 2018. Developing and proving a framework for reaction time experiments in VR to objectively measure social interaction with virtual agents. In *2018 IEEE conference on virtual reality and 3D user interfaces (VR)*. IEEE, 191–198.

[12] Carolin Wienrich and Marc Erich Latoschik. 2021. eXtended Artificial Intelligence: New Prospects of Human-AI Interaction Research. *Frontiers in Virtual Reality* (2021).

[13] Carolin Wienrich, Clemens Reitelbach, and Astrid Carolus. 2021. The Trustworthiness of Voice Assistants in the Context of Healthcare: Investigating the Effect of Perceived Expertise on the Trustworthiness of Voice Assistants, Providers, Data Receivers, and Automatic Speech Recognition. *Frontiers in Computer Science* 3 (2021), 53.

[14] Carolin Wienrich, Kristina Schindler, Nina Döllinqer, Simon Kock, and Ole Traupe. 2018. Social presence and cooperation in large-scale multi-user virtual reality-the relevance of social interdependence for location-based environments. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 207–214.