CENTRIC 2021

The Fourteenth International Conference on Advances in Human oriented and Personalized Mechanisms, Technologies, and Services

ISBN: 978-1-61208-896-9

October 3 -7, 2021

Barcelona, Spain

CENTRIC 2021 Editors

Lasse Berntzen, University of South-Eastern Norway, Norway

Stephan Böhm, University of Applied Sciences Wiesbaden, Germany

Jeff Stanley, The MITRE Corporation, USA

CENTRIC 2021

Forward

The Fourteenth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC 2021), held on October 3 - 7, 2021 in Barcelona, Spain, addressed topics on human-oriented and personalized mechanisms, technologies, and services, commonly known as I-centric.

There is a cohort of technologies that favored the so called “user-centric” services and applications. While some of them reached some maturity, others are to prove their economics (WiMax, IPTV, RFID, etc). The human-oriented and personalized technologies and services rely on a key set of features, some to be deployed, others getting more mature (personal profiles, preferences, identity, proximity, personal devices, etc.). Following, advanced applications covering human related activities benefit from personalized and human-oriented networks and services, especially preventive and personalized medicine, body networks and devices, or anticipative systems.

The conference provided a forum where researchers were able to present recent research results and new research problems and directions related to them. The conference sought contributions presenting novel result and future research in all aspects of user-centric mechanisms, technologies, and services.

Similar to the previous editions, this event continued to be very competitive in its selection process and very well perceived by the international community. As such, it attracted excellent contributions and active participation from all over the world. We were very pleased to receive a large amount of top quality contributions.

We take here the opportunity to warmly thank all the members of the CENTRIC 2021 technical program committee as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and efforts to contribute to the CENTRIC 2021. We truly believe that thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the CENTRIC 2021 organizing committee for their help in handling the logistics and for their work that is making this professional meeting a success.

We hope the CENTRIC 2021 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in personalization research.

CENTRIC 2021 Steering Committee

Stephan Böhm, RheinMain University of Applied Sciences - Wiesbaden, Germany
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Yasushi Kambayashi, NIT - Nippon Institute of Technology, Japan
Jeff Stanley, The MITRE Corporation, McLean, USA

CENTRIC 2021 Publicity Chair

José Miguel Jiménez, Universitat Politecnica de Valencia, Spain
Lorena Parra, Universitat Politecnica de Valencia, Spain

CENTRIC 2021

Committee

CENTRIC 2021 Steering Committee

Stephan Böhm, RheinMain University of Applied Sciences - Wiesbaden, Germany
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Yasushi Kambayashi, NIT - Nippon Institute of Technology, Japan
Jeff Stanley, The MITRE Corporation, McLean, USA

CENTRIC 2021 Publicity Chair

José Miguel Jiménez, Universitat Politecnica de Valencia, Spain
Lorena Parra, Universitat Politecnica de Valencia, Spain

CENTRIC 2021 Technical Program Committee

Youssef A. Attia, King Abdulaziz University, Saudi Arabia
Stefania Bandini, RCAST - Research Center for Advanced Science & Technology | The University of Tokyo, Japan
Samir Brahim Belhaouari, Hamad Bin Khalifa University, Doha, Qatar
Lasse Berntzen, University of South-Eastern Norway, Norway
Stephan Böhm, RheinMain University of Applied Sciences - Wiesbaden, Germany
Daniel B.-W. Chen, Monash University, Australia
Sabine Coquillart, INRIA, France
Marco Costanzo, Università degli Studi della Campania "Luigi Vanvitelli", Italy
Carlos Cunha, Polytechnic Institute of Viseu, Portugal
Rui Pedro Duarte, Polytechnic institute of Viseu, Portugal
Luciane Fadel, Federal University of Santa Catarina, Brazil
Rainer Falk, Siemens AG Corporate Technology, Germany
Filipe Fidalgo, Polytechnic Institute of Castelo Branco, Portugal
Alicia García-Holgado, GRIAL Research Group - University of Salamanca, Spain
Faisal Ghaffar, IBM Ireland / University College Dublin, Ireland
Till Halbach, Norwegian Computing Center, Norway
Qiang He, Swinburne University of Technology, Australia
Koen Hindriks, Vrije Universiteit Amsterdam, Netherlands
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Shigao Huang, University of Macau, Macao SAR, China
Takeshi Ikenaga, Kyushu Institute of Technology, Japan
Imène Jraidi, Advanced Technologies for Learning in Authentic Settings (ATLAS) Lab | McGill University, Montreal, Canada
Christos Kalloniatis, University of the Aegean, Greece
Yasushi Kambayashi, NIT - Nippon Institute of Technology, Japan
Mazaher Kianpour, Norwegian University of Science and Technology (NTNU), Norway

Boris Kovalerchuk, Central Washington University, USA
Cun Li, Eindhoven University of Technology, Netherlands
Stanislav Mamonov, Feliciano School of Business, USA
Ângela Cristina Marques de Oliveira, Instituto Politécnico de Castelo Branco, Portugal
Pierre-François Marteau, Université Bretagne Sud / Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA), France
Célia Martinie, Université Paul Sabatier Toulouse III, France
José Martins, LIAAD - INESC TEC / Polytechnic of Leiria, Portugal
Thomas Marx, TH Bingen, University of Applied Sciences, Germany
Erik Massarczyk, RheinMain University of Applied Sciences Wiesbaden Rüsselsheim, Germany
David Melhart, Institute of Digital Games | University of Malta, Malta
Pedro Merino, ITIS Software | University of Malaga, Spain
Toshiro Minami, Kyushu Institute of Information Sciences, Japan
Eduardo Miranda, University of Plymouth, UK
Fatma Najar, Concordia University, Montreal, Canada
Areolino Neto, Federal University of Maranhão, Brazil
Helder C. R. Oliveira, University of Calgary, Canada
Monica Perusquía-Hernández, NTT Communication Science Laboratories, Japan
Stefan Pickl, Universität der Bundeswehr München, Germany
Melissa Ramos da Silva Oliveira, University of Vila Velha, Brazil
Jagat Jyoti Rath, Instituteof Infrastructure Technology Research and Management (IITRAM), Ahmedabad, India
Valentim Realinho, Instituto Politécnico de Portalegre, Portugal
Ann Reddipogu, RKode Ltd, UK
Michele Risi, University of Salerno, Italy
Armanda Rodrigues, NOVA LINCS | Universidade NOVA de Lisboa, Portugal
Aurora Saibene, University of Milano - Bicocca, Italy
Sandra Sanchez-Gordon, Escuela Politécnica Nacional, Ecuador
Jungpil Shin, The University of Aizu, Japan
Marie Sjölinder, RISE, Sweden
Alfredo Soeiro, University of Porto - FEUP, Portugal
Jeff Stanley, The MITRE Corporation, McLean, USA
Mu-Chun Su, National Central University, Taiwan
Patricia Torrijos Fincias, University of Salamanca, Spain
Carlos Travieso González, University of Las Palmas de Gran Canaria, Spain
Alberto Vergnano, University of Modena and Reggio Emilia, Italy
Christina Volioti, Aristotle University of Thessaloniki / University of Macedonia / International Hellenic University, Greece
Hao-Chun Yang, National Tsing Hua University, Taiwan
Christos Zaroliagis, University of Patras, Greece
Alejandro Zunino, ISISTAN-CONICET | Universidad Nacional del Centro (UNICEN), Argentina

Copyright Information

Table of Contents

Inclusive Personalities for Conversational User Interfaces: A Preliminary Discussion

Jennifer Strickland, Jeff Stanley
The MITRE Corporation
Bedford, MA, USA
email:{jstrickland, jstanley}@mitre.org

***Abstract*—Conversational user interfaces (CUIs) such as chatbots and voice assistants are increasingly used to deliver services not just in industry but in government. Therefore, it is increasingly important for CUIs to provide good experiences for constituents with diverse backgrounds and abilities. Existing research on CUI personality focuses on engaging typical target users. Synthesizing existing literature on CUI personalities with principles for inclusive design, we discuss how to design CUI personalities that provide good experiences for diverse users. Key considerations are to consider the user's unique situation, their expectations and preferences toward technology, and their purpose in using the technology. Our intent is to identify challenges for future research and to move towards a set of guidelines for inclusive CUI design.**

Keywords- chatbot; personality; inclusive design; equitable design; cross-cultural design; accessibility.

I. Introduction

Conversational User Interfaces (CUIs) such as text-based chatbots and voice-based assistants have become a popular solution for commercial services and are increasingly used to deliver government services as well. While companies are motivated to design CUI personalities that reflect their brands and engage target customers, government services must be accessible to all constituents. Therefore, in this discussion, we go beyond the question of how to craft a CUI personality that relates well to users: We focus on how to craft a CUI personality that relates well to diverse groups of users with disparate needs, wants, and expectations. In other words, how does a CUI's personality include or exclude sections of the population, and what research questions should be answered to ensure CUIs do not unintentionally alienate the people being served?

Some existing research explores how a CUI's content and interface should account for diverse needs, often by adapting existing web content standards to the complexities of CUIs [1][2]. These include standards for fonts and colors on the screen, reading level for text content, how elements can be navigated on a webpage, and how they should be labeled and placed. However, existing standards do not explicitly address the novel problem space of artificial personality.

In this paper, we bring together research on CUI personality with principles for inclusive design and introduce topics to consider when designing inclusive CUI personalities. Our goal is to take a step towards guidelines for CUI personalities that serve all people.

Section II introduces inclusive design principles and CUI personality and describes how the former can be applied to the latter. Section III discusses some of the challenges involved in designing CUI personalities that satisfy inclusive design principles. Section IV offers recommendations to help manage those challenges. Section V concludes the paper.

II. Background

A. Inclusive Design

Inclusive design is an approach seeking to ensure all can access and are included in the design and outcome of a service or product. This perspective encompasses ability, age, gender, culture, ethnology, race, socioeconomics, power, and vulnerability, among other characteristics. Inclusive design practitioners are expected to investigate their own biases, hire diverse teams, and consider "design for/with/by" approaches to process. Using design tools, frameworks, language, and processes that are accessible and inclusive is a key tenet of inclusive design. At the start of a project, it is a best practice to define a list of who the outcome may exclude, then use that to guide decision-making. Being mindful of who is included or excluded is a key guidepost.

In service of inclusive design, it is helpful to take a "design by" approach, bringing the service consumer into the design of outcomes. Hiring for lived experience is a tremendous asset to bringing awareness to inclusive processes. Design workshops that bring consumers into the design and development processes are incredibly valuable in ensuring outcomes serve diverse audiences.

Government services are especially relevant for inclusive design due to the range of experiences served. Where else does a service have an audience with such diversity in geography, culture, economics, ability, etc.? Most commercial products are willing to exclude those with low incomes, yet those are some of the critical users for whom government services may be a matter of life or death. Many government agencies already use CUIs to deliver or supplement key public services [3].

ISBN: 978-1-61208-896-9

B. CUI Personality

Personality for CUIs, in a broad sense, is a topic of interest for researchers and industry. Personality shapes a CUI's response content, either by carefully designing each piece of content [4] or by training the CUI's language model on a particular data set [5]. Web-based CUIs often have a visual component like a headshot that can reflect a particular kind of personality. When considering voice-based systems, different voice types can similarly reinforce different kinds of personalities [6]. Some industry experts offer strategies for how to design CUI personalities. These include identifying personality traits the CUI should have, which can be based on established models of personality [7] or brand values [8]; and identifying kinds of people to use as models for the CUI's behavior. Persson et al. [9] refer to these two strategies as *trait schemas* versus *social role schemas*; though it is possible to use both together, for instance as recommended by Google [10] when developing for Google Assistant.

C. Application of Inclusive Design to CUI Personality

Community experts provide six Inclusive Design Principles [11]; here, we give examples to illustrate their applicability to CUI personality. As we discuss challenges in this paper, we will refer to the principles most applicable to each.

1. **Provide comparable experience**: A CUI should use simple straightforward language so that people who cannot fluently read the CUI's language can complete tasks with success similar to those who can.
2. **Consider situation**: A CUI should use empathy if users are likely to be under pressure.
3. **Be consistent**: A CUI should adhere to familiar conversational conventions, such as Grice's maxims (see [12]).
4. **Give control**: A CUI should give the user plenty of opportunities to steer the conversation.
5. **Offer choice**: A CUI should be responsive to different language styles and registers.
6. **Prioritize content**: A CUI should convey only content most relevant to the conversation topic so the user can stay focused.
7. **Add value**: A CUI should not engage in talk or offer conversation paths that do not improve user experience or satisfaction.

III. Challenges for Inclusive CUI Personalities

A. Grace, Respect, Empathy, and Mindful Language

What sort of personality will best serve the user's purpose and scenario? That is likely to vary depending on the individual's perspective, which may itself vary based on culture, gender, age, ability, or any of several factors. To bring grace, respect, and empathy to the CUI personality, the design team must conduct inclusive research with a broad range of human experience to design mindful, effective (and possibly affective) conversation.

Empathy can improve adoption of CUIs and improve human mood [13][14]. However, inaccurate empathy such as unmerited sympathy can decrease the user's trust [15].

Consider how a person's background may influence the perception of personality, and how that might impact the acceptance of a CUI. Taking a casual tone may be perceived as disrespectful or create comfort; using dark humor could build rapport or offend; over time the bot's personality could adapt to the relationship's evolution or maintain a purely transactional perspective, depending upon the goal of the CUI service and user needs.

Follow the Inclusive Design Principles, "provide comparable experience," "consider situation," "be consistent," "give control," "offer choice," and "add value."

B. User's Self-Identification

Imagine, if you will, that a CUI refers to you regularly as a different gender than you identify as, or refers to abilities that you do not have; how would you feel? An individual's identity is a personal statement reflecting their history, experience, values, and mission. How might a CUI welcome the full range of human identity, which may vary in language, lingo, tone, and even code switching?

When designing a CUI's personality, the development team should be aware of any biases and stereotypes informing the design and how this could interact with users' self-identification. For instance, a digital assistant modeled after a young female secretary might appeal to certain users but offend others [16].

Follow the Inclusive Design Principles, "consider situation," "give control," and "offer choice."

C. User's Situation and Mood

There are situations that may be particularly stressful for people, such as navigating an unfamiliar city. Google Maps anticipated this by offering character voices such as Morgan Freeman or Santa Claus, which can defuse tension. Additionally, conversations between passengers and drivers tend to be simple and concise to account for their divided attention [17][18].

The user's mood, like situation, affects conversational priorities. While an impatient user needs answers quickly, other users might appreciate additional content acknowledging their emotional state, such as potential targets of fraud [19].

Follow the Inclusive Design Principles, "consider situation," "be consistent," and "prioritize content."

D. Politeness

What level of formality and politeness should a CUI show its human user? The wrong level of politeness in language and behavior can easily offend or annoy, such as over-politeness among friends or rudeness among acquaintances.

Politeness theory distinguishes between positive and negative *face*. Positive face can be thought of as the desire for affirmation and acceptance, while negative face can be thought of as the desire to maintain personal autonomy. Polite language such as "if you don't mind" appeals to negative face, allowing room to politely refuse [20]. However, politeness is more than specific phrases. It is important to identify the range of face needs for the CUI's intended users. Someone reporting a scam may feel ashamed of having been fooled. The CUI can consider the user's positive face by showing empathy and understanding [19]. Meanwhile, technological assistants for people with disabilities need to consider negative face and assist only as needed and requested [21].

Humans expect the politeness of an interaction to be appropriate to the social relationship between the two parties [20]. Therefore, it is important to ask first whether users are likely to approach the CUI as a social partner, and if so whether the CUI is viewed as a close peer or as a formal representative of some organization.

Follow the Inclusive Design Principles, "consider situation," "give control," "offer choice," and "add value."

E. Different Interaction Styles and Preferences

When speaking with CUIs assisting with chronic disease management, patients preferred different healthcare provider interaction styles, such as paternalistic, informative, and deliberative, based on their ages and the nature of their disease [22]. In domains like healthcare that have clear taxonomies of interaction styles, CUI designers need to determine what user attributes will influence their preferences, or simply test a range of interaction styles with a large representative sample of target users to understand which are preferred.

Follow the Inclusive Design Principles, "consider situation," "give control," "offer choice," and "prioritize content."

IV. Recommendations for Inclusive CUI Personalities

A. Know Your Users, and Be Aware of Who You Are Including and Excluding

When designing a CUI, understand your audience through user research, interviews, and contextual inquiry. Some teams document a list of those they are willing to exclude (for example, users of Internet Explorer 7 since it is well-past the sell-by date) and keep the list in mind throughout the design and development to guide decision-making. Providing a text-based chatbot along with any audio is a way to be inclusive of those with hearing considerations. For Veteran survivors of military sexual trauma, future research may reveal that some personality features may be too "soft" and make the Veteran feel they are not understood. Get to know your audience, and provide personalities that suit their needs.

B. Offer a Range of Personalities for a Range of People

Offering a selection of personalities is one avenue that some interfaces offer. For example, Siri offers a selection of voices, as well as languages from a range of countries and regions. Each has a slightly different personality, and some users select their language from a particular region because of the personality they associate with it, such as a U.S. user choosing a U.K. accented voice. Microsoft's Clippy virtual assistant evolved to offer alternative avatars with different personalities. An important rule of thumb, though, is: "No matter what you choose, avatars won't cure bad interactions. Just ask Clippy" [23]. In other words, personality choices must be targeted and not just for the sake of variety.

C. Make Sure the Bot's Personality Enhances Its Purpose

Understanding the user's purpose is key in designing suitable services. Depending on the audience, the bot may need to be formal or casual; humor and even conflict may be used to provoke critical thinking, such as with "Bots of Conviction" [24]. In this case study, the bot asked the user if they would bury their loved ones beneath their bed. Users generally were surprised, which allowed the bot to reveal that in some ancient cultures they did this to keep their loved ones close. The bot's personality is confidently of another culture, eliciting discourse and reflection. In helping Veterans ready for life after active duty, a bot may need to be both compassionate and challenging, as it reminds users to go to training, submit forms, and attend to other tasks. In contrast, the Amazon customer service bot is friendly, upbeat, and apologetic as it addresses customer service issues. If it took a humorous approach, that would likely offend some customers already upset about a product issue.

D. Understand Users' Tendency to Anthropomorphize

Some of the challenges mentioned in this paper depend on whether users are likely to view the CUI as a social partner or a transactional means to an end. Factors affecting a user's tendency to anthropomorphize technology include age, gender, computer anxiety, and need for interaction [25]. Users likely to anthropomorphize CUIs can be expected to appreciate social conventions such as appropriately polite and empathic language.

E. Involve Diverse People in the Development Process

Because people from different cultures and backgrounds have different expectations for conversations, the surest way to accommodate a range of people is to involve them in product design and testing. Politeness conventions, for example, differ between individualistic and collectivistic cultures [26].

Radar Pace, a virtual coach developed by Oakley and Intel, adjusts its personality by locale. In Spanish-speaking locales, the coach's voice is female and gives responses that are "firm and authoritative", while in French-speaking locales it has a male voice and is "encouraging and cooperative" [8]. Cross-cultural feedback was necessary to create an application that could be taken seriously as a coach by a variety of users.

V. Conclusion

In this paper we presented challenges that should be systematically addressed in research to move toward inclusive CUI personalities, as well as some overarching recommendations or themes to guide development. Studies exploring the impacts of empathy and politeness in conversational robots and software need to be integrated with studies of how diverse users respond to manifestations of social cues in technology. CUI development teams should take full advantage of user-centered research and design tools, such as personas, user stories, and structured interviews [27], to understand and anticipate the range of needs, attitudes, and expectations of their users.

Most CUIs take an initially neutral personality and when an interaction becomes more complex transfers the conversation to a human being. Until a CUI can precisely adapt to a user's preferences, that approach remains among the most inclusive. However, ambitious research, synthesis, and tool development can bring us closer to CUIs that serve all potential users at all times of day.

Approved for public release. Distribution unlimited 20-03275-5.

References

[1] K. Lister, T. Coughlan, F. Iniesto, N. Freear, and P. Devine, "Accessible conversational user interfaces: considerations for design," in *Proceedings of the 17th International Web for All Conference*, New York, NY, USA, Apr. 2020, pp. 1–11. doi: 10.1145/3371300.3383343.

[2] J. Stanley, R. ten Brink, A. Valiton, T. Bostic, and B. Scollan, "Chatbot Accessibility Guidance: A review and way forward," in *Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London*, vol. 3, X.-S. Yang, S. Sherratt, N. Dey, and A. Joshi, Eds. Brunel University, London: Springer Nature Singapore Pte Ltd., 2021, pp. 919–942. Accessed: Apr. 14, 2021. [Online]. Available: https://www.springer.com/gp/book/9789811617805

[3] J. Davis, "Government CIOs Prioritize Chatbots in Pandemic," *InformationWeek*, Jan. 06, 2021. Accessed: Aug. 12, 2021. [Online]. Available: https://www.informationweek.com/leadership/government-cios-prioritize-chatbots-in-pandemic

[4] L. Avanessian, "Bank of America: Designing Erica's chatbot personality," 2018. https://loricavanessian.com/project/erica (accessed Mar. 17, 2021).

[5] Á. Callejas-Rodríguez, E. Villatoro-Tello, I. Meza, and G. Ramírez-de-la-Rosa, "From Dialogue Corpora to Dialogue Systems: Generating a Chatbot with Teenager Personality for Preventing Cyber-Pedophilia," in *Text, Speech, and Dialogue*, Cham, 2016, pp. 531–539. doi: 10.1007/978-3-319-45510-5_61.

[6] C. Edwards, A. Edwards, B. Stoll, X. Lin, and N. Massey, "Evaluations of an artificial intelligence instructor's voice: Social Identity Theory in human-robot interactions," *Comput. Hum. Behav.*, vol. 90, pp. 357–362, Jan. 2019, doi: 10.1016/j.chb.2018.08.027.

[7] S. Katz, "The ultimate guide to chatbot personality," *Medium*, May 18, 2020. https://chatbotsmagazine.com/the-ultimate-guide-to-chatbot-personality-b9665ab5e99d (accessed Jan. 04, 2021).

[8] A. Danielescu and G. Christian, "A Bot is Not a Polyglot: Designing Personalities for Multi-Lingual Conversational Agents," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC Canada, Apr. 2018, pp. 1–9. doi: 10.1145/3170427.3174366.

[9] P. Persson, J. Laaksolahti, and P. Lonnqvist, "Understanding socially intelligent agents - a multilayered phenomenon," *IEEE Trans. Syst. Man Cybern. - Part Syst. Hum.*, vol. 31, no. 5, pp. 349–360, Sep. 2001, doi: 10.1109/3468.952710.

[10] Google, "Create a persona - Conversation design process - Conversation design," *Designing Actions on Google*, 2021. https://designguidelines.withgoogle.com/conversation/conversation-design-process/create-a-persona.html (accessed Oct. 21, 2020).

[11] H. Swan, I. Pouncey, H. Pickering, and L. Watson, "Inclusive Design Principles," *Inclusive Design Principles*, 2021. https://inclusivedesignprinciples.org/ (accessed Aug. 09, 2021).

[12] B. Jacquet, A. Hullin, J. Baratgin, and F. Jamet, "The Impact of the Gricean Maxims of Quality, Quantity and Manner in Chatbots," in *2019 International Conference on Information and Digital Technologies (IDT)*, Jun. 2019, pp. 180–189. doi: 10.1109/DT.2019.8813473.

[13] T. Bickmore, A. Gruber, and R. Picard, "Establishing the computer–patient working alliance in automated health behavior change interventions," *Patient Educ. Couns.*, vol. 59, no. 1, Art. no. 1, Oct. 2005, doi: 10.1016/j.pec.2004.09.008.

[14] M. de Gennaro, E. G. Krumhuber, and G. Lucas, "Effectiveness of an Empathic Chatbot in Combating Adverse Effects of Social Exclusion on Mood," *Front. Psychol.*, vol. 10, p. 3061, Jan. 2020, doi: 10.3389/fpsyg.2019.03061.

[15] H. Cramer, J. Goddijn, B. Wielinga, and V. Evers, "Effects of (in)accurate empathy and situational valence on attitudes towards robots," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Osaka, Japan, Mar. 2010, pp. 141–142. doi: 10.1109/HRI.2010.5453224.

[16] F. D. Rosis, C. Pelachaud, and I. Poggi, "Transcultural believability in embodied agents: a matter of consistent adaptation," in *Agent Culture: Human Agent Interaction in a Multicultural World*, S. Payr and R. Trappl, Eds. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2004, pp. 75–106.

[17] J. Maciej, M. Nitsch, and M. Vollrath, "Conversing while driving: The importance of visual information for conversation modulation," *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 14, no. 6, pp. 512–524, Nov. 2011, doi: 10.1016/j.trf.2011.05.001.

[18] F. A. Drews, M. Pasupathi, and D. L. Strayer, “Passenger and cell phone conversations in simulated driving,” *J. Exp. Psychol. Appl.*, vol. 14, no. 4, pp. 392–400, 2008, doi: 10.1037/a0013119.

[19] J. Guo, J. Guo, C. Yang, Y. Wu, and L. Sun, “Shing: A Conversational Agent to Alert Customers of Suspected Online-payment Fraud with Empathetical Communication Skills,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan, May 2021, pp. 1–11. doi: 10.1145/3411764.3445129.

[20] P. Brown and S. C. Levinson, *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.

[21] J. R. Wilson, N. Y. Lee, A. Saechao, and M. Scheutz, “Autonomy and Dignity: Principles in Designing Effective Social Robots to Assist in the Care of Older Adults,” 2016.

[22] C. Gross *et al.*, “Personalization of Conversational Agent-Patient Interaction Styles for Chronic Disease Management: Results from two studies with COPD patients (Preprint),” Journal of Medical Internet Research, preprint, Dec. 2020. doi: 10.2196/preprints.26643.

[23] C. Platz, *Design Beyond Devices: Creating Multimodal, Cross-Device Experiences*, 1st edition. New York: Rosenfeld Media, 2020.

[24] M. Roussou, S. Perry, A. Katifori, S. Vassos, A. Tzouganatou, and S. McKinney, “Transformation through Provocation?,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–13. Accessed: Aug. 12, 2021. [Online]. Available: https://doi.org/10.1145/3290605.3300857

[25] M. Blut, C. Wang, N. V. Wünderlich, and C. Brock, “Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI,” *J. Acad. Mark. Sci.*, vol. 49, no. 4, pp. 632–658, Jul. 2021, doi: 10.1007/s11747-020-00762-y.

[26] P. Longcope, “The universality of face in Brown and Levinson’s politeness theory: A Japanese perspective,” *Univ. Pa. Work. Pap. Educ. Linguist.*, vol. 11, no. 1, pp. 69–79, 1995.

[27] E. Adiseshiah, “Personas, scenarios, user stories and storyboards: what’s the difference?,” *JUSTINMIND*, Jul. 28, 2017. https://www.justinmind.com/blog/user-personas-scenarios-user-stories-and-storyboards-whats-the-difference/ (accessed Aug. 12, 2021).

ISBN: 978-1-61208-896-9

Supporting Observability through Social Cues

Natalie Friedman
Cornell Tech
New York City, NY, USA
e-mail: nvf4@cornell.edu

Patricia L. McDermott, Jeff Stanley
The MITRE Corporation
McLean, VA, USA
{pmcdermott, jstanley}@mitre.org

***Abstract*—For improved acceptance of robots in social spaces, it is important to have a strong mental model of what the robot can do, what the robot is currently doing, or what the robot is about to do. How do social cues help people understand what is going on 'under the hood'? Imagine this: a machine perks up if someone enters the room. This lets you know it is socially aware, awake, and ready to interact. Drawing from a pre-existing taxonomy of social cues for conversational agents, we reviewed 40 papers with instances of robot or software agent personality traits influencing observability. This survey led us to elaborate on six particular cues, clarify their relationship to observability and provide examples, with the intent to advance discussion and encourage research on the relationship between social cues and observability.**

Keywords-observability; social cues; personality; human-robot interaction; human-computer interaction; trust; predictability

I. Introduction

How do social cues help people understand what is going on “under the hood” of robots? Many fields including Human-Computer Interaction (HCI), dance, and animation confirm that expressivity can reveal functionality [1][2]. However, there is limited research presenting design guidelines that link individual social cues with a robot’s internal state, i.e., what is happening “under the hood.” Observability is defined as appropriate transparency into what an automated partner can do and is doing relative to task progress [3]. Amy LaViers, a Human-Robot Interaction (HRI) researcher who studies movement, describes that in the Laban movement community there is an "indivorcibility of function and expression" and that "more expressive robots are more functional robots" [1]. HRI researchers, studying movement and expressivity, divide functional task motions (e.g., “grabbing the doorknob”) from expressive motions (e.g., “looking around the door handle and scratching its head”). While they attempt to separate the two, they write “we do not subscribe to the idea that these are completely separate concepts” [2]. Similarly, we argue that there is an indivorcibility of observability and expressivity. In this paper, we contribute an assessment of how an agent's social cues [4] can help improve robot observability. Discussion of social cues for robots is nothing new, and we aim to advance the discussion by clarifying ways to view social cues through the lens of observability.

A. *Observability*

Observability and transparency have been used synonymously [3]. A transparent system communicates feedback about the system reliability and situational factors; it can establish appropriate trust and improve team performance [5]. Through transparency, teammates generate a shared understanding of the task and calibrate trust based on the team members' capabilities [6]. It is especially important for humans to develop calibrated trust because inappropriate trust often leads to misuse or disuse of automation [5]. Transparency also gives the human team members situational awareness of the task, robot, and environment [7][8].

The benefits of observability come with challenges. Communication among human-machine teams (or teams of humans and automated agents) is restricted due to the limited ways that automated agents give and receive communication. For example, humans use non-verbal communication, and can adapt communication to situations outside the nominal task [8]. Some of the non-verbal ways that humans communicate are via social cues [4], which have been implemented on robots, as we will discuss.

B. *Personality and Social Cues*

Human personality is defined as “characteristic sets of behaviors, cognitions, and emotional patterns that evolve from biological and environmental factors” [9]. While robots are not influenced by biological factors, they have software and environmental sensors. In designing robots, personality can be operationalized by social behaviors and cues that are designed to be appropriate for the robot’s role and environmental factors.

A social cue is defined as “a cue that triggers a social reaction towards the emitter of the cue” [4][10][11][12]. In human-robot interaction, a social cue has been defined as “features that `act as channels of social information’” [4][13].

Which social cues should robots use? In human-computer interaction, performative behaviors [14] have been used to display a machine's reaction to its actions [2], system state, or to demonstrate its social standing [15]. For example, researchers in collaboration with animators designed a slouching behavior for a robot when the robot could not open the door [2]. This performative reaction demonstrates to the bystanders that it is aware and embarrassed of its failure. Slouching after failure is an example of how social cues can improve observability, and ultimately the human understanding of a robot.

Copyright (c) IARIA, 2021. ISBN: 978-1-61208-896-9

Performative behaviors are also used to change the social dynamic within a group. For example, Jung et al. used repair interventions ("hey, not cool") to achieve an awareness of conflict within a human group dynamic to prevent personal attacks [15]. Similarly, Sebo et al. found that designing a robot to be vulnerable (“Sorry guys, I made the mistake this round. I know it may be hard to believe, but robots make mistakes too”) can make others more likely to share their failure to the group and laugh together [16]. In this case, this performative behavior gives the user information about its knowledge of failure and displays personality. Here, we see performative social cues can help people understand that the robot is aware of its mistake by showing shame and vulnerability and aware of the social dynamic of the situation by reacting to it.

In this paper, we assess specific social cues and how they can make a robot’s internal state more observable. The paper has five sections. Section II briefly describes the methodology and sources that informed our analysis. Section III introduces six social cues and illustrates how they can contribute to observability, as summarized in Table I at the end of that section. In Section IV we highlight some overall themes and open questions. Section V concludes the paper.

II. Methods

We intend to develop a dataset of social cues that are useful to designers for improving observability of robots and automated agents. Although we primarily refer to robots in this paper, the examples and findings are also applicable to embodied agents or digital assistants. We draw from the paper “A Taxonomy of Social Cues for Conversational Agents” by Feine et al. [4], which systematically identifies 48 social cues from the literature. In parallel, we have searched for these social cues in relationship to observability. We used a combination of ACM Digital Library and Google Scholar searches to learn about the effects of personality on observability. Searches for articles included keywords like, “observability”, “mental model”, “status”, “common ground”, “predictability”, and “machine personality.” These search terms originated from literature about observability [3].

Once we identified research that investigated how these social cues could support observability, we narrowed down the list of social cues using the following criteria:

- Cue should display system state dynamically (as opposed to static design choices like gender and name).
- Cue can be applied to both physical and virtual agents.

These criteria enabled us to focus on cues that were applicable to designers of a wide variety of systems, from assistive robots to digital agents in planning applications. After reviewing 40 sources, we selected cues with sources illustrating their relationship to observability.

We included cues from three different modality types: posture, voice, and dialogue. See Table 1 for the social cues, the definition of the cues from Feine et al. [4] and their relationship to observability.

III. Analysis of social cues

A. *Head Movement*

Head movement refers to a gesture or position of the head and can include nodding, shaking, tilting, looking towards or looking away from a human. Common straightforward cues include nodding to indicate agreement and shaking head to indicate disagreement [17]. Pairing a nod with an affirmative statement can make an agent’s behavior seem consistent and reassure a human observer. Conversely, if the head movement is inconsistent with verbal statements, such as nodding while denying a request, the human may perceive the robot unreliable.

Poggi [18] presents more nuanced cues that can indicate a speaker’s beliefs and goals, such as: nodding head to show certainty, looking up to show careful thinking, looking down obliquely to indicate trying to remember. A human may have an expectation that an agent is able to respond instantaneously. If an agent provides a cue that it is thinking, either by looking up or looking obliquely down, the human can understand that the agent needs time to construct a response. The human can adjust his or her expectations and be less frustrated by a pause in the dialogue. Humans are more likely to be patient if they know that their problem or question is being carefully considered.

Head movement may not occur in isolation, as the only non-verbal cue. Often head movement cues are combined with facial expression and eye movement to communicate beliefs and intentions. An example is a robot that combines a fixed stare with raising the inner parts of the eyebrows and a bent head to indicate “I implore you.” [18].

Takeaway: Head movement can be used for more than just agreement (nod), disagreement (shake). Head movements can convey that an agent is thinking (looking upwards), is certain (nodding) or trying to remember (looking obliquely downwards).

B. *Facial Expression*

A facial expression, a form of nonverbal communication, is a movement of the face to communicate internal state like surprise, sadness, happiness, anger, etc. Internal state is demonstrated through specific facial expressions including a nod, smile, shake, frown, tension of the lips, tilt, or raise of the eyebrows [19].

For example, Cassell designs an avatar to be expressive during turn taking; when the agent is letting the person speak, it raises its eyebrows and relaxes its hands [20].

Bevacqua et al. studied the efficacy of backchanneling (a non-interrupting acoustic or visual signal demonstrated by the listener during a speaker’s turn) in creating meaning through different facial expressions and acoustic cues [17]. They tried to learn which cues together could communicate interest and understanding. They found that ‘interest’ was conveyed by the following combination of cues: a smile and a verbal “okay,” a nod and a verbal “okay,” and a nod and a verbal “ooh.” Likewise, ‘understanding’ was conveyed by nodding as well as the following combinations: raising eyebrows and a verbal “ooh,” nodding and a verbal “ooh,” nodding and a verbal “really,” nodding and a verbal “yeah.”

They note that only cues together could compose meaning. In general, it is not possible to select one cue from their findings and expect it to communicate interest or understanding on its own [17].

A risk of using a facial expression could be using the wrong expression for the environment or role. For example, if a robot is showing a happy facial expression in a serious or sad environment, like a funeral, the robot might be deemed as socially inappropriate. Conversely, we speculate that a more serious facial expression in a librarian role, for example, might help make the robot belong in the role and context.

Takeaway: Facial expressions, like raising eyebrows (in conjunction with relaxing hands) can show that an agent is letting a person speak. Smiling and saying “okay,” can show interest.

C. Voice Tempo

Voice tempo refers to speech rate and pacing. It can be measured as seconds per syllable, for example, or as the length of pauses between words when spoken by a synthesized voice.

Speech rate, as well as pauses in speech, can communicate confidence level, which is important for observability of machines. Research on human speech perception has found that a moderately fast speech rate (ideally similar to or slightly faster than the listener’s speech rate) conveys competence, while slower pacing and pauses longer than 5 seconds indicate that the speaker is not sure of the content [21]. Therefore, voice tempo could be a valuable social cue for machines that need to communicate confidence levels and manage the human’s trust in content because trust can be calibrated appropriately by altering the tempo.

Abnormally fast speech can be associated with nervousness or urgency. Jang has shown that the speed of computerized speech does in fact convey urgency of a situation in a predictable way [22]. On the other hand, a slower speech rate might help humans to remain calm during an emergency. For instance, when a semi-automated car notices an approaching obstacle, a fast speech rate may be needed to communicate urgency, followed by a slower speech rate to orient and support the human.

Speech rates that are too fast or too slow can contribute to comprehension problems [23]. When manipulating voice tempo, in non-extreme situations, synthetic voices should tend away from extremes and toward everyday average human speech rates to be appropriate interaction partners.

Takeaway: Voice tempo can speed up to communicate urgency, but in non-emergency situations should be similar to a human’s voice tempo for easy understandability. Slowness or pauses can be used to communicate lower confidence in information. This is valuable because correctly communicating uncertainty can improve the human’s overall calibrated trust in the machine’s judgments.

D. Pitch Range

Pitch range refers to how high and low a synthesized voice varies from its average pitch frequency.

One study found evidence that for both Italians and Americans, having a wider pitch range made communications for people seem more exciting, interesting, and credible [24]. Similarly, a study of consumers’ impressions indicated that a wider pitch range contributes to more exciting and memorable commercials [25]. These findings suggest that an exaggerated pitch variation should be used to present important high-confidence information.

While pitch range increases the perceived competence of a message, research on vocal styles has found an inverse relationship between competence and benevolence; and there is some evidence that exaggerated pitch variation decreases perceived benevolence of a speaker, along with respect and fairness [26]. Therefore, a wider-than-normal pitch range should only be used when needed to draw the user’s attention; and should perhaps be avoided especially for systems that need to be seen as fair, such as recidivism predictors [27].

Cowell and Stanney [28] tested the effects of non-verbal cues expressed by a conversational digital assistant for helping to sort photos into albums. Characters with an appropriate pitch variation, a moderately fast speech rate (50-70 words per minute), facial expressions, and eye gaze were rated as significantly more trustworthy and credible than characters without any intelligent management of these features.

Takeaway: Use exaggerated pitch range only to draw attention to important or high-confidence information. Otherwise, use humanlike pitch range as appropriate to convey the trustworthiness and competence of the machine.

E. Greetings and Farewells

Greetings and farewells are expressions, or “ritual behaviors” [29], marking an agent entering or leaving an interaction. These social cues can improve trust [30] and have been found to help users to perceive an agent as more reliable, competent, and knowledgeable. Examples include saying, “Welcome”, “Nice to meet you”, “Hello”, “See you later.”

A greeting demonstrates that the target has entered the agent’s realm of activation and is being sensed or tracked. We speculate that the timing and manner of an agent greeting a person could show awareness of when a person is bored or busy. A well-timed farewell is an effective way for a person to know that the robot understands that the interaction is over.

Risks of using this social cue could be saying a farewell too early in an interaction, which could be perceived as rude or socially unaware. This is a challenge, because of the lack of social intuition that robots have. It is often difficult for a robot to know when to interrupt, which has been explored by Semmens et al., in which researchers in a car periodically asked, “Is now a good time?” They found that a system can access automotive data for knowing when to ask if it is a good time [31]. This goes to show that while robots have trouble sensing social situations, there is an opportunity to

ISBN: 978-1-61208-896-9

TABLE I. SOCIAL CUES AND RELATIONSHIP TO OBSERVABILITY

Social Cue	Feine's Definition [4]	Relationship to Observability	Source
Head movement [Posture]	The agent moves its head. (I.e., nodding and turning)	In addition to more obvious indicators of agreement and disagreement, head movement can be used to indicate nuanced beliefs and goals such as confidence, thinking, and remembering	[18]
Facial expression [Posture]	The agent expresses a gesture by executing one or more motions with his facial muscles (i.e., smile or eyebrow raise)	Facial expressions, like raising eyebrows (in conjunction with relaxing hands) can show that an agent is letting a person speak.	[30]
Voice Tempo [Voice]	The pace of the agent's voice.	The speed of computerized speech conveys urgency of a situation in a predictable and systematic way, and speech pacing conveys confidence.	[21][22]
Pitch Range [Voice]	The degree of variation from the agent's average pitch. (I.e., monotone, animate voice)	Exaggerated pitch range can draw attention to important or high-confidence information. Humanlike pitch range should be appropriate to the trustworthiness and competence of the machine.	[25]
Greetings and farewells [Dialogue]	The agent expresses a word of welcome or marks someone's departure.	Small talk, which include greetings and farewells, can improve perception of an agent's good will and credibility.	[30]
Ask to start/ pursue dialogue [Dialogue]	The agent requests the user's permission to start, continue, or end the conversation	Asking to start or pursue a dialogue communicates that the human is in charge and is in support of the Human-Machine Teaming theme of Directability	[3][32]

leverage other sensors, like proximity sensors or lidar, for detecting things like if someone has left the space.

Takeaway: Greetings and farewells signal awareness of the user and that a new interaction is beginning/ending.

F. Ask to Start / Pursue Dialogue

Ask to start/pursue dialogue refers to behavior in which the agent seeks permission to interact with a human partner. This could take the form of initiating a conversation, as in, "My name is Indira and I can help plan tourist activities. Would you like me to look for available excursions?" It could also involve the seeking of approval to continue an interaction, such as "Would you like me to keep searching?" This social cue is tightly related to the Human-Machine Teaming theme of Directability [3], by which humans are easily able to direct and redirect an automated partner's resources, activities, and priorities. Asking to start or pursue dialogue signals that the human is in control, which is important as the human should not be removed from the command role [32]. When humans perceive a lack of control, they can become frustrated. Letting the human know that they can discontinue the agent's help shows that the agent is directable, potentially decreasing frustration and increasing user adoption.

Takeaway: When an agent asks to start or continue dialogue, it signals that the human is in control of the interaction. Provide multiple choice points in which the human can decide whether to continue dialoguing with the agent to minimize user frustration.

IV. DISCUSSION

In assessing the social cues and their relationship to observability, we found that one common risk of using social cues is setting wrong expectations. For example, if there are moving eyes on a robot, it might be perceived that a robot can see. We speculate that if it cannot see, but has eyes, then that could lead to a mistrust of an agent, which could be worse than not including the social cue at all. The use of social cues can increase the "human-ness" of a robot or software agent, but if expectations are violated it would be better to not include the social cue. Another danger of making a robot human-like is that it can enter the "uncanny valley" [33], in which robots that look a lot like humans, but not quite human, are perceived as creepy and cause revulsion. Examples include the characters in the Polar Express movie and the version of Sophia from Hanson Robotics that debuted at the 2016 South by Southwest (SXSW) conference. We posit that the social cues described in this article would not by themselves enter the uncanny valley. Rather, their use on a highly humanoid (but not convincingly human) robot platform could be disconcerting.

We also noticed that in most research about social cues, one social cue alone often does not express one piece of information; but instead, it is multiple cues in a row, or different modalities demonstrated in parallel, which communicate the desired observable behavior. For instance, one facial expression will not convey interest, but instead, a facial expression in parallel with other gestures will show interest.

Social cues may be sensitive to the cultural context in which they are used. Some cues may be universal in humans, such as the combination of lowered eyebrows, lips firmly pressed, and bulging eyes to convey anger. Other cues could be interpreted differently in separate cultures. What is perceived as friendly in the United States could be perceived as intrusive in other countries. Interpretability of social cues should be tested across cultures to ensure the gesture, movement, or vocal characteristic conveys the intended information. This could lead to the identification robot social cues that are universal.

We recommend that future work focus on the combination of social cues to convey information. In human-to-human encounters, social cues typically occur simultaneously and across modalities and are a natural part of communication. Mapping combinations of social cues to observability would likewise enhance the richness of robot-to-human communication.

V. Conclusion

In this paper, we began with discussing the importance of people understanding what is going on "under the hood" of machines and the opportunities of social cues to help uncover system status for a user. Next, we shared our process of finding social cues that have the potential to improve observability in both physical and virtual agent design. We focused on cues that display system state dynamically and had supporting literature. Lastly, we did a deep dive into six social cues that met our criteria by including definitions of the cue, examples of the cue, the relationship to observability, and the associated risks of using the cue inappropriately. See table 1 for the cues organized. We found that for maximum success at communicating observability, the cues should be used in parallel with other cues.

Acknowledgment

Thank you to MITRE for funding this work; thank you to Kit Mandeville, Kelly Horinek, Alexandra Valiton, Dr. Ozgur Eris, and Dr. Beth Elson for their discussions about these topics.

Approved for Public Release; Distribution Unlimited. Public Release Case Number 21-1848.

References

[1] I. Pakrasi, N. Chakraborty, and A. LaViers, "A design methodology for abstracting character archetypes onto robotic systems," in *Proceedings of the 5th International Conference on Movement and Computing*, 2018, pp. 1–8.

[2] L. Takayama, D. Dooley, and W. Ju, "Expressing thought: improving robot readability with animation principles," in *Proceedings of the 6th international conference on Human-robot interaction*, 2011, pp. 69–76.

[3] P. McDermott et al., "Human-machine teaming systems engineering guide," The MITRE Corporation, Bedford, MA, USA, 2018. [Online]. Available: https://www.mitre.org/publications/technical-papers/human-machine-teaming-systems-engineering-guide [Accessed Sept. 13, 2021].

[4] J. Feine, U. Gnewuch, S. Morana, and A. Maedche, "A Taxonomy of Social Cues for Conversational Agents," *International Journal of Human-Computer Studies*, vol. 132, pp. 138–161, Dec. 2019, doi: 10.1016/j.ijhcs.2019.07.009.

[5] K. A. Hoff and M. Bashir, "Trust in automation: integrating empirical evidence on factors that influence trust," *Hum Factors*, vol. 57, no. 3, pp. 407–434, May 2015, doi: 10.1177/0018720814547570.

[6] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human factors*, vol. 39, no. 2, pp. 230–253, 1997.

[7] J. B. Lyons, "Being transparent about transparency: A model for human-robot interaction," presented at the AAAI Spring Symposium: Trust and Autonomous Systems, 2013.

[8] J. Joe, J. O'Hara, H. Medema, and J. Oxstrand, "Identifying requirements for effective human-automation teamwork," presented at PSAM, 2014.

[9] P. J. Corr and G. Matthews, Eds., *The Cambridge Handbook of Personality Psychology*, 2nd ed. Cambridge: Cambridge University Press, 2020. doi: 10.1017/9781108264822.

[10] A. Vinciarelli and G. Mohammadi, "A Survey of Personality Computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, Jul. 2014, doi: 10.1109/TAFFC.2014.2330816.

[11] B. J. Fogg, "Persuasive technology: using computers to change what we think and do," *Ubiquity*, vol. 2002, no. December, p. 2, 2002.

[12] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *Journal of Social Issues*, vol. 56, pp. 81–103, 2000.

[13] E. J. Lobato, S. F. Warta, T. J. Wiltshire, and S. M. Fiore, "Varying social cue constellations results in different attributed social signals in a simulated surveillance task," presented at FLAIRS, 2015.

[14] L. Holzman, "Lev Vygotsky and the new performative psychology: Implications for business and organizations," *The social construction of organization*, pp. 254–268, 2006.

[15] M. F. Jung, N. Martelaro, and P. J. Hinds, "Using robots to moderate team conflict: the case of repairing violations," in *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 2015, pp. 229–236.

[16] S. S. Sebo, M. Traeger, M. F. Jung, and B. Scassellati, "The Ripple Effects of Vulnerability: The Effects of a Robot's Vulnerable Behavior on Trust in Human-Robot Teams," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, Feb. 2018, pp. 178–186. doi: 10.1145/3171221.3171275.

[17] E. Bevacqua, S. Pammi, S. J. Hyniewska, M. Schröder, and C. Pelachaud, "Multimodal backchannels for embodied conversational agents," in *International Conference on Intelligent Virtual Agents*, 2010, pp. 194–200.

[18] I. Poggi, "Mind markers," *The Semantics and Pragmatics of Everyday Gestures. Berlin Verlag Arno Spitz*, 2001.

[19] D. Heylen, E. Bevacqua, M. Tellier, and C. Pelachaud, "Searching for prototypical facial feedback signals," in *International Workshop on Intelligent Virtual Agents*, 2007, pp. 147–153.

[20] J. Cassell, "Embodied conversational interface agents," *Communications of the ACM*, vol. 43, no. 4, pp. 70–78, 2000.

[21] R. L. Street and R. M. Brady, "Speech rate acceptance ranges as a function of evaluative domain, listener speech rate, and communication context," *Communication Monographs*, vol. 49, no. 4, pp. 290–308, Dec. 1982, doi: 10.1080/03637758209376091.

[22] P.-S. Jang, "Designing acoustic and non-acoustic parameters of synthesized speech warnings to control perceived urgency," *International Journal of Industrial Ergonomics*, vol. 37, no. 3, pp. 213–223, Mar. 2007, doi: 10.1016/j.ergon.2006.10.018.

[23] E. Rodero, "Do Your Ads Talk Too Fast To Your Audio Audience?: How Speech Rates of Audio Commercials Influence Cognitive and Physiological Outcomes," *Journal of Advertising Research*, vol. 60, no. 3, pp. 337–349, Sep. 2020, doi: 10.2501/JAR-2019-038.

[24] M. Urbani, "The Pitch Range of Italians and Americans. A Comparative Study," University of Padua, 2013. [Online].

Available: http://paduaresearch.cab.unipd.it/5976/ [Accessed Sept. 13, 2021].

[25] E. Rodero, R. F. Potter, and P. Prieto, “Pitch Range Variations Improve Cognitive Processing of Audio Messages,” *Human Communication Research*, vol. 43, no. 3, pp. 397–413, Jul. 2017, doi: 10.1111/hcre.12109.

[26] P. Rockwell, “The effects of vocal variation on listener recall,” *J Psycholinguist Res*, vol. 25, no. 3, pp. 431–441, May 1996, doi: 10.1007/BF01727001.

[27] J. Dressel and H. Farid, “The accuracy, fairness, and limits of predicting recidivism,” *Science Advances*, vol. 4, no. 1, p. eaao5580, Jan. 2018, doi: 10.1126/sciadv.aao5580.

[28] A. J. Cowell and K. M. Stanney, “Manipulation of non-verbal interaction style and demographic embodiment to increase anthropomorphic computer character credibility,” *International Journal of Human-Computer Studies*, vol. 62, no. 2, Art. no. 2, Feb. 2005, doi: 10.1016/j.ijhcs.2004.11.008.

[29] J. Cassell and H. Vilhjálmsson, “Fully embodied conversational avatars: Making communicative behaviors autonomous,” *Autonomous agents and multi-agent systems*, vol. 2, no. 1, pp. 45–64, 1999.

[30] J. Cassell and T. Bickmore, “External manifestations of trustworthiness in the interface,” *Communications of the ACM*, vol. 43, no. 12, pp. 50–56, 2000.

[31] R. Semmens, N. Martelaro, P. Kaveti, S. Stent, and W. Ju, “Is now a good time? An empirical study of vehicle-driver communication timing,” in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1–12.

[32] C. E. Billings, *Aviation automation: The search for a human-centered approach*. CRC Press, 2018.

[33] M. Mori, K. F. MacDorman, and N. Kageki, “The uncanny valley [from the field],” *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012.

Intelligibility of Responsive Webpages: User Perspective

Vanessa Hönig
Faculty of Computer Science
Nuremberg Institute of Technology
Nuremberg, Germany
email: vanessa_hoenig@gmx.de

Alexander Kröner
Faculty of Computer Science
Nuremberg Institute of Technology
Nuremberg, Germany
email: alexander.kroener@th-nuernberg.de

***Abstract*— Nowadays, websites are considered to be responsive as a matter of course. Relatable literature focuses on the technical implementation of the associated changes and rarely addresses the potential negative usability implications. Therefore, this article addresses user expectations concerning the presentation of a website on different devices. A series of experiments focusing on selected visual and functional aspects of a website requiring adaptation is presented for this purpose and their findings are discussed.**

Keywords-RWD; responsive; webdesign; mental model; user tests; usability; user experience; first-click-test.

I. Introduction

Responsive Web Design (RWD) can enhance user experience significantly – but it can also be detrimental. If a user knows the presentation of a website from a particular presentation environment, then modifying that website to a different presentation environment may cause, for instance, a loss of orientation [1][2]. A standardized realization of responsive behavior could counter such effects to some extent. However, responsive behavior may strongly depend on application context and device-specific constraints, and thus be highly diverse. This might explain why standards such as DIN EN ISO1 9241-151 recommend to consider contextual aspects for user interface design, but provide few hints regarding the realization of responsive behavior. Platform-specific user interface guidelines (e.g., [3]) fill this gap to some extent, but cannot necessarily be applied to manufacturer-independent scenarios.

The rest of this paper is organized as follows. Section II addresses related work. Section III deals with the performed usability tests in context of this article. The results and goals of these tests are discussed in Section IV.

II. Related Work

Human expectations (or behavioral patterns) relate to so-called "mental models", a central concept of human-computer interaction [4]. Such models assume that humans compile personal experiences into abstract models, which are then used to predict real-world behavior. Based on the stored knowledge a quick interpretation and reaction to external and internal events is possible [5]. Therefore, a presentation and interaction model of a website, which is close to the user's mental model of that site, can support understanding and operation [6][7] and may contribute to an anticipatory design process [33].

Acquiring a mental model can be challenging. People do not have access to their entire belief structure. Consequently, they have no direct conscious access to the mental model they have constructed from previous experiences. Thus, an interview may result in an incomplete or wrong model [8]. Therefore, hints concerning users' mental model of a responsive web page's behavior must be acquired in a different way. To some extent, this is addressed by studies investigating the usability of responsive webpages (e.g., a webpage for a conference [31] or the tourism domain [32]). In order to continue such efforts for a wider application focus, the work presented in the following combines research concerning current trends in responsive behavior with experiments in a given application (a web page concerned with job offers) and results from cognitive science.

III. Experiments

The following experiments share the primary goal of improving the understanding of today's users' mental model of responsive web pages. Derived goals include insights to the relevance of device-specific features for the mental model and people's understanding of responsive behavior.

TABLE I: AGE AND GENDER DISTRIBUTION OF THE USABILITY TEST PARTICIPANTS

	Survey	Preference-Test	First-Click-Test
Quantity	70	20	50
18 to 24	17%	20%	31%
25 to 34	43%	20%	51%
35 to 44	6%	20%	0%
45 to 59	23%	20%	10%
60 and older	11%	20%	8%
Male	43%	60%	41%
Female	57%	40%	59%

A. Preparation

A preparatory study targeted user expectation concerning responsive behavior, which are likely due to the widespread adoption of responsive design. Based on the 30 most visited websites in Germany (category: online shops and news sites) [9][10][12], an analysis on the visualization of websites on different devices was conducted. Its focus was on position and presentation of the main menu. In summary, this analysis indicates a wide adoption of changeable menus. In desktop format, 80% of the menus are positioned in the upper area,

ISBN: 978-1-61208-896-9

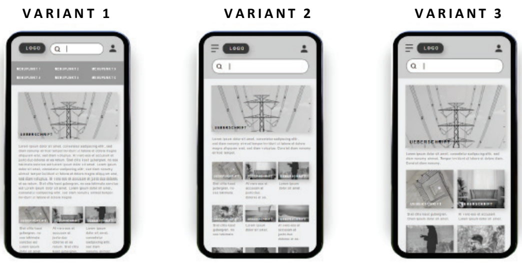
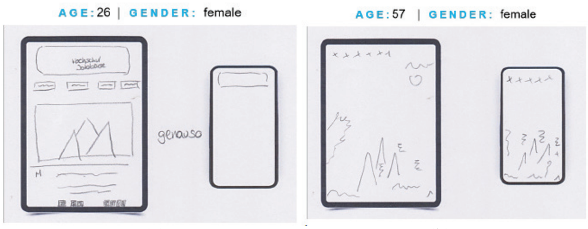


Figure 1. Three different mobile variant options.



(a) (b)

Figure 2. Mobile version sketch of a young (a) and an older participant (b).

followed by positioning on the left with 40%. The smaller the device, the more this distribution changes. On smartphones, for example, 63% of the menus are positioned on the left. On small displays, 90% of the menus are displayed in form of a hamburger icon. The format of the menu usually stays consistent and is not affected by how the device is held in vertical or horizontal format.

B. Survey

To get a basic understanding of users themselves in the context of RWD, a survey conducted via an online form was performed (see Table 1, column "Survey").

1) General web user capabilities

All subjects stated they are competent in the German and 83% also in the English language. The older the participants are, the more their ability to understand English terms decreases. After the age 45, the English skills drops from 96% to 58%. To ensure users understand website texts, English expressions should be avoided as the age of the target group increases.

Next, the participants were asked, which devices they have that can that enable website access. 97% said they own a smartphone, followed by a laptop with 79% a tablet with 64% and desktop computer with 43%. The survey also revealed that older subjects are more likely to own a tablet device than a laptop. They also prefer to use a combination of tablet and smartphone, while the younger generation tends to use a smartphone and laptop. The older subjects' preference could be explained with the more intuitive human-computer interaction of the tablet compared to desktop computers. Especially for people with special needs, such as a limited field of vision, this is attractive. This includes, among others, the older generation [11].

With regard to the operation of the various devices, the majority of users rated themselves as confident to very confident (e.g., 81% in operating mobile devices). Older test persons tended to give a poorer estimate of neutral to confident. This can probably be attributed to the time spent using websites. While the younger participants stated that they spent an average of two to four hours a day on websites, the older generation spent just under one hour. In addition, the fundamental experience with the digital world plays a role here (digital immigrants) [13].

2) Web user expectations

This test showed that good usability and uncomplicated navigation were particularly important website features to all participants. Furthermore, it showed that a fast-loading time is especially relevant for younger test persons. Whereas for older participants it is uncomplicated navigation.

Based on the laptop representation of a website, participants were asked to choose one of three given mobile variants (see Figure 1). Variant 2 in Figure 1 is visually most similar to the initial device, but to match the desktop look, the text was shortened. The test persons are therefore mostly unaware that a change has been made, because only variant 1 is unchanged. The majority of the participants chose variant 3, in which changes were made to both the content and the layout. It is assumed this decision is made based on experience (mental model) [14].

3) Comprehension for adaptive behavior

In general, most participants indicated an understanding of website visual customization. When it comes to changes in function, the results are different. Here, more than half of all subjects had no understanding for adjustments. Nevertheless, only 21% said that an understanding of website customization is very important to them. In contrast, 74% stated that reaching their goal for visit the site quickly, was very important to them. This suggests that the scrutability of a website customization is irrelevant to the participants if it helps them reach their destination faster.

C. Preference test

A preference test in form of an interview was realized [15]. Since a small number of respondents was expected, an even distribution of age groups was chosen for the test persons (see Table 1, column "Preference-Test").

1) Visualizing web users mental model

Based on a printed screenshot of the desktop version of the start page of Hochschuljobbörse [16], participants were asked to outline its tablet and smartphone version. Figure 2 shows sketches of two participants. These sketches reveal that younger test persons are aware of the entire website. The respondents only adapted the content to the width of the device without changing the data itself. For display on a smartphone, expectations differed. Some participants expected an identical presentation on the smartphone as well as on the tablet (see Figure 2 (a)). Other respondents changed the layout of the website and adapted the content to the compressed display size. Although all test persons received the same instruction, some participants only focused on sketching the big image. The test showed that this behavior only occurs among older test persons.

This observation may relate to human brain evolution. An impact on the executive function can be expected with increasing age. Part of this is the control of attention, which is used to perceive external stimuli. Some stimuli are noticed instinctively; others require awareness and attention [17].

Due to these findings, we assume that the triggered stimulus of the image might be too strong for older participants. As a result, they were less aware of the other web elements shown. Beyond, these observations suggest that a single mental model can hardly represent the expectations of all age groups.

2) *Expectations on content and layout*

The participants were presented the desktop version of a dummy website and had to choose one of two given mobile versions. 65% favored a customized version with shortened text, cropped images, and the menu as a hamburger icon. The older the respondent was, the stronger the preference was for the unchanged version.

Furthermore, 95% of users want access to all content, regardless of what device is used. If this is not realizable (e.g., small display) the participants preferred additional pages over reduction. An opposite result was obtained for the adaption of the layout. When switching the device position, 35% of the respondents prefer an unchanged layout. These participants called the reason for “being used to” this. Due to a changed device orientation, the website layout is not adjusted here either. In this Case, 65% preferred a change in the data layout for an improved overview.

Additionally, the participants were asked about the presentation of texts and images on different display sizes. Their responses indicated that they were generally unwilling to read texts on digital devices. For images, the relevant elements depicted should be easily recognizable. Furthermore, a good overview of the site content should also be guaranteed. Beyond, the larger the screen area the more images and text should be displayed. However, one test participant reported that he prefers reading text on large surfaces – but uses a smartphone for reading in the first place. The readiness of mobile devices might explain such contradictive behavior [18].

3) *Mouseover effects*

Mouseover effects can be used to highlight interactive elements. Most of the test persons perceive these effects as positive. Mouseovers also frequently display tooltips. This additional data cannot be activated on devices without mouse input. As mentioned before, web users want access to the same information no matter which device they use. By applying tooltips, this is not given. In summary, the use of this additional information should be applied wisely, or an alternative presentation should be considered for other input techniques.

4) *Device orientation*

Asked for their preferred device orientation of mobile devices, 90% answered they mostly use their smartphone vertically. Also among the six typical cell phone holding positions, the phone is only horizontal in one [19]. In contrast to the smartphone, the participants in this preference-test preferred to operate the tablet primarily horizontally. Explained can this by the similarity to the desktop presentation. Some Internet users already use the tablet as a laptop replacement. With the help of magnetic keyboards, the tablet can be quickly converted into a mini laptop at any time [20].

5) *Expectations on menu*

According to the test participants, there should be an everlasting access to the menu. However, for smaller display areas, menu fixation can be counterproductive. Often the browser buttons require a lot of screen space. Sticking the website menu in place reduce the area for displaying the content, this way interactions may become harder to perform. For example, scrolling requires a certain space to perform the desired gesture. If this area is small, this can become difficult [21]. Taken together, this suggests avoiding fixation of menus on small displays.

In addition, the participants were asked to evaluate navigation elements that can appear and disappear compared to a permanent presentation. With large display areas, all participants favor the permanent version. On devices with small display, a collapsed menu variant is preferred. In relation to a changed device orientation, the menu display should remain constant in the opinion of the users, especially older participants. Younger users prefer a changed menu presentation, when all menu items can be shown directly.

Furthermore, instead of overwriting menu items, the respondents preferred an extension with sub items. They explained this with the possibility to compare menu options. Jumping back and forth between different subpages will also be avoided. Only two users disagree to the addition of menu sub items. They stated an overload resulting from the choice of elements. This phenomenon is called Paradox of Choice. Paradox since a larger selection is intuitively regarded as positive. In practice, however, too many options can be considered difficult or frustrating [22][23].

6) *Thumb zone*

Subsequently, the participants should define the menu placement they expect on different devices (out of 6 predefined menu positions). For the laptop and tablet (vertical and horizontal), the menu in the upper area received most votes, followed by the arrangement left. Similarly, most smartphone users prefer the menu at the top and left. A few expected the arrangement at the bottom of the screen. One of the respondents justified this decision with the phone’s handling. He unconsciously refers to the so-called thumb zone. The thumb performs most phone interactions. As a result, only one third of the screen, the so-called "thumb zone", can be reached effortlessly. For frequently used interaction elements (which are not limited to menus [21]), it is therefore recommended to place them within reach. For smartphones, this means at the bottom of the page.

In summary, from an ergonomic point of view, placing a menu at the bottom can lead to an improvement in usability. This additionally makes clear, that the mental model alone does not increase the usability of a website.

D. *First-click-test*

As the name suggests, a first-click test analyzes the users' first click on a user interface [24]. The probability of successfully completing a task on a website is twice as high if the first click was correct [25]. The first-click-test of this article (see Table 1, column "First-Click-Test") was set up in the form of the A/B-test concept [26].

For Original Version A and adapted Version B, the web presentation of the Hochschuljobbörse was used again. In the context of this article, the focus lies not only on the correct click but also on the time needed for it. Consequently, it can be identified how far the composition of the website corresponds to the mental model of the test person.

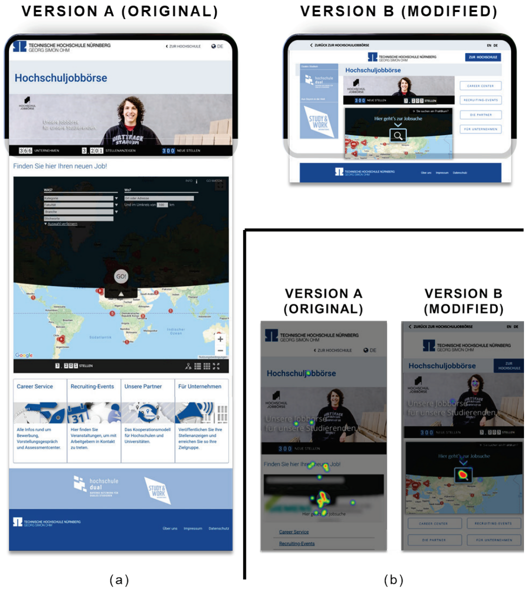


Figure 3. Horizontal (a) and vertical (b) smartphone version of the job search subpage.

1) *Right amount of information*

Different results between the test versions can already be seen in the first task. The respondents in the original version A found the menu 5 seconds faster on average than the respondents in version B. Due the compression in Version B, more elements are visible on one display area. As a result, the viewer may be flooded with information, which influences the information processing.

With increasing information input, the performance of human information processing increases linearly up to a threshold value. Above this threshold, however, performance degrades dramatically [27].

Consequently, more information input does not always have a positive effect for the viewer. However, avoiding information overload should not result in too less displayed information. These negative effects can be shown by looking at the results of the mobile representations.

In the horizontal position (see Figure 3 (a) right), the layout of the page has not been adapted to the changed device position, instead it has simply been scaled. The header image is enlarged to such an extent that the user can hardly recognize any further information. Consequently, the user lacks an overview of the site. In the modified version B, the layout was changed to fit the new device position (see Figure 3 (b)). As a result, nearly the same number of elements are displayed in both orientations. Thus, the sum of the information to be processed does not change or only minimally for the viewer. Due to this, respondents in test version B find the searched element three times faster than participants in group A.

2) *Webpage length*

As can be seen in Figure 3 (a), the page length is minimized by adjusting the layout. Consequently, participants had to perform fewer interactions to retrieve the desired data. This may also be a cause for the quick finding of the searched element. However, shortening the page length does not generally improve the user experience. For example, hiding content in order to reduce page load time [28] may shorten a webpage. A first-click test showed that participants needed 13 seconds longer to find a partially hidden section. This time loss is high when compared to the desired performance improvement [28]. Collapsing small elements should therefore be avoided.

3) *Orientation and recognition*

If the user can process the information presented more quickly, he or she will also orientate faster on the website [27]. The test also showed that the users' orientation can be guided with the help of highlighted elements. By highlighting, users found the desired element 5 times faster.

Additionally, a familiar presentation of interactive elements allows the user to recognize these faster. If a click is not followed by the expected action, a user may get frustrated. This effect can be reduced by following web design conventions. For example, buttons should be designed as rectangles with a three-dimensional appearance [29]. As a result, web users not only recognize interactive elements better, but the subjects of this test also clicked more precisely on the correct them (Figure 3 (b)).

The findings of this test further suggest the use of "anchor points" that do not or hardly change their position and form, regardless of the layout or design of the page. How relevant such an anchor can be is shown in the mobile version of the job search window. In Figure 4 (b), the light blue header represents this anchor with a button to return to the previous page. In the original version A (see Figure 4 (a)), all test subjects wrongly chose the X in the upper left corner of the browser window which closes the window instead of the correct one in right bottom corner.

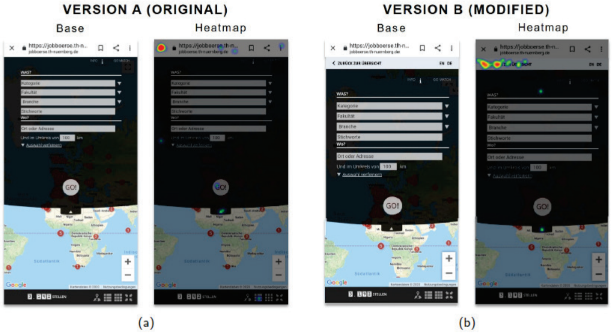


Figure 4. Vertical smartphone version of the job search formular subpage.

In addition, the test subjects needed an average of 16 seconds for this incorrect decision. Due to the everlasting anchor in the form of the header, the participants of group B needed only three seconds to click on the correct button.

4) Popups

Pop-up windows offer a way to convey important information without losing the context of the current screen [30] and to attract the user’s attention. However, this may be disadvantageously for the user. For instance, if the user wants to search for a job at the Hochschuljobbörse website first a window with internships information pops up. Once closed, however, the user can no longer access the pop-up without reloading the page.

To avoid such potentially confusing behavior, both test groups were asked to assess a modified version of the web page, where information about internships is not directly visible or overlap with the job search form. Only a field that refers to further information is displayed. In both test groups, most participants opted for no overlap version.

5) Menu features

A mental model can be changed, which is shown by the mobile menu presentation of the test. In the original version, the menu is presented as a hamburger icon and the position is unchanged (in the middle under the header). Based on the website analysis, menu icon is predominantly positioned in the top left corner. This placement was presented to the participants of the modified test variant B. Although the original version deviates from the usual arrangement, the subjects recognized both menu variants at the same rate. This suggests that the participants adjusted their mental model based on the desktop variant shown earlier.

Participants were also asked to choose one of two smartphone menu variants (seen in Figure 5). In test group A (see Figure 5 (a)), 96% opted for variant 1. The majority in group B (see Figure 5 (b)) also voted for version 1, but only 57%. Reasons for this can be found in the preference-test results. Users indicated to prefer the overview of all menu items. Since the menu items of test version A, variant 2 do not fit on one screen area, the participants presumably opt for variant 1. The reasoning is again confirmed by the narrow decision of the participants of group B. This is because all items are clearly visible on the screen in both versions. The fact that most of the participants in group B nevertheless tend towards variant 1 can be attributed to the fact that the initially collapsed menu items prevent the user from feeling overwhelmed.

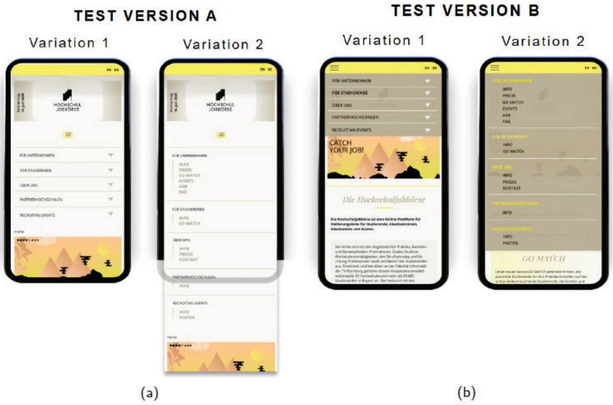


Figure 5. Mobile menu variants.

In the horizontal representation on the smartphone, the participants in test group A were shown a folded-out menu variant. Test group B still showed the hamburger icon in the upper left corner. It showed that the participants of the unchanged variant needed five seconds less to click on the menu than in the changed one. Due to the changed menu form, the user must first process the newly information, this takes time. The fact that older users emphasize with consistent menu representations (preference-test) is also reflected in this test. Participants between 18 and 34 years of age needed on average only 4 seconds longer to recognize the menu in the adapted version. For the older test subjects, it is 8 seconds more. From an information processing point of view, users benefit from a consistent website presentation.

IV. Conclusion

This article seeks to providing insights to the relation between adaption of webpages and human cognition. A series of experiments indicated, for instance, that users do not notice small adaptations if they lead to a web site that looks "the same" on different devices. For fast processing of information, the right amount of information should be presented on a display surface. Due to the increasing perceptual limitations in old age, this turned out to be especially relevant for older web users. In general, the test results usually differed due to age groups. For example, while the older generation prefers a constant website display across all devices, younger web users welcome an adjustment if it improves the web experience for them.

These results are limited in the following ways. First, only German participants have taken part in the test. Since experiences may differ due to cultural background, the results in this article should only be used for the German region. Second, the results are based on purely visual experiments and theoretical questions. Third, responsiveness should not be considered as an isolated variable when assessing the general user-friendliness of a website, which is also affected, e.g., by visual design and content selection.

Future work may build on these findings in various ways. Beside additional verification of the presented results by further studies, users' changing expectations concerning adaptive behavior suggest a continuous repetition of such studies – and eventually a representation of derived recommendations able to comprise the evolution of user expectations.

REFERENCES

[1] B. V. Usabilla, "User Experience Report: Is Responsive Design the Answer to the Growth in Mobile Devices?," 2014. [Online]. Available from: https://usabilla.com/blog/user-experience-report-responsive-design-answer-growth-mobile-devices/ [retrieved: 03/10/2021].

[2] N. B. Sarter, D. D. Woods, and C. E. Billings, "Automation Surprises," in Savendy, G. (Ed.), Handbook of Human Factors and Ergonomics (2nd Ed.), Wiley, NY, USA, 1997.

[3] Apple, "Human Interface Guidelines – Visual Design," 2021. [Online]. Available from: https://developer.apple.com/design/human-interface-guidelines/ios/visualdesign/adaptivity-and-layout/ [retrieved: 09/01/2021].

[4] J. Nielsen, "Mental models," Nielsen Norman Group, 2010. [Online]. Available from: https:/www.nngroup.com/articles/mental-models/ [retrieved: 09/01/2021].

[5] S. P. Roth, A. N. Tuch, E. D. Mekler, J. A. Bargas-avila, and K. Opwis, "Location matters, especially for non-salient features–an eye-tracking study on the effects of web object placement on different types of websites," International Journal of Human-Computer Studies, vol. 71, no. 3, pp. 228–235, ELSEVIER, 2013.

[6] L. Fischer, "Mental models as a central concept in the field of usability," netnodeblog.de, 2019. [Online]. Available from: https://www.netnode.ch/blog/mentale-modelle-im-bereich-usability [retrieved: 09/01/2021].

[7] A. Cooper, "About Face: The Essentials of Interaction Design," John Wiley & Sons, Inc., 2014.

[8] G. Dedre and L. S. Albert, "Mental Models", Psychology Press, 1983.

[9] statista, "Survey on the most popular types of websites in Germany 2015," Statista Research Department, 2015. [Online]. Available from: https://de.statista.com/statistik/daten/studie/486804/umfrage/beliebteste-arten-von-webseiten-nach-nutzungshaeufigkeit-indeutschland/ [retrieved: 09/01/2021].

[10] A. Poleshova, "Reach of mobile online offerings in Germany in October 2020," Statista Research Department, 2020. [Online]. Available from: https://de.statista.com/statistik/daten/studie/164615/umfrage/mobile-facts-2010-top-25-mobile-enabled-websites/ [retrieved: 09/01/2021].

[11] H. S. Tsai, R. Shillair, S. R. Cotten, V. Winstead, and E. Yost, "Getting Grandma Online: Are Tablets the Answer for Increasing Digital Inclusion for Older Adults in the U.S.?," Educational Gerontology, vol. 41, no. 10, pp. 695–709, 2015.

[12] A. Kunst, "Germany: Types of online services used in 2020," Statista Research Department, 2020. [Online]. Available from: https://de.statista.com/prognosen/999829/deutschland-genutzte-artenvon-online-angeboten [retrieved: 09/01/2021].

[13] Q. E. Wang, M. D. Myers, and D. Sundaram, "Digital natives and digital immigrants: Towards a model of digital fluency," Proc. of ECIS 2012 Proceedings, (39), 2012.

[14] akamai, "How loading time affects your bottom line," Neilpatel, 2011. [Online]. Available from: https://neilpatel.com/blog/loading-time/ [retrieved: 09/01/2021].

[15] K. Klingsieck, "research in the internship semester - test procedures/tests," blogs.uni-paderborn.de, 2015. [Online]. Available from: https://blogs.uni-paderborn.de/fips/category/durchfuehrung/befragung/ [retrieved: 09/01/2021]

[16] TH Nuremberg, "Hochschuljobbörse - our job portal for our students," 2020. [Online]. Available from: https://www.hochschuljobboerse.de/ [retrieved: 09/01/2021].

[17] A. M. Fjell and K. B. Walhovd, "Structural brain changes in aging: courses, causes and cognitive consequences", Rev Neurosci, 2010.

[18] Saarbrücker Zeitung, "Internet usage 2019 - mobile vs. Desktop," sz-medienhaus.de, 2019. [Online]. Available from: https://www.sz-medienhaus.de/internetnutzung-2019-mobile-vs-desktop/ [retrieved: 09/01/2021].

[19] S. Hoober, "Design for Fingers, Touch, and People, Part 1," UXmatters, 2017. [Online]. Available from: https://www.uxmatters.com/mt/archives/2017/03/design-for-fingers-touch-and-people-part-1.php [retrieved: 09/01/2021].

[20] D. Kunde, "The tablet replaces the laptop," CaptainGadget, 2015. [Online]. Available from: https://www.captain-gadget.de/tablets-ersetzen-laptops/ [retrieved: 09/01/2021].

[21] J. Clark, "Designing for Touch," A Book Apart, 2016.

[22] B. Schwartz, "The Paradox of Choice: Why More Is Less", New York: Ecco Press, 2004.

[23] A. M. Grant, and B. Schwartz, "Too much of a good thing: The challenge and opportunity of the inverted U," Perspectives on Psychological Science, SAGE, vol. 6, no. 1, pp. 61–76, 2011.

[24] UsabilityHub, "An introduction to first click testing," usabilityhub.com, 2020. [Online]. Available from: https://usabilityhub.com/guides/first-click-testing [retrieved: 09/01/2021].

[25] B. Bailey, "Firstclick usability testing," Web Usability, 2013. [Online]. Available from: http://webusability.com/firstclick-usability-testing/ [retrieved: 09/01/2021].

[26] E. Dixon, E. Enos, and S. Brodmerkle, "A/b testing," FMR LLC, 2011.

[27] M. Volnhals and B. Hirsch, "Information overload and controlling," Controlling & Managment, vol. 52, pp. 50–57, 2008.

[28] F. Maurice, "Better than display: None - hide content on smartphones correctly," maurice-web.de, 2014. [Online]. Available from: https://maurice-web.de/besser-als-display-none-inhalte-auf-smartphones-richtig-ausblenden/ [retrieved: 09/01/2021].

[29] H. Loranger, "Beyond blue links: Making clickable elements recognizable," Nielsen Norman Group, 2015. [Online]. Available from: https://www.nngroup.com/articles/clickable-elements/ [retrieved: 09/01/2021].

[30] K. Whitenton, "Overuse of overlays: How to avoid misusing lightboxes," Nielsen Norman Group, 2015. [Online]. Available from: https://www.nngroup.com/articles/overuse-of-overlays/ [retrieved: 09/01/2021].

[31] J. Bernacki, I. Błażejczyk, A. Indyka-Piasecka, M. Kopel, E. Kukla, and B. Trawiński, "Responsive Web Design: Testing Usability of Mobile Web Applications," Proc. of Asian Conference on Intelligent Information and Database Systems, pp. 257–269, Springer, 2016.

[32] A. Groth, and D. Haslwanter, "Perceived Usability, Attractiveness and Intuitiveness of Responsive Mobile Tourism Websites: A User Experience Study," Proc. of Information and Communication Technologies in Tourism, pp. 593–606, Springer, 2015.

[33] C. Mullins, "Responsive, Mobile App, Mobile First: Untangling the UX Design Web in Practical Experience," Proc. of the 33rd Annual International Conference on the Design of Communication, pp. 1–6, ACM, 2015.

An Accessible Portal to Teach Computer Science Modules to Typical and Special Needs Children: A Prototype

Davis Ward
Computer Science and Software Engineering
Auburn University
Auburn, AL USA
e-mail: dzw0042@auburn.edu

Quinterious Hall
Computer Science and Software Engineering
Auburn University
Auburn, AL USA
e-mail: qdh0003@auburn.edu

Daniela Marghitu
Computer Science and Software Engineering
Auburn University
Auburn, AL USA
e-mail: marghda@auburn.edu

***Abstract*— Society's increased reliance on technology has simultaneously increased the demand for people who can develop and design these new advancements. This has led to an influx of students looking to learn how to code and gain the technological skill set that is currently among the most marketable. Learning to code is challenging; without the right tools, resources, and assistance, it can be tough to build the foundation needed to understand key computer science fundamentals. The existing web platforms focused on assisting K-12 learners are competitive from an educational and technical perspective. There is a huge lack of virtual educational platforms that can deliver resources to students with disabilities through innovative accessible features and provide guidance to K-12 teachers that are trying to support this area. This lack of guidance is especially evident when examining resources available to teachers about increasing access and engagement of struggling learners including students with disabilities. The motivation of this paper is to introduce a prototype of a centralized portal, Accessible Virtual Learning, that implements user experience strategies and accessible usability principles aiming to be accessible to any student and also educators who need guidance on finding suitable materials. The success of this portal relies heavily on its ability to allow teachers and self-directed learners to facilitate curriculums effectively while maximizing student engagement, ease of learning, and digital assistance for students at various ages with different learning abilities, both physical and cognitive.**

Keywords-teaching computer science; accessibility; students with disabilities; human computer interaction.

I. Introduction

Over the next decade, the U.S. will have to adapt to technological advances (AI, Big Data, and Cybersecurity) by creating structures and implementing coordination strategies that take full advantage of the opportunities they present. This situation will only become more urgent: by 2026, Science and Engineering (S&E) jobs are predicted to grow by 13% compared with 7% growth in the overall U.S. workforce [1]. Yet, even as Science Technology Engineering and Math (STEM) competencies have become more essential, U.S. K-12 mathematics and science scores are well below those of many other nations and have stagnated [2].

Women, underrepresented minorities, and people with disabilities remain inadequately represented in S&E relative to their proportions in the U.S. population. The rapid growth of S&E jobs and demographic changes have outpaced the progress that has been made in the participation of these groups in S&E. Increasing STEM skills and opportunities for all Americans requires local, state, and federal governments, public and private educational institutions, community organizations, and industry to step up their efforts. Earlier intervention is needed to advance STEM education and careers [3].

Society's increased reliance on technology has simultaneously increased the demand for people that can develop and design these new advancements. This has led to an influx of students looking to learn how to code and gain the technological skill set that is currently among the most marketable. Learning to code is challenging and without the right tools, resources, and assistance, it can be tough to build the foundation needed to understand key Computer Science (CS) fundamentals.

In the past decade, there has been increased awareness of the importance of teaching CS basics to students prior to college. Many high schools now offer the opportunity to at least experience the rudimentary principles in developing software. This has led to more students being prepared to tackle a college curriculum in STEM and being more successful in developing the skills necessary to pursue a career in computer CS. However, there seems to be a lack of resources that allow special needs students to achieve this knowledge as efficiently as typical students can [4]. Despite the commitment of the CS education field to increasing equity within CS education, there is still limited guidance for K-12 teachers on how to support a broad range of learners in CS education. This lack of guidance is especially evident when examining resources available to teachers about increasing access and engagement of struggling learners including students with disabilities.

The existing web platforms focused on assisting K-12 schoolers in their CS endeavors are competitive from an educational and technical perspective. The next step is to make these platforms effective for a wide range of learners (e.g., students with special needs). This requires compiling several accessibility features to cater to both special needs and typical students to ensure an equal opportunity to learn [5]. In this paper, we:

1. Evaluate the basic needs of both typical and special needs students
2. Discuss the design goals for making sure these needs are met, and
3. Explain our development process and how it improves CS learning modules for K-12 special needs students.

Learning needs of students differ, and the combination of accessibility needs for digital tools and the gap in digital literacy across socioeconomic and racial/ethnic lines create an inequitable environment in early CS education. These challenges are evident in the annual reports from the K-12 education community declaring the lack of diversity and equity in CS classrooms and a call for action [6]. The web portal will implement the Universal Design for Learning (UDL) framework toward providing a centralized location for learning by compiling several common coding platforms designed for high schoolers and beginner learners. The design of this portal will require the implementation of significant accessibility features such as voice navigation and other assistive elements to make these resources accessible for students with any form of disability or condition that could affect their learning.

The success of this portal relies heavily on its ability to allow teachers and self-directed learners to facilitate curriculums effectively while maximizing student engagement, ease of learning, and digital assistance for students at various ages with different learning abilities, both physical and cognitive. The development of this website has always focused on the user first, before design or development. The Accessible Virtual Learning (AVL) portal will be able to deliver collections of resources in a visually pleasing, accessible, and engaging way.

In Section II we will discuss related platforms and accessibility tools. In Section III we will discuss the study of Human-Computer Interaction. In Section IV we will discuss the design approach, architecture of the platform, and accessibility features. Lastly, In Section V we will discuss the conclusion of our work and task moving forward to improve the prototype.

II. BACKGROUND AND RELATED WORK

There are several disabilities that can be a challenge to effectively use computers and other technologies. Many accessibility features have been introduced since the commercialization of the personal computers to help people with disabilities use technology more easily.

Closed captioning is the display of text on a screen from the audio portion of a video. This allows a user to read any spoken dialogue, music, or even register sound effects and has been an instrumental accessibility feature to ensure material is available to individuals who are deaf or have impaired hearing. Closed captioning differs from subtitles as it provides greater accuracy and includes dialogue, an explanation of sound effects, and identification for who or what is currently speaking.

Keyboard shortcuts were introduced as an accessibility feature to allow users to access a site in its entirety using only typed commands. Many users with disabilities are not able to use a mouse or pointer to navigate the interface of a computer. Keyboard shortcuts have also become common among typical users, who usually only use keyboard shortcuts for certain tasks. An ongoing issue with keyboard shortcuts is that computer interfaces are normally designed to work best with the combination of both a keyboard and mouse being used. Navigating a site using the keyboard exclusively can easily become more cumbersome than using a mouse, seemingly creating an entirely separate user experience for users with disabilities. The task of creating a system that has seamless integration of navigation accessibility features will have to overcome the challenge of making sure navigating a site via keyboard exclusively has comparable utility as using a mouse and keyboard combination.

In the mid-twentieth century, barrier-free design and accessible design terms were introduced to illustrate efforts to remove physical barriers to people with disabilities [7]. Over the years with the technology advancements, there have been many improvements on how information is being presented to students with vision-impairment, Screen Readers being the dominant mechanism.

A Screen Reader is a software application that converts text and/or images from a screen to the speech format that visually impaired people can understand and interact with. Many screen readers are also compatible with the websites developed under accessibility standards. The main disadvantage of Screen Readers is that blind users need to go through an abundance of irrelevant content before they find what they were looking for [8]. This problem can be resolved by using an interactive JavaScript speech recognition library that gives the speech control of the application to the user. This allows an application to use the device's microphone and receive speech. The speech is then converted to text that is subsequently matched against a list of commands that would initiate a corresponding action for the user navigating the site.

Text-to-speech is a commonly used feature that ultimately allows text from your mobile device or your computer to be read to you aloud. Text-to-speech has drastically improved the access of information for the visually impaired specifically. An issue with text-to-speech, since its introduction as an accessibility feature. has been that the text is usually read by a computerized voice that can frequently mispronounce or distort the natural phonics of a word, making it difficult at times for someone who is using the feature to accurately interpret what is being said and affecting their overall literacy over time. Over the years, this issue has been addressed by developing more natural sounding text-to-speech systems that are almost indistinguishable from humans.

Moodle is a platform that has made it is to be fully accessible and be able to accommodate all users regardless of what their learning needs may be [18]. Their interface is tested with a range of screen reader software and is developed to comply with most accessibility standards. Totara's corporate e-learning platform provides the same accessible learning modules for business and organizations to perform training needs and employee onboarding [19]. E-learning platforms need to be available to a wide audience; and implementing as many accessible tools as possible only increases the potential audience that the platform could reach.

III. LITERATURE STUDY METHODOLOGY

The way technology has revolutionized the world socially, economically, and politically has been seismic and is clearly only scratching the surface. In a matter of a few years, the Internet has become one of the widely used technologies that has changed the way we communicate, learn, or do business. A 2019 report by Internet World Stats shows that the number of internet users has increased by almost 1150% since 2000 and 4.39 billion active internet users in 2019 [9].

Human-Computer Interaction (HCI) is a study of design, implementation, and evaluation of an interactive computing system for human use and for studying the major phenomena surrounding them. The accelerating growth of the Internet and the technology boom has led a number of schools and universities to provide courses and degree programs via distance education. HCI research has made it possible for students with disabilities to have the necessary accommodations for an equal opportunity to gain an education online. In most STEM fields, it is imperative for students to at least be moderately proficient at math. Students with disabilities are often at a disadvantage when it comes to understanding complex formulas and interpreting important visualizations. As HCI has evolved, students with vision-impairments have been able to close the gap with MyA+ Math, an accessible learning platform that has interactive resources to help the visually impaired learn key math concepts [17]. With the evolution of HCI, developers have also been able to explore new ways to make the interaction between humans and computer easier [10].

Software engineers have very quickly risen to the top of the totem pole in job outlook, and technology companies can only hope that the supply of skilled developers will one day match the demand. Developing software is a strenuous task. Learning to develop software is even more difficult and compounds the challenge of knowing what code to write on top of knowing how exactly to write it. Therefore, it is important to identify and alleviate any additional challenges that are not inherent in the process of learning CS. Making CS easier to learn is not the objective. The objective is to ensure we are not making it more difficult to learn than it already is.

When analyzing the challenges presented to students learning to code, it is clear the learning curve gets steeper for students with special needs. This is simply due to the fact that learning resources have not catered their curriculums or platforms for this specific demographic and lack even basic components necessary to ensure special needs students can learn just as efficiently as typical students. Coding is extremely visual and intellectual. If there are students with visual impairments or cognitive disabilities, it presents several obstacles that may make it difficult for these students to even begin their learning process.

Special needs students are frequently provided the opportunity for accommodations for face-to-face and traditional instruction methods. It is imperative to activate the same policies for online learning platforms. The W3C Web Content Accessibility Guidelines(WCAG) [11] provide a framework for ensuring basic accessibility needs are met; and all platforms should be complying with these to meet the needs of students and reach a larger demographic of learners with their resources.

IV. DESIGN AND IMPLEMENTATION

We designed the AVL portal with the following objectives:

- Provide a clean and easy to understand user interface for the user to create an account and get the wanted resources.
- Adhere to W3C Accessible standards that allows the use of Screen Readers and other accessibility tools.
- Provide a clean color scheme and font sizes that are accessible to visual disabilities.

A. *Technology Used*

Accessible Virtual Learning is implemented using Hypertext Markup Language 5 (HTML), Cascading Style Sheets (CSS), Embedded JavaScript Templates (EJS), and JavaScript for the front-end. For the back end, Node.js runtime engine [12] along with Express framework is used. A MySQL database is used for storing user's information such as name, email, encrypted password, user role such as 'educator' or 'student', and foreign keys for module ownership. There also exist tables that store module information for educator resource allocation. The blog portion of AVL uses ghost.io, an online publishing platform that makes content administration tasks secure and straightforward. It also uses a RSS feed to add articles to the AVL blog that are related to our content space, along with the articles that our content creators publish. The backend also uses the following open-source JavaScript packages and middle-wares [13]:

- Sequelize is a promise-based Node.js object-relational mapper that is used for the MySQL database models and querying.
- Bcrypt.js is a JavaScript package that allows proper password hashing for privacy and security of user profiles.
- Passport.js is a Node.js authentication middleware that facilitates the AVL login system.
- Annyang.js is a JavaScript speech recognition library used for voice navigation on web apps. Custom voice commands and actions can be created that allows AVL to be more accessible to users with

visual disabilities. It is especially useful for user navigation purposes.

- Connect-flash is a JavaScript package that is used in AVL to create robust user feedback related to form interaction. This facilitates all the back-end form validation messages to the user.
- Express-validator is a JavaScript package that assists in the back-end form validation and sanitation logic.

B. AVL Portal Architecture

The AVL portal follows a model-view-controller architecture [14]. There are models that are representations of data that are being posted and manipulated by the controllers. The program logic and database manipulation are done in the controllers that pass on data to the view, in the form of EJS templates that serve HTML/CSS/JS pages to the user. The architecture is facilitated through Express routes that are used for knowing what the user wants to see or interact with and calls the appropriate controllers to interact with the data, and then sends the appropriate view with that data. In the following, we detail the implementation of different components implemented in AVL

1) Dashboard Component

The dashboard delivers content and functionality to both educator and student accounts and is the first thing the user sees after logging into AVL. The dashboard features a list of resources, called modules, that educators can add – such as an article, resource, or course along with an URL to the resource. Educators can create, edit, and delete modules. Modules are stored in the MySQL database within their own table, with a foreign key connecting it to the educator who created it. Students can then view and sort through the modules that all educators have created. Modules are contained by a card user-interface that is in a list that can be navigated by keyboard, which is especially important for screen readers. Users can also consume modules by the author, through the educator page. This is important if a student is using the website to get resources specifically from their educator.

2) Blog

The blog allows educators to create blog posts that surround the topics of accessibility and CS education. These blog posts are meant to be read by both educators and students. Educators can create, edit, and delete blog posts. The posts are then displayed in an accessible way. The blog also generates content from news feeds on relevant technological topics in order to maintain a fresh collection of articles to read whenever a user logs in. Students can view the entire archive of blog posts and articles but are not allowed to post, edit, or delete any content. Blog content is strictly informative and should act as an extension of the learning modules within AVL to facilitate extracurricular learning not directly related to coursework. The blog was implemented using ghost.io [14] to facilitate the type of content but also the content authors. Since this is a public facing portal and ultimately anyone can create an educator account, we decided to have the ability to choose which educators can create content for the AVL blog.

3) Resources Component

The resources page is a collection of resources that the site creators collected. These are resources that are notable in usability and popularity in the education and computer science space. The resources are also displayed with a card user interface.

4) Voice Navigation

The voice navigation feature allows users to explore the different features of the portal non visually. When a user logs in, a large voice icon on the bottom right of the screen is presented, which will also be accessible to a screen reader. This button displays a banner over the whole web app that explains how to use the voice navigation. The only purpose of this button is to display those instructions. When a user is on any given page, they can speak any of the following commands to be redirected to the desired page. This component was implemented with Annyang.js, where all of the voice commands and their desired actions were added. With this library, more custom voice commands can be added in the future to extend the scope of the portal. The voice commands can be seen below in Table 1.

TABLE I. VOICE COMMANDS

Voice Commands	*Action*
‘Home’	Redirect to the index page
‘Dashboard’	Redirect to the dashboard page
‘Educators’	Redirect to the educators’ page
‘Blog’	Redirect to the AVL blog
‘Resources’	Redirect to the resources page
‘Log out’	Log the user out of the portal

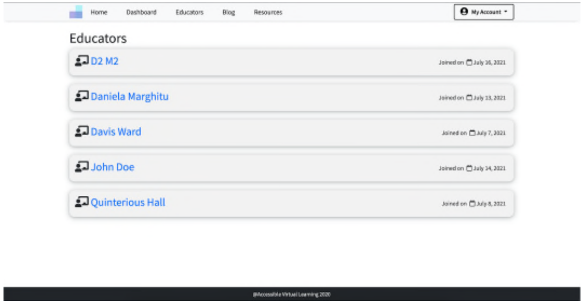
C. User Interface

This section will contain screenshots of the pages, features, and functionality of the portal that has been developed. As this is still a prototype, the look and scope of the portal may change in the future. They may be changed based on feedback from students and educators.

The user interface uses a CSS framework, Bootstrap 5, to aid in the development of the views. This framework is especially helpful when creating mobile first applications. The use of Bootstrap 5 components does speed up the development process of user interfaces, but developers must be careful to add extra html attributes and hidden text, as not all bootstraps are natively accessible to W3C standards.

1) User’s View

Figure 1, Figure 2, and Figure 3. are screenshots of the educators’ page, resources page, and AVL blog. These are pages that all users can access, and do not change based on the user type.



Educators

D2 M2 Joined on July 16, 2021

Daniela Marghitu Joined on July 13, 2021

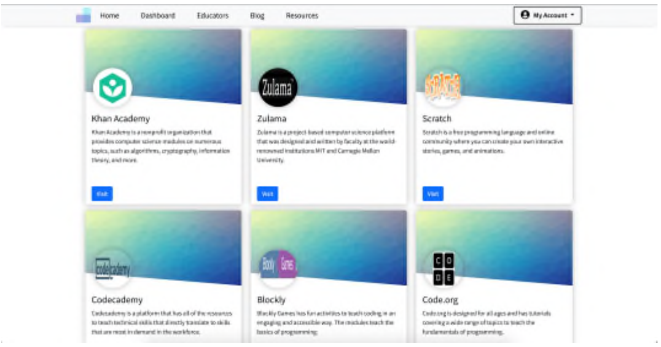
Davis Ward Joined on July 7, 2021

John Doe Joined on July 14, 2021

Quinterious Hall Joined on July 8, 2021

@Accessible Virtual Learning 2020

Figure 1. Educators' page.



Khan Academy

Khan Academy is a nonprofit organization that provides computer science modules on numerous topics, such as algorithms, cryptography, information theory, and more.

Zulama

Zulama is a project-based computer science platform that was designed and written by faculty at the world-renowned institutions MIT and Carnegie Mellon University.

Scratch

Scratch is a free programming language and online community where you can create your own interactive stories, games, and animations.

Codecademy

Codecademy is a platform that has all of the resources to teach technical skills that directly translate to skills that are most in demand in the workforce.

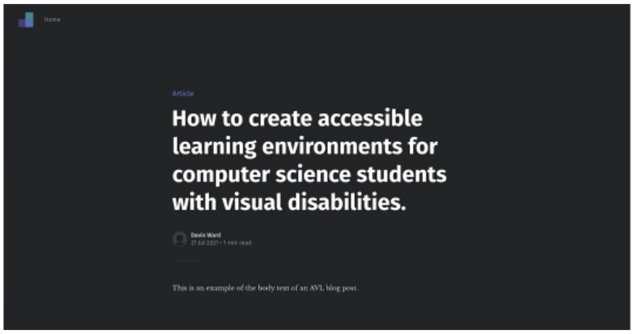
Blockly

Blockly Games has fun activities to teach coding in an engaging and accessible way. The modules teach the basics of programming.

Code.org

Code.org is designed for all ages and has tutorials covering a wide range of topics to teach the fundamentals of programming.

Figure 2. Resources page.



Article

How to create accessible learning environments for computer science students with visual disabilities.

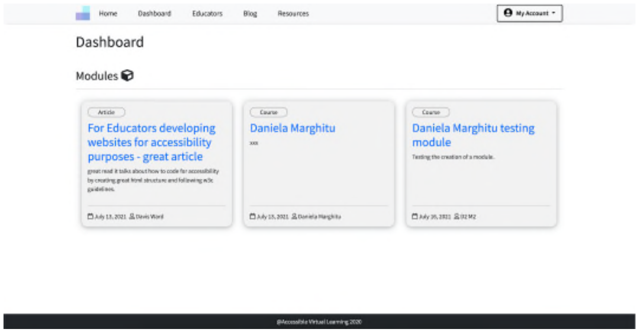
Davis Ward
27 Jul 2021 • 1 min read

This is an example of the body text of an AVL blog post.

Figure 3. AVL Blog page.

2) *Student's View*

Students can interact and explore resource modules but cannot create modules. The cards are keyboard focusable and navigable. The view also dynamically sizes the card based on the width of the viewing screen, the number of modules, and the amount of content within each module. Figure 4. shows the card user interface for the student's dashboard page.



Dashboard

Modules

Article
For Educators developing websites for accessibility purposes - great article
great read it talks about how to code for accessibility by creating great html structure and following w3c guidelines.
July 13, 2021 Davis Ward

Course
Daniela Marghitu
xxx
July 13, 2021 Daniela Marghitu

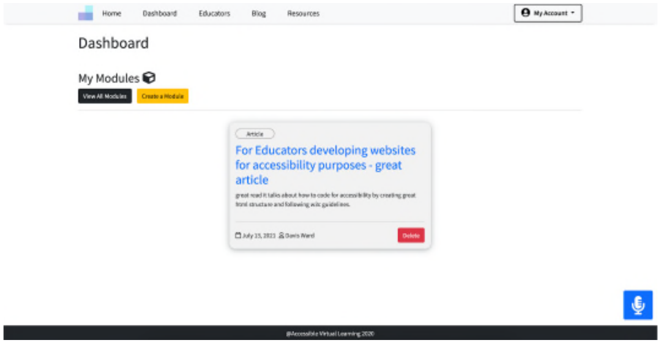
Course
Daniela Marghitu testing module
Testing the creation of a module.
July 16, 2021 D2 M2

@Accessible Virtual Learning 2020

Figure 4. Student's view.

3) *Educator's View*

Educators can consume the resource modules, but also view them as the student would. Figure 5 shows the card interface showing the educators own modules, and Figure 6 the form for creating a module. All of the educator's views are accessible in the same way that the student's views are. This is important because it demonstrates the opportunity for students with disabilities to become educators with disabilities.



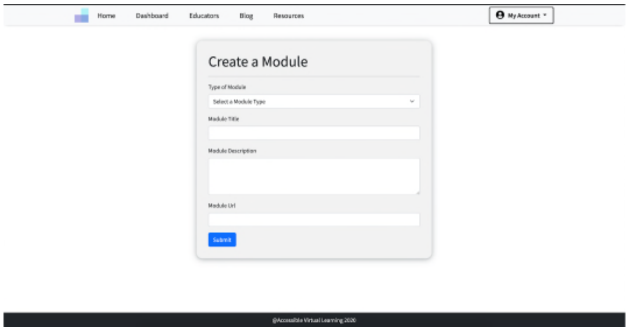
Dashboard

My Modules

Article
For Educators developing websites for accessibility purposes - great article
great read it talks about how to code for accessibility by creating great html structure and following w3c guidelines.
July 13, 2021 Davis Ward

@Accessible Virtual Learning 2020

Figure 5. Educators' view.



Create a Module

Type of Module
Select a Module Type

Module Title

Module Description

Module Url

@Accessible Virtual Learning 2020

Figure 6. Educators' form for creating a module.

D. Testing and Accessibility

1) Design for Accessibility

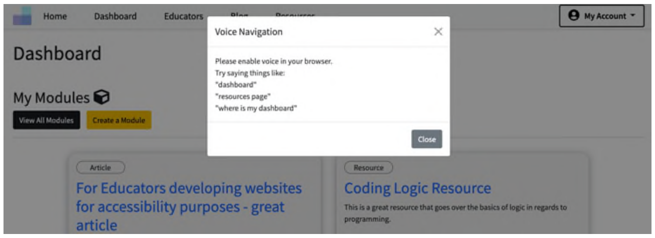
According to the Web Content Accessibility Standard 2.0 (WCAG 2.0), the following are important requirements for making web apps accessible:

- Text alternatives that serve equivalent purpose for all non-text content
- Text can be resized up to 200% without losing content functionality
- Users can operate the site using keyboard-based navigation options
- Users can access content with the use of assistive tools like screen readers
- Text to background contrast must be a 4.5:1 ratio at a minimum

An accessible portal that adheres to W3C standards must be designed with strong, semantic, and structural HTML that closely follows the guidelines. When designing the user interfaces, the HTML is the first thing that was focused on, as styles can be added after to create a better-looking view. Many Accessible Rich Internet Application (ARIA) [15] attributes were introduced natively to HTML 5. For example, many of the buttons on the site are instead used as anchor tags, but with a role attribute of the button. This functionality means to screen readers that it is not a link, but a button that a user clicks.

The use of icons is also used heavily on the portal, both semantically and decoratively. For decorative icons, the aria attribute of aria-hidden should be set to true, so that a screen reader will simply skip over the icon tag. Since it does not display any meaning, it is not necessary for visually impaired users to digest. However, for semantic icons, is it extremely important to use accessible html because these icons display important meanings for the content that is next to, or below them. A span tag must be added after the icon that contains the textual meaning of the icon and is hidden to visual users, but not to screen readers. This way a visually impaired user can have the same experience as a visual user. Content images can also enhance visual user experience, but they must have appropriate alternative text for visually impaired users to receive the same experience.

Adding native voice navigation to a web portal is a huge advantage. Screen readers are advanced enough to make navigating a website using auditory and physical sense inputs and responses easy but having the ability to navigate pages instantly through speech makes it even more accessible for these users. With a feature like this, it is paramount that the instructions to use the voice navigation are easily consumed by the screen reader, or the feature itself would be unusable without third party assistance. When a student clicks on the voice navigation symbol, instructions pop up that also dim the rest of the page. The voice navigation feature can be shown below in Figure 7.



Please enable voice in your browser.
Try saying things like:
"dashboard"
"resources page"
"where is my dashboard"

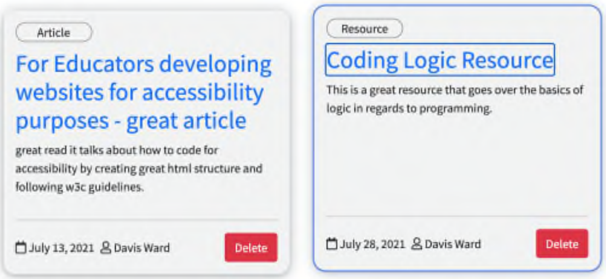
For Educators developing websites for accessibility purposes - great article

Coding Logic Resource

This is a great resource that goes over the basics of logic in regards to programming.

Figure 7. Voice navigation feature.

Additionally, making sure the web portal was navigable by keyboard was an important standard. It should not only be navigable by keyboard, but when a user is focused on a certain user interface element, the element should show a visual cue to let the user know where they are on the structure. For example, when a user is on the AVL dashboard, and navigates to focus on one of the module cards through keyboard action, the card is moved in an upright position and a colored border appears. The user can then click enter to navigate to the URL of the module they are focused on. The unfocused and focused states can be seen in Figure 8. The hover effect is also the same as the focused effect for users navigating by mouse.



For Educators developing websites for accessibility purposes - great article

great read it talks about how to code for accessibility by creating great html structure and following w3c guidelines.

July 13, 2021 Davis Ward

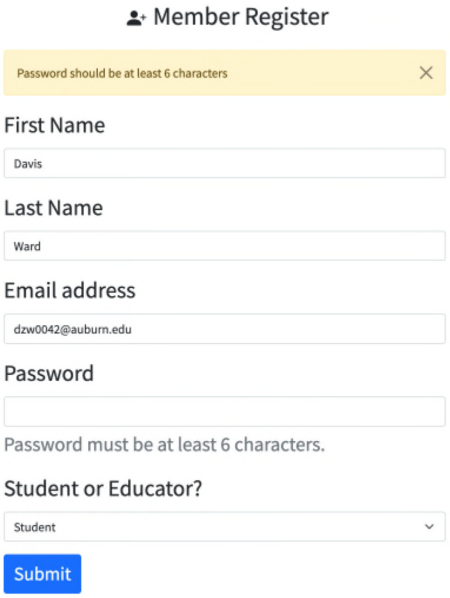
Coding Logic Resource

This is a great resource that goes over the basics of logic in regards to programming.

July 28, 2021 Davis Ward

Figure 8. Unfocused and focused states of modules.

Forms are also a very important part of accessible web apps. They must have clear labels that correspond to each input area so that the user knows what each input is for, and also so that screen readers can correctly convey the form. AVL forms use server-side validation and user notifications to display success, warning, and error messages regarding the submission of the form. The notification should clearly state which label-input field was not sufficient to let the user know exactly how to fix it to successfully submit the form. Figure 9 shows the user interface of the member registration form, and what happens when the password field input is not sufficient. It states what label was incorrect, and what was incorrect with it.



Member Register

Password should be at least 6 characters ×

First Name

Davis

Last Name

Ward

Email address

dzw0042@auburn.edu

Password

Password must be at least 6 characters.

Student or Educator?

Student

Figure 9. Validation of a member registration form.

2) *Testing for Accessibility*

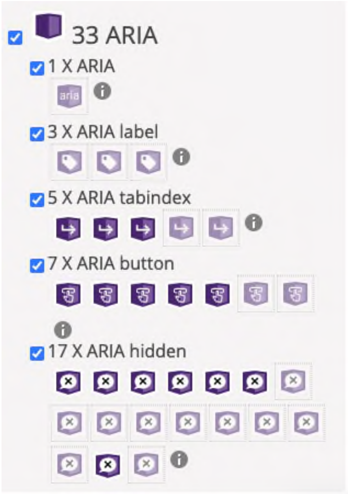
Testing for accessibility is just as important as designing for it. The testing process for the portal includes the following main steps:

1. Code review of the HTML structure and correct attributes, alt tags, and hidden ARIA text.
2. Testing of the live site by interacting with all features and assuring expected behavior.
3. Testing with the Web Accessibility Evaluation Tool (WAVE) [16], a chrome extension that displays any errors against the W3C guidelines by parsing the HTML structure of each page.
4. Comparing the web portal against the WCAG 2.0 Web Accessibility Checklist to assure all requirements were met.

Going through this type of testing is much more robust then only doing the first step. It assures that the site is indeed accessible, as it is very easy to skip over critical accessibility problems.

Out of the 47 guidelines that are specified in the WCAG 2.0 Web Accessibility checklist, all were met except 4. These are planned to be fixed in the near future before testing with actual student subjects. Specifically, the readability of the site can be improved, with alternate text for information that is past a lower-secondary reading level. These areas will be highlighted with the actual test subjects and alternate text will be provided. Additionally, text-based help needs to be added for the module functionality. We do think it is intuitive enough to be used without discrete instruction, but this may prove untrue in subject testing.

The test results using WAVE were very promising as we used it concurrently throughout the development phase. Each page was updated to ensure a result of 0 errors in WAVE. It also shows all of the aria attributes and ensures color contrast is acceptable. Some results from WAVE can be seen below in Figure 10.



☑ 33 ARIA

☑ 1 X ARIA

☑ 3 X ARIA label

☑ 5 X ARIA tabindex

☑ 7 X ARIA button

☑ 17 X ARIA hidden

Figure 10. WAVE ARIA attribute results.

V. CONCLUSION AND FUTURE WORK

The AVL platform meets the original objectives and design goals. As a prototype, it needs to be tested and expanded on. In any online platform, accessibility should be a requirement, not a design goal. If platforms continue to strive to implement key features that allow inclusion for all, the online education space will make great progress. We hope we made a difference by examining accessibility on the web and turning our findings into a working prototype.

In the future, we plan on testing the portal with a range of different students to ensure that the user interface is easy to follow and understand. We would also like to see different students using the site with screen readers and testing our voice navigation. The input from this type of testing would be insightful and lead us to making smart changes. Testing would be done by selecting a range of students with different disabilities ranging from visual to developmental. We would not be allowed to assist the student at all with creating an account and interacting with learning modules. If a student ever became stuck or confused, then we would note the point of frustration and acknowledge a change needed for that functionality or content. This would ensure that not only is our site accessible, but usable by a large range of students and educators with different needs.

Technology continues changing the way we learn. If the goal is to maximize the potential that technology has as a resource for knowledge, it is imperative to ensure that this resource is available to everyone and can be used by any demographic to gain new skills, talents, and abilities.

A different approach to the need of more accessible resources for learning platforms relating to students with disabilities could have been a tool, rather than a portal. A tool that could be used for multiple purposes might have reached more learners, but after the initial research of the idea, creating a portal seemed to be a more direct approach. One of the biggest challenges for this project was deciding what specific needs the leaners would require, and how to implement them into our portal through functionality. We believe we did a good job of this in respect to the time and knowledge of our work, but after subject testing, the portal could be vastly improved with more specific accessible functionality that scopes past WCAG requirements.

This virtual learning platform is created to take the first step into providing a universal learning experience. As technology continues to evolve, the resources available to enhance learning will also advance. In the future, the following additions could be included to improve the site:

- Providing new and updated learning material is key for maintaining interest in a skill area. We are planning to add an RSS feed to the blog portion of the site in order to keep a constant flow of new content to keep learners attracted. Articles will be relevant to learning modules on the platform.
- The platform will give students the opportunity to be content creators on the site's blog after they have displayed a certain level of proficiency in their learning. Their content will be moderated by their respective educators.
- The platform will allow students to create a profile based on their interests and learning objectives. This would then be used to recommend public other relevant resources.
- The current version of the platform uses voice navigation and has an established list of commands a user is allowed to use. A future version of the site will have custom voice commands added to expand the utility of this feature and improve the overall user experience.
- The site will be updated once extensive user subject testing trials are completed, to ensure usability and to make sure the site meets all the needs of the target audience.

ACKNOWLEDGMENT

This work was made possible due to NSF Grant # 1842092.

REFERENCES

[1] National Science Board, National Science Foundation. 2020. Science and Engineering Indicators 2020: The State of U.S. Science and Engineering. NSB-2020-1. Alexandria, VA. [Online]. Available from: https://ncses.nsf.gov/pubs/nsb20201/ 20021.07.22

[2] NSB, “Elementary and Secondary Mathematics and Science Education,” Science & Engineering Indicators 2020.

[3] S. Grover, S. Cooper, and R. Pea, “Assessing Computational Learning in K-12.” ITiCSE '14 (p. 5). Uppsala, Sweden, June 2014.

[4] R. Ladner, M. Israel, “For All in Computer Science for All.” Communications of the ACM 59, no. 9 pp. 26-28, 2016.

[5] M. Ray, M. Israel, C. Lee, and V. Do, “A Cross Case Analysis of Instructional Strategies to Support Participation of K-8 Students with Disabilities in CS for All.” In Proceedings of the Association for Computing Machinery (ACM) Technical Symposium on Computer Science Education. (SIGCSE), pp. 900-905, 2018.

[6] J. Wang, H. Hong, J. Ravitz, and S. Hejazi Moghadam, “Landscape of K-12 Computer Science Education in The US: Perceptions, Access, and Barriers.” Paper presented at the Proceedings of the 47th ACM Technical Symposium on Computing Science Education, pp. 645-650, 2016.

[7] E. Ostro,: Universal design: an evolving paradigm. Universal design handbook 2,34{42 2011

[8] I. Ramakrishnan, V. Ashok, and S. M.Billah: Non-visual web browsing: Beyond web accessibility. In: International Conference on Universal Access in Human-ComputerInteraction. pp. 322{334. Springer 2017.

[9] Miniwatts Marketing Group. Internet Usage Statistics: The Internet Big Picture World Internet Users and Population Stats 2019. [Online]. Available from: https://www.internetworldstats.com/stats.htm. 2021.07.22

[10] Lawrence, D.O., Ashleigh, M.: Impact of human-computer interaction (hci) onusers in higher educational system: Southampton university as a case study. International Journal of Management Technology 6(3), 1{12 2019.

[11] Web Content Accessibility Guidelines (WCAG) [Online]. Available from: https://www.w3.org/WAI/standards-guidelines/wcag 20021.07.22

[12] EDUCBA [Online]. Available from: https://www.educba.com/javascript-vs-node-js/ 20021.07.22.

[13] JavaScripting [Online]. Available from: https://www.javascripting.com/, 20021.07.22.

[14] Codecacademy [Online]. Available from: https://www.codecademy.com/articles/mvc 2021/07/22.

[15] Accessible Rich Internet Application (ARIA) [Online]. Available from: https://developer.mozilla.org/en-US/docs/Web/Accessibility/ARIA , 20021.07.22.

[16] WAVE Web Accessibility Evaluation Tool [Online]. Available from:

https://wave.webaim.org/, 20021.07.22.

[17] Jariwala, A., Marghitu, D., Chapman, R.: A multimodal platform to teach mathematics to students with vision-impairment. In: Antona, M., Stephanidis, C. (eds.) Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments. pp. 109–117. Springer International Publishing, Cham (2021).

[18] Moodle [Online]. Available from: https://docs.moodle.org/311/en/Accessibility 2021.07.22

[19] Totara [Online]. Available from: https://help.totaralearning.com/display/TPD/Accessibility+at+Totara 2021.07.22

Neural Speech Synthesis in German

Based on Tacotron 2 and Multi-Band MelGAN

Johannes Wirth, Pascal Puchtler, René Peinl
Research Group System Integration
Hof University of Applied Sciences
Alfons-Goppel-Platz 1, 95028 Hof, Germany
e-mail: Johannes.Wirth.3@iisys.de, Pascal.Puchtler@iisys.de, Rene.Peinl@iisys.de

***Abstract*—While many speech synthesis systems based on deep neural networks are thoroughly evaluated and released for free use in English, models for languages with far less active speakers like German are scarcely trained and most often not published for common use. This work covers specific challenges in training text to speech models for the German language, including dataset selection and data preprocessing, and presents the training process for multiple models of an end-to-end text to speech system based on a combination of Tacotron 2 and Multi-Band MelGAN. All model compositions were evaluated against the mean opinion score, which revealed comparable results to models in literature that are trained and evaluated on English datasets. In addition, empirical analyses identified distinct aspects influencing the quality of such systems, based on subjective user experience. All trained models are released for public use.**

Keywords: Text-To-Speech; German; Tacotron 2; Multi-Band MelGAN.

I. Introduction

The quality of speech synthesis or Text To Speech (TTS) systems has leaped since deep neural networks are being leveraged. Whereas such systems acted as a niche technology a few years ago, today every voice assistant and a large number of car models are equipped with their own, manufacturer-specific, synthetic but increasingly natural-sounding voices. However, smaller companies interested in using TTS in their products or services mostly have to rely on large-scale software providers, or alternatively, freely available models as investments in in-house solutions would often be financially unfeasible.

Since state-of-the-art models with permissive licenses exist almost exclusively for English, several model compositions based on Tacotron 2 [1] and Multi-Band MelGAN [2] were trained for the German language and published for free use. This work describes the processes that were carried out to train these neural networks and provides a corresponding evaluation based on the Mean Opinion Score (MOS), setting an initial benchmark for future systems in German. The described models and results are part of the development of a smart speaker system.

The rest of the paper is structured as follows. In Section II, state-of-the-art of deep neural networks used for TTS are presented and available datasets for German TTS are reviewed in section 3. In section 4, key learnings from training of selected network models are described further. Section 5 describes how the evaluation of synthetic voices was implemented and presents the results, which are interpreted in a subsequent discussion in section 6 and put into perspective by limitations in section 7. Lastly, the work is concluded with a summary and an outlook in section 8.

II. BACKGROUND

Most state-of-the-art systems for speech synthesis based on neural networks consist of two components: an acoustic model and a vocoder. The acoustic model generates an intermediate representation called mel spectrogram from input characters or phonemes, while the vocoder converts this representation into a final audio signal. The following subsections describe the general principles of operation of both components in more detail and presents several architectures. An overview of model compositions already evaluated in literature is given in TABLE II.

A. *Acoustic Model*

Acoustic modelling defines the task of encoding an input sequence of characters to a hidden representation and the subsequent prediction of mel spectrogram frames per time step. The formerly common models for mel spectrogram generation based on Hidden Markov Models (HMMs) [3] have been increasingly replaced by approaches based on deep learning in recent years. In particular, Tacotron [4] and its successor Tacotron 2 [1] have led to a dramatic increase of quality in speech synthesis research. While Tacotron still uses a Griffin-Lim vocoder as a second stage, only reaching a MOS of 3.82, Tacotron 2 succeeds in achieving a MOS value of 4.53, which is very close to the value of human speakers (4.58), by using a continuous deep learning-based process. For the latter, a modified version of WaveNet [5] was used as a vocoder.

While Tacotron is based on Recurrent Neural Networks (RNNs), which are commonly used for speech synthesis, Transformer TTS [6] successfully applied the transformer architecture [7], which became well-known from the domain of natural language processing with models such as BERT [8], to speech synthesis, achieving similar or slightly better scores than Tacotron 2. Transformer TTS [9] achieves a MOS value of 4.39 compared to 4.44 of human speakers and is thus on par with Tacotron 2.

Autoregressive models such as Tacotron 2 and Transformer TTS achieve state-of-the-art quality but can hardly be parallelized, leading to longer processing times. A few minutes of audio quickly take hours to generate [10]. Therefore, most of the research in 2019 and 2020 has focused on exploring architectures that are significantly faster and provide similarly good MOS values, rather than continuing to work on even better speech quality. Both Tacotron2 and TransformerTTS also incorporate certain attention mechanisms, which can lead to word omissions or even repetitions in outputs.

Non-autoregressive models can be further categorized into those using knowledge distillation like FastSpeech [10] and others utilizing differing technologies. Flow-TTS [11] and Glow-TTS [12] are examples for the latter. Interestingly, while many of the more recent publications presenting non-autoregressive models claim to be better than Tacotron 2 in a direct comparison, none of them were able to achieve comparably good MOS values close to the ground truth. Parallel Tacotron [13], Flow-TTS [11] and Fastpitch [14] are closest with MOS values above 4.0 and less than 0.5 worse than the ground truth.

B. Vocoder

Neural vocoders receive a mel spectrogram and predict audio signal frames for each spectrogram frame. A mel spectrogram can be generated directly from an audio file, as opposed to acoustic models, requiring audio-transcript-pairs. Therefore, it is comparably easy to generate training data, which results in a broad selection of well performing vocoders that can produce high quality audio hardly distinguishable from real human voices. The main reference is WaveNet [5], which achieved 4.21 on the MOS scale from 1 to 5 in the original publication [5] and 4.53 MOS in a later publication [1]. This is very close to the ground truth of 4.58 and still the state-of-the-art reference value up until now. Since WaveNet is autoregressive, it is both comparably slow and requires significant resources. To compensate these weak points, several alternatives have been suggested.

Parallel WaveNet [15] uses knowledge distillation to derive a much faster network from WaveNet in a student-teacher manner. It can generate 20s of audio in 1s (real-time factor RTF 0.05), whereas WaveNet requires 1,000s to generate 20s of audio (RTF 50).

WaveGlow [16] is a representative of flow-based networks, which can be parallelized well in contrast to auto-regressive networks like WaveNet. It achieves RTF 0.04 on an Nvidia Tesla V100 GPU. It is also commonly implemented as acoustic model, i.e., in [12], [17], [18].

Multi-Band MelGAN [2] is also worth mentioning, being based on a different approach. Its architecture utilizes a Generative Adversarial Network (GAN) and achieved a MOS of 4.34 in empirical analysis. However, this was achieved for the Chinese language instead of English and is therefore not directly comparable.

Best results based on the popular LJspeech dataset [19] are reported by Hifi-GAN [20] and WaveGrad [21] with 4.36 and 4.55 respectively. The latter is identical to the ground truth MOS value.

Finally, WaveRNN [22] achieves MOS 4.46 and is therefore the closest competitor to WaveNet and WaveGrad.

III. DATASETS

The selection of suitable datasets was based on metadata from LJSpeech. Strict criteria for the minimum length of audio-transcript pairs (>20 hours) and text normalization (no leftover digits or symbols) were set. The sampling rate of 22.05kHz was not considered to be a hard criterion, merely regarded preferable, so not to further reduce the scope of the already limited number of existing datasets.

Selected datasets were further processed in preparation of the subsequent training processes.

A. Selection

Besides the acoustic model and vocoder, the quality and quantity of the dataset used for training are the main factors influencing the quality of the resulting synthetic voice. The following datasets were evaluated regarding their suitability and partially selected for subsequent model training. The final selection of datasets is presented in TABLE **I**.

1) M-AILABS

The M-AILABS speech dataset is based on data from LibriVox [23], a platform providing free audio books by voluntary, mostly amateur speakers, and consists of five single speaker datasets. Their durations range from 19h to 68h of speech and respective texts. Despite a comparatively low sampling rate of 16kHz for each recording, two speakers, Karlsson (male, 40h) and Eva K (female, 29h) were chosen for model training. Ramona (female, 68h) was discarded due to her subjectively unpleasant voice.

2) Thorsten Voice

Specifically created for the creation of TTS applications, the Thorsten neutral dataset consists of more than 23 hours of audio-transcript pairs from a single male voice, recorded with a sampling rate of 22.05kHz [24]. It was first released in March 2021 and, to the authors' knowledge, has not been evaluated in any scientific publication yet.

3) HUI Audio Corpus

Similar to M-AILABS, the recently released HUI audio corpus [25] also consists of freely available audio data from LibriVox and transcripts from gutenberg.org [26], but provides a much larger quantity of audio-transcript pairs per speaker and a higher sampling rate of 22.05kHz. The speakers Bernd Ungerer (male, 97h) as well as Hokuspokus full (female, 43h) and Hokuspokus clean (female, 27h; subset of Hokuspokus full, containing less noise) were chosen for model training.

TABLE I. DATASETS USED FOR FURTHER PROCESSING.

Dataset	Speaker	Sampling Rate	Hours
HUI Audio Corpus	Bernd Ungerer (m)	22 kHz	97 h
	Hokuspokus clean (f)	22 kHz	27 h
	Hokuspokus full (f)	22 kHz	43 h
Thorsten neutral	Thorsten Müller (m)	22 kHz	23 h
M-AILABS	Eva K (f)	16 kHz	29 h
	Karlsson (m)	16 kHz	40 h

B. Further Processing

To reduce the range of phrases and punctuation marks acoustic models receive as input, transcript sentences of all datasets were filtered and adjusted using several mechanisms. Also, since phoneme-based models generally perform better than character-based models due to their unambiguousness in terms of pronunciation, transcript data was converted to this type of representation beforehand.

1) Text Modification

Since many punctuation symbols have very similar effects on emphasis in German, a subset was defined onto which all further symbols were mapped. This resulted in a subset consisting only of the characters [“.”, “,”, “?”, “!”], which significantly reduced of the vocabulary size.

Additionally, datasets based on LibriVox mostly consist of audio books of which the transcripts were written in the early 20th century and earlier, as German licensing rights require authors to have been deceased for at least 70 years, before copyright of their works expires. Transcripts of such ages were written according to obsolete orthographic standards, but the models to be trained were intended to be used in modern contexts. For this reason, a dictionary has been created semiautomatically (partly by crawling [27], a website providing common mappings between orthographic conventions, partly through manual identification of obsolete phrasing inside transcript sentences). Utilizing regular expressions, the outdated transcripts were adapted to currently applicable orthographic principles.

2) Phonemization

As no publicly available mapping tools or dictionaries seemed to be performing well enough for phonemization in German, a custom dictionary was created by crawling Wiktionary German [28], a website providing over 640,000 German word pairs with notations based on character as well as the International Phonetic Alphabet (IPA) including nouns in multiple grammatical cases and verbs in multiple tenses.

To convert composites which are not exactly contained within the phoneme dictionary into phoneme notation, a bidirectional search algorithm was implemented, which splits words into substrings if no exact match is found. The longest substrings found are individually converted to phoneme symbols and merged back together afterwards.

Since compounds and nominalizations by using different suffixes are widely used in the German language, a major proportion of the vocabulary can be covered by this approach.

While this algorithm handles borderline cases, names and words from other languages rather poorly, most German words as well as composites can be mapped to their respective phoneme representation quite efficiently. To reduce suboptimal mappings to a minimum, a large fraction of unknown words contained in the selected training datasets was added manually to the phoneme dictionary.

IV. MODEL TRAINING

The following subsections present and justify the final selection of model architectures for both stages of a full TTS system and describe all conducted training workflows on a detailed level. Both acoustic models and vocoders were trained independently.

A. Model Selection

Since a wide range of architectures exists for both acoustic models and vocoders, several test trainings were conducted to determine a viable composition.

Tacotron 2 and TransformerTTS were considered as acoustic models due to their excellent evaluations in literature as well as their inclusion into the ESPnet [29] framework, a toolkit for speech processing, offering simple mechanisms for building TTS training pipelines. First trainings showed that stop token prediction clearly performed better with Tacotron 2 than TransformerTTS, thus the final choice was made in favor of this architecture. AlignTTS was considered as well, but preexisting implementations were badly documented and training with reasonable effort was unfeasible.

For the vocoder stage, it was intended to test several architectures in sequence. However, Multi-Band MelGAN, as first architecture to be evaluated, already achieved subjectively satisfactory results in initial tests and was selected as the vocoder architecture for subsequent trainings. It was refrained from testing other vocoders, since subjectively, the quality of the acoustic model had a larger impact on overall output quality.

B. Tacotron 2

To optimize the training process, minor adjustments were made to the default hyperparameter configuration before the training process. In addition, the most suitable decoder configuration at inference time was determined through manual evaluation.

1) Training

The specific model architecture and training configuration for Tacotron 2 were derived from the existing recipe for LJSpeech incorporated in the ESPnet framework and adapted to fit the available hardware in terms of batch size (or number of batch bins, as implemented in ESPnet). This recipe differs from the original implementation of Tacotron 2 in the usage of guided attention loss. While training with datasets based on a sampling rate of 16kHz resulted in fast loss convergence, models trained on 22.05kHz audio data quickly reached a stage of oscillating loss. This was remedied by the use of AMSGrad [30]. All other parameters were maintained. In order to utilize ESPnet, the datasets used were converted into the Kaldi [31] format.

All models were trained for 200 epochs. Training processes took ~4 days each for 22.05kHz datasets and ~2 days each for 16kHz datasets respectively on two Nvidia Titan RTX GPUs.

2) Inferencing

The decoder configuration can be dynamically adjusted at inference time. In order to find the best possible configuration for all speakers, several suitable values were defined for each adjustable parameter and output audio was generated for each combination of parameters. Any of the variables may cause word repetition or deletion errors, if misconfigured.

The following parameters were determined:

- Minimum Length Ratio: 0.08
- Maximum Length Ratio: 10
- Backwards Attention Window: 2
- Forwards Attention Window: 3
- Stop-Token-Threshold: 0.1

While optimal values varied slightly between all speakers, the specified configuration generally yielded good results. This rendered the following model evaluations independent of speaker-specific decoder configurations.

C. Mutli-Band MelGAN

The implementation used was the publicly available version by Tomoki Hayashi [30] and the standard configuration was retained. Each model was trained according to this for 800,000 steps. Training took ~3 days per model using the same hardware as for the Tacotron 2 models.

For the speaker Hokuspokus no separate vocoder with the clean subset was trained, instead the vocoder from the full dataset was reused.

V. Empirical Analysis

The trained model compositions were evaluated through a survey, collecting MOS values for original speakers, full two-level inferences, and inferences of vocoders based on algorithmically generated mel spectrograms of original recordings. Additionally, the survey included further questions regarding the "best" fully synthetic voice, according to individual ratings of the respondents. Furthermore, demographic parameters, as well as audio output devices used during the survey were queried.

A. Questionnaire Design

The core components and structure of the survey are described in more detail in the following subsections.

1) MOS

Each respondent could listen to three audio files per voice, which were to be rated qualitatively on a scale of 1 to 9 without further instructions. No text labels for the individual numbers were provided on purpose, it was merely indicated that 9 meant very good and 1 very bad quality. Fully synthetic voices (acoustic model + vocoder), ground truths of all voices as well as vocoder-only inferences derived from mel spectrograms of ground truth data were evaluated in order to gain insights into the general performance of the model combinations as well as the sole influence of the trained vocoders on speech quality. The judgements of the mean opinion scores thus included 16 different voices and 48 audio recordings with 5-8 seconds length per recording.

During the evaluation, the rating scale was rescaled to the range 1 to 5 (in 0.5 increments) to enable direct comparison to the MOS values of other publications.

To avoid a bias regarding the order of the heard speakers, the sequence in which respondents were to rate them was randomized.

2) Detailed "Best" Speaker

After all MOS values had been filled in, the best-rated, fully synthetic voice was automatically determined, its corresponding recordings were played again, and more in-depth questions were asked regarding the characteristics of this voice.

- Did you notice any anomalies in pronunciation you found annoying? (*Very many* to *None*) (Q1)
- How would you describe the effort needed to understand the message? (*Nothing understood* to *Everything understood*) (Q2)
- How did you perceive the pace of speech? (*Too slow* to *Too fast*) (Q3)
- How did you perceive the naturality of the voice? (*Very unnatural* to *Very natural*) (Q4)
- Did you find certain words difficult to understand? (*Very many* to *None*) (Q5)
- How would you describe the voice? (*Very unpleasant* to *Very pleasant*) (Q6)
- Would you find it easy or difficult to listen to this speaker for an extended period of time? (*Very easy* to *Very difficult*) (Q7)

These questions were intended to provide insight into which aspects of the synthetic voices were subjectively perceived as suboptimal. The selection of questions was based on [32]. Posterior characters represent references to the questions in TABLE **V**.

3) Demographic Data

To derive further conclusions from previously collected scores, participants were additionally asked regarding their native language and age. As described in CrowdMOS [33], the audio device used while answering the survey was also asked for.

B. MOS Results

The survey was conducted over the internet. Invitations were sent to students from the University of Applied Sciences Hof, the research institute employees, as well as to a network of company partners. It was also circulated on the internet via Twitter and Linked.in.

A total of 193 participants was recorded of which 101 finished the survey. 94 of this subset were native German speakers. Answers and ratings of those were used for further analysis. Around half of the leftover respondents used a smartphone or PC with built-in speakers. 34 were using headphones, 11 dedicated loudspeakers. The age of participants was 30.1 years on average with a median of 26 and a range from 18 to 74 years.

TABLE **III** summarizes the results. Synth represents the MOS for the synthetic voice, created using both the trained acoustic model and vocoder. Vocoder represents the MOS for the synthetic voice that was generated based on the mel spectrograms derived from the ground truth. GT represents the MOS for the human speaker used as training data. Δ GT is the difference between the MOS of the ground truth and the MOS of the synthetic voice. TABLE **IV** puts the results and training datasets in relation to each other.

C. *Speaker-Specific Analysis*

The more detailed, speaker-specific analysis shown in TABLE **V** presents an overview of the advanced evaluation, including certain characteristics of speech, which primarily revealed a persistent deficit of naturalness in the voices, where no synthetic voice reached an average score over 4.0. This is supported by comparable scores for anomalies in pronunciation and how pleasant the voice is perceived. Comprehensibility of individual words was rated slightly better. Pace of speech and effort required to understand the message of utterances were rated very positively. Ultimately, scores for the difficulty of listening to a speaker over an extended period of time were consistently mediocre. Bernd Ungerer especially stood out regarding naturalness of the synthetic voice, whereas there was no large difference to other voices regarding anomalies in pronunciation and ease of understanding compared to Thorsten. The pace of speech was also similar.

VI. Discussion

The empirical survey affirmed the preexisting subjective impression that the fully synthetic TTS system trained on data from the speaker Bernd Ungerer produced the best results among all evaluated model compositions. However, the overall scores were lower than expected. This is partly due to a large variation in answers with participants voting 2.3 on average for all 16 voices and others voting 4.6 (avg: 3.55, median: 3.62). With 94 qualified answers, the empirical survey is much larger than the ones in other TTS papers that frequently use less than two dozen participants.

Interestingly, the speaker Thorsten Müller achieved best results for vocoder only and a similar distance between synthetic voice and ground truth as Bernd Ungerer, despite having only a quarter of the training data. This indicates that data quality is at least equally important, if not more important than total size of the dataset. The same conclusion can be drawn from the results of Hokuspokus clean and full. Although the clean subset contains only 27 hours of voice data, the MOS results are slightly better than those of networks trained on the full 43 hours of data available. Which amount of (qualitatively high) training data would actually be needed for a well performing acoustic model remains to be determined. Matsubara et al. [34] found that as few as one hour of training data is sufficient for achieving MOS values of 3.8 with LPCnet and 9 hours for MOS values of 4.06 with WaveNet, with a ground truth of only 4.18. However, this could not be reproduced using Tacotron 2 and Multi-Band MelGAN, which may be caused by the chosen model composition. Stop token prediction proved problematic, which resulted in additional babbling sounds as part of the generated audio files. This mainly occurred with models trained on less than 20 hours of audio-transcription pairs.

The ground truth values of 4.25 and 4.27 for speakers Bernd Ungerer and Hokuspokus (both full and clean) are similar to the values reported in literature for English language, e.g., 4.27 for FastSpeech 2 [35] and 4.31 for TalkNet [36]. However, they are significantly worse than the 4.58 reported in the Tacotron 2 paper [1] or 4.55 for Flow-TTS [11]. This indicates that there is still potential for improvement since neither Bernd Ungerer nor Hokuspokus are professional speakers. Accordingly, the recordings were not professionally produced and processed, which in consequence lead to inconsistent narration styles and noise. A delta of 0.5 between ground truth MOS and synthetic voice (Bernd Ungerer, Thorsten Müller) is only topped by very few of the well-known English TTS results published. It can therefore be concluded that the chosen model architectures can generally be equally well trained on datasets in German as in the English language (or Chinese for Multi-Band MelGAN).

Vocoder MOS values are significantly lower than expected for all speakers except Thorsten Müller. A delta of 0.22 for Thorsten Müller is among the best in published English results. However, for Bernd Ungerer (0.50) and Hokuspokus (0.66), values are worse than the average published in English publications concerning TTS, which is around 0.35. For Multi-Band MelGAN, the published results are 4.22, which is 0.36 worse than the ground truth on the MOS scale. However, these results were gathered in Chinese. Switching the vocoder should be investigated for future experiments.

Differently than suggested in literature [37], the female voices are not judged better than the male voices, but worse. This is especially unexpected for the direct comparison of Hokuspokus clean with Thorsten Müller. Hokuspokus has a better GT score and slightly more training data in the clean dataset (27h vs. 23h). Therefore, a better MOS value for Hokuspokus than for Thorsten was expected. There are two major differences between the datasets. The Thorsten neutral set consists of one (short) sentence per audio sample having an average duration of 3.3s with a maximum of 12s and only few audio files with more than 5s (see Figure **1**), whereas audio-transcripts from Hokuspokus (and other sets from the HUI audio corpus) were split based on duration with a minimum length of 5s and an average of 9s with some audio files at over 20s, regardless of sentence cohesion.

Utterances in the Thorsten neutral dataset are continuously very clearly emphasized as it was specifically generated for the creation of TTS systems, while recordings by Hokuspokus do not contain any special emphasis, sounding generally more natural (which possibly led to comparably higher MOS values for ground truth). However, this aspect seems to render the Hokuspokus datasets less suitable for speech synthesis applications. Additionally, the average silence loudness in dB is slightly lower in the Thorsten neutral dataset (-58.3 dB) compared to Hokuspokus clean (-56.6 dB, see Figure **2**), indicating less noise. It would be interesting to see, whether a further cleansing of Thorsten speech samples yield better training results. Due to the generally low amount of training data contained in the Thorsten neutral dataset, no further investigation was conducted.

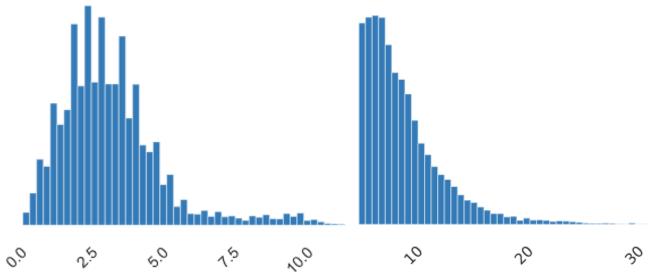


Figure 1. Thorsten (l) and Hokuspokus (r) length of audio in seconds.

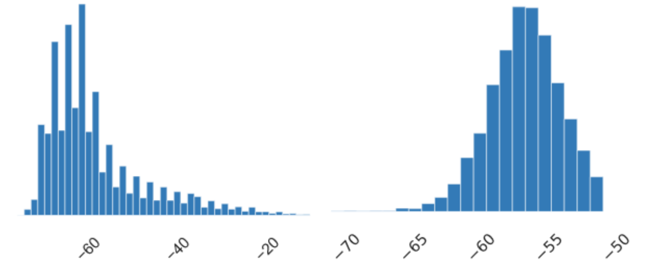


Figure 2. Thorsten (l) and Hokuspokus (r) min. silence in dB.

It is also surprising, that the speaker Karlsson achieved a comparably high MOS for the ground truth despite being based on a sampling rate of 16 kHz. Also, the vocoder MOS is among the best with 3.76, whereas the fully synthetic voice merely achieved 2.96 (-0.8 compared to the vocoder).

Moreover, it is remarkable that the loss in MOS from GT to vocoder and full synthetic voice is split relatively equally for Thorsten Müller and Hokuspokus, whereas there is nearly no loss for the acoustic model for Bernd. In contrast to that, Karlsson and Eva have most of the loss in acoustic model and a much smaller one for the vocoder. Looking at published results for LJspeech, examples of both described discrepancies can be found. For an equal split, there is AlignTTS, FlowTTS und TalkNet with WaveGlow vocoder, as well as TalkNet 2 with Hifi-GAN vocoder. A larger loss for the vocoder can be observed for Glow-TTS and Fastspeech with WaveGlow vocoder, as well as Reinforce-Aligner and Diff-TTS with Hifi-GAN vocoder. Finally, EFTS-CNN with Hifi-GAN has a higher loss in the acoustic model than the vocoder. Therefore, it could be a matter of tuning the hyperparameters for the training process that makes a difference, but it could also be characteristics of the dataset in this case. It is assumed that the acoustic model benefits more from large amounts of training data, whereas the vocoder benefits more from a high audio quality.

Furthermore, speaker-specific analysis confirmed that basic conditions for natural speech, such as pace and correct as well as clear pronunciation of individual words, are generally met. However, fully synthetic outputs still contain too many irregularities, which reduces the acceptance of users to listen over longer periods of time. Additionally, none of the recordings contained in the training datasets were made by a professional speaker, which is reflected in the mediocre scores on how pleasant the different voices were perceived.

VII. LIMITATIONS

Audio files, which were used for the empirical analysis were specifically chosen to be comparable across all speakers as well as comparable with the ground truth. Although sentences that proved to be difficult during the training process were included, they are still somehow cherry-picked. When generating speech from arbitrary texts from news websites, some problems with the synthesized voices were encountered that are not reflected in the test audio. Negative examples can be found on the webpage, presenting results ([38]).

Although these cases are seldom, the quality of the generated speech output still needs to be double-checked, since Tacotron 2 performs in a non-deterministic way, which is intended in order to vary stylistic attributes in output mel spectrograms. However, this feature sometimes leads to very bad output quality.

Additionally, the choice of vocoder and acoustic model are somewhat arbitrary. Although there was a systematic analysis of available models, no detailed evaluation with multiple candidates was performed. Instead, the first models subjectively producing good results were used for the empirical study. Finally, the vocoder should have been trained with the Hokuspokus clean subset as well, instead of reusing the one from the full subset in order to explore the full potential of data cleansing.

VIII. CONCLUSION AND OUTLOOK

In this work, the training processes of several deep neural networks for speech synthesis in the German language was reported along with an evaluation based on the MOS. A MOS of 3.74 was achieved for the best rated model (using the speaker Bernd Ungerer), which is comparable to recently published results for speech synthesis systems in English like 3.79 for FastSpeech 2 [35] or 3.66 for Flowtron [18].

However, they are far away from the best published results like 4.53 for Tactron 2 with Wavenet [1] or 4.19 for Flow-TTS with WaveGlow [11]. On the other hand, Tacotron 2 also achieves only 3.52 on the MOS scale in the Flow-TTS paper. To the best of the authors' knowledge, results are the best published MOS results for German TTS and can serve as a benchmark for future publications. In the years before neural TTS systems, MaryTTS has been a well-known option for German [39] and multi-lingual speech synthesis [40]. However, even in explicit quality analysis [41], no MOS values are reported.

In addition, deeper insights were gained regarding distinct aspects of different synthetic voices, which suggest actions regarding further optimization of future models. At dataset level, alignment of audio transcript pairs, recording quality and its homogeneity, as well as prosody can be improved. Regarding the definition of hyperparameters, values were set based on comparisons. A thorough hyperparameter search could lead to better results. In addition, the phoneme dictionary needs to be extended to include a larger number of terms in order to cover as many words as possible.

All compared models and respective recipes for ESPnet are released for public use.

For further research, it is intended to continue experimenting with internal voice datasets of higher quality but smaller size, as well as different network architectures. Especially for the vocoder, a broader range of alternatives to Multi-Band MelGAN will be considered, including Hifi-GAN [20], WaveGrad [21] and Wave RNN [22], which all have published results well over 4.3 MOS in English language and differences to ground truth below 0.1.

Additionally, it needs to be investigated which aspects of the training data differentiate a very good from an average dataset. A few aspects like good recording conditions and trained speaker are well known. However, there is little information regarding speaking style, choice of sentences and words, diversity of the vocabulary, etc. Those aspects are expected to influence dataset quality. Moreover, the preexisting processing pipeline for the generation of datasets from [25] will be altered to shorten the minimum and maximum duration of audio snippets contained in training data to a scope 2s minimum, 6s mean and 15s maximum.

Curriculum learning [42] represents another promising method, which would be worth investigating in the context of TTS. It is dangerous to draw conclusions from humans to DNNs. Despite some similarities, DNNs still work different from human brains. Nevertheless, human children usually learn to speak short utterances first, as opposed to words like "Frühsommer-Meningoenzephalitis" (FSME), a complex German word from the medical domain, which is part of an internal test dataset. Therefore, it could be also helpful to run trainings of model architectures with audio-transcription pairs of short sentences or even single words and gradually increase the length of labeled audio files. There is already evidence that this method increases robustness of TTS models for longer input texts during inference [43]. It could potentially also improve loss convergence during training as well as output speech quality.

The findings presented in this work will be incorporated into the development of an independent smart speaker, whereby the performance of TTS systems on edge devices, primarily resource requirements and RTF, will be a major challenge.

References

[1] J. Shen *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[2] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band MelGAN: Faster Waveform Generation for High-Quality Text-to-Speech," *arXiv preprint arXiv:2005.05106*, 2020.

[3] S. Kayte, M. Mundada, and J. Gujrathi, "Hidden Markov model based speech synthesis: A review," *International Journal of Computer Applications*, vol. 130, no. 3, pp. 35–39, 2015.

[4] Y. Wang *et al.*, "Tacotron: Towards End-to-End Speech Synthesis," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

[5] A. van den Oord *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[6] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, no. 01, pp. 6706–6713.

[7] A. Vaswani *et al.*, "Attention Is All You Need," *arXiv:1706.03762 [cs]*, Dec. 2017, Accessed: Sep. 01, 2021. [Online]. Available: http://arxiv.org/abs/1706.03762

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Neural Speech Synthesis with Transformer Network," 2019.

[10] Y. Ren *et al.*, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems*, 2019, pp. 3171–3180.

[11] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, "Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7209–7213.

[12] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," *arXiv preprint arXiv:2005.11129*, 2020.

[13] I. Elias *et al.*, "Parallel Tacotron: Non-Autoregressive and Controllable TTS," *arXiv preprint arXiv:2010.11439*, 2020.

[14] A. Łańcucki, "FastPitch: Parallel Text-to-speech with Pitch Prediction," *arXiv preprint arXiv:2006.06873*, 2020.

[15] A. Oord *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International conference on machine learning*, 2018, pp. 3918–3926.

[16] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3617–3621.

ISBN: 978-1-61208-896-9

[17] Z. Zeng, J. Wang, N. Cheng, T. Xia, and J. Xiao, “Aligntts: Efficient feed-forward text-to-speech system without explicit alignment,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6714–6718.

[18] R. Valle, K. Shih, R. Prenger, and B. Catanzaro, “Flowtron: an Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis,” *arXiv preprint arXiv:2005.05957*, 2020.

[19] “The LJ Speech Dataset.” https://keithito.com/LJ-Speech-Dataset (accessed Sep. 01, 2021).

[20] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” *Advances in Neural Information Processing Systems*, vol. 33, n. pag., 2020.

[21] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, “WaveGrad: Estimating gradients for waveform generation,” *arXiv preprint arXiv:2009.00713*, 2020.

[22] N. Kalchbrenner *et al.*, “Efficient neural audio synthesis,” *arXiv preprint arXiv:1802.08435*, 2018.

[23] “LibriVox | free public domain audiobooks.” https://librivox.org/ (accessed Sep. 01, 2021).

[24] T. Müller, “Thorsten Open German Voice Dataset”. https://github.com/thorstenMueller/deep-learning-german-tts (accessed Sep. 01, 2021).

[25] P. Puchtler, J. Wirth, and R. Peinl, “HUI-Audio-Corpus-German: A high quality TTS dataset,” Berlin, Germany, Sep. 2021.

[26] “Projekt Gutenberg”. https://www.projekt-gutenberg.org/ (accessed Sep. 01, 2021).

[27] J. von Heyl, “korrekturen.de - Portal für Rechtschreibung”. https://www.korrekturen.de/ (accessed Sep. 01, 2021).

[28] “Wiktionary, das freie Wörterbuch”. https://de.wiktionary.org/wiki/Wiktionary:Hauptseite (accessed Sep. 01, 2021).

[29] T. Hayashi *et al.*, “Espnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 7654–7658. doi: 10.1109/ICASSP40776.2020.9053512.

[30] S. J. Reddi, S. Kale, and S. Kumar, “On the Convergence of Adam and Beyond,” *arXiv:1904.09237 [cs, math, stat]*, Apr. 2019, Accessed: Sep. 01, 2021. [Online]. Available: http://arxiv.org/abs/1904.09237

[31] D. Povey *et al.*, “The Kaldi speech recognition toolkit,” *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Jan. 2011.

[32] M. Viswanathan and M. Viswanathan, “Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale,” *Computer Speech & Language*, vol. 19, no. 1, pp. 55–83, 2005, doi: https://doi.org/10.1016/j.csl.2003.12.001.

[33] F. Protasio Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, “CROWDMOS: An Approach for Crowdsourcing Mean Opinion Score Studies,” May 2011, Accessed: Sep. 01, 2021, ICASSP. [Online]. Available: https://www.microsoft.com/en-us/research/publication/crowdmos-an-approach-for-crowdsourcing-mean-opinion-score-studies/

[34] K. Matsubara *et al.*, “Investigation of training data size for real-time neural vocoders on CPUs,” *Acoustical Science and Technology*, vol. 42, no. 1, pp. 65–68, 2021.

[35] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “FastSpeech 2: Fast and High-Quality End-to-End Text-to-Speech,” *arXiv preprint arXiv:2006.04558*, 2020.

[36] S. Beliaev, Y. Rebryk, and B. Ginsburg, “TalkNet: Fully-Convolutional Non-Autoregressive Speech Synthesis Model,” *arXiv preprint arXiv:2005.05514*, 2020.

[37] J. Cambre and C. Kulkarni, “One voice fits all? Social implications and research challenges of designing voices for smart devices,” *Proceedings of the ACM on human-computer interaction*, vol. 3, no. CSCW, pp. 1–19, 2019.

[38] J. Wirth, “iisys Audio Samples for German Speech Synthesis Tacotron 2 + MultiBand MelGAN”. http://narvi.sysint.iisys.de/projects/tts/results (accessed Sep. 01, 2021).

[39] M. Schröder and J. Trouvain, “The German text-to-speech synthesis system MARY: A tool for research, development and teaching,” *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[40] M. Schröder, M. Charfuelan, S. Pammi, and I. Steiner, “Open source voice creation toolkit for the MARY TTS Platform,” presented at the Twelfth annual conference of the international speech communication association, 2011.

[41] F. Hinterleitner, C. Norrenbrock, S. Möller, and U. Heute, “What makes this voice sound so bad? A multidimensional analysis of state-of-the-art text-to-speech systems,” in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 240–245.

[42] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[43] S.-W. Hwang and J.-H. Chang, “Document-Level Neural TTS Using Curriculum Learning and Attention Masking,” *IEEE Access*, vol. 9, pp. 8954–8960, 2021.

TABLE II. MOS OVERVIEW OF COMPARABLE TTS SYSTEMS.

Model	**Vocoder**	**GT**	**Vocoder**	**Synth**	**GT-MOS**	**GT-voc**	**Voc-synth**
Fastspeech	WaveGlow	4.41	4.00	3.84	0.57	0.41	0.16
AlignTTS	WaveGlow	4.53	4.28	4.05	0.48	0.25	0.23
Glow-TTS	WaveGlow	4.54	4.19	4.01	0.53	0.35	0.18
Flow-TTS	WaveGlow	4.55	4.35	4.19	0.36	0.20	0.16
TalkNet	WaveGlow	4.31	4.04	3.74	0.57	0.27	0.3
TalkNet 2	Hifi-GAN	4.32	4.2	4.08	0.24	0.12	0.12

TABLE III. MOS COMPARISON OF ALL TRAINED SPEAKERS.

Dataset	**Speaker**	**Synth**	**Δ GT**	**Vocoder**	**GT**
HUI Audio Corpus	Bernd Ungerer	3.74	0.51	3.75	4.25
	Hokuspokus clean	2.98	1.29	x	4.27
	Hokuspokus full	2.88	1.39	3.60	4.27
Thorsten neutral	Thorsten Müller	3.49	0.50	3.78	3.99
M-AILABS	Eva K	2.13	1.60	3.33	3.72
	Karlsson	2.96	1.18	3.76	4.14

TABLE IV. OVERVIEW OF DATASETS USED FOR MODEL TRAINING AND CORRESPONDING MOS EVALUATIONS.

Speaker	**GT**	**Δ GT-synth**	**Δ GT-Vocoder**	**Δ Vocoder-synth**	**Amount of data (hours)**	**Training Loss (Acoustic Model)**	**Sampling Rate**
Bernd Ungerer	4.25	0.51	0.50	0.01	97	0.52	22.05 kHz
Thorsten Müller	3.99	0.50	0.22	0.28	23	0.48	22.05 kHz
Hokuspokus Clean	4.27	1.29	0.66	0.62	43	0.44	22.05 kHz
Hokuspokus Full	4.27	1.39	0.66	0.72	27	0.46	22.05 kHz
Karlsson	4.14	1.18	0.38	0.80	40	0.43	16 kHz
Eva K.	3.72	1.60	0.39	1.20	29	0.56	16 kHz

TABLE V. SPEAKER-SPECIFIC ANALYSIS (OPTIMAL SCORES IN BRACKETS).

Speaker	**Votes**	**Q1 (5.0)**	**Q2 (5.0)**	**Q3 (0.0)**	**Q4 (5.0)**	**Q5 (5.0)**	**Q6 (5.0)**	**Q7 (5.0)**
Bernd Ungerer	54	3.6	4.4	-0.2	4.0	4.1	3.9	3.5
Thorsten Müller	14	3.7	4.3	-0.2	3.1	4.0	3.5	3.0
Hokuspokus Clean	3	3.2	4.2	±0	3.2	4.2	3.3	3.5
Hokuspokus Full	23	3.0	4.1	-0.3	3.3	3.6	3.6	3.0

- Did you notice any anomalies in pronunciation you found annoying? (*Very many* to *None*) (Q1)
- How would you describe the effort needed to understand the message? (*Nothing understood* to *Everything understood*) (Q2)
- How did you perceive the pace of speech? (*Too slow* to *Too fast*) (Q3)
- How did you perceive the naturality of the voice? (*Very unnatural* to *Very natural*) (Q4)
- Did you find certain words difficult to understand? (*Very many* to *None*) (Q5)
- How would you describe the voice? (*Very unpleasant* to *very pleasant*) (Q6)
- Would you find it easy or difficult to listen to this speaker for an extended period of time? (*Very easy* to *Very difficult*) (Q7)

Automatic Emotions Analysis for French Email Campaigns Optimization

Alexis Blandin
IRISA, EXPRESSION,
Université Bretagne-Sud
UNEEK - Kosmopolead
France
email: alexis.blandin@univ-ubs.fr

Farida Said
IRISA, EXPRESSION,
LMBA
Université Bretagne-Sud
France
email: farida.said@univ-ubs.fr

Jeanne Villaneau, Pierre-François Marteau
IRISA, EXPRESSION,
Université Bretagne-Sud
France
email: jeanne.villaneau@univ-ubs.fr,
email: pierre-francois.marteau@univ-ubs.fr

***Abstract*—Email communication and newsletter campaigns remain a significant concern for companies. The main question addressed here is how to optimize the form and content of a newsletter so that it is not interpreted as spam or annoyance by the recipient. We address this question by analyzing the emotions and opinions conveyed by emails and evaluating how they affect their open and click rate performance. We first describe a new dataset of French newsletters, and then we use emotional embeddings to analyze the associations between emotions and email performance. We finally derive clues on how to write effective email campaigns.**

***Keywords*—Algorithm; Artificial intelligence; Sentiment Analysis; Emotion prediction; Emotion recognition; Email campaign**

I. Introduction

Artificial intelligence is developing in many areas and is increasingly used to determine and optimize business and marketing strategies. In particular, Natural Language Processing (NLP) techniques are widely used for the automatic analysis of human interactions, and we exploit them to optimize email communication by analyzing the content of newsletters.

More precisely, we focus on how emotions and opinions conveyed in an emailing campaign can influence its performance. To address this question, we first built a dataset of more than 900 French newsletter campaigns provided by various companies or associations.

We first proposed vector representations of newsletters that reflect emotion and sentiment using NLP techniques. We then statistically analyzed the relationships between the emotions and opinions conveyed by the newsletters and their performance indicators, i.e., click and open rates. Finally, we used the proposed vectorizations to evaluate the prediction of a newsletter's performance based on the emotions and opinions in its text.

II. Related work

A. Marketing studies

We first review some hypotheses formulated in marketing science regarding email marketing optimization and their potential links to emotions. A study proposed in 2008 by K. Byron [1] suggests that the lack of face-to-face interaction due to email communication can lead to misinterpretation of emotions.

According to the author, the lack of cues that allow the recipient to determine the intended emotions generally leads to a *neutrality effect*. The design of the email may even increase the likelihood that the recipient will perceive the email negatively, resulting in a *negativity effect*. The author argues that when the email contains few cues about emotions, the ambiguity of the emotional tone increases the salience of all negative information. For instance, sarcasm may be perceived more negatively than in face-to-face interaction because of the lack of context and tone ambiguity. The study also highlights the importance of the social context of email communication and the socio-demographic characteristics of the sender and recipient (gender, age, relative status in the company, Etc.) in interpreting emotions. Although the author points out some positive consequences of the negativity effect, such as "*using less niceties or not "sugarcoating" the message*," it should be noted that the negativity effect can be problematic in the context of marketing communication in which it is crucial to elicit positive emotions such as pride [2] in order to expect better actions from the customer, especially in western culture.

Furthermore, when the sender and recipient do not know each other, there is even less contextual information to help the recipient interpret the emotions correctly. Thus, there is an increased risk of misinterpreting the emotions conveyed by the email. On the other hand, in a recent study conducted in the French context of the COVID-19 pandemic [3], commercial communication by email seems to evolve from purely informative content to more entertaining and emotional content. Therefore, it is becoming crucial for companies that the recipients do not misinterpret the emotions contained in their emails.

B. Email content analysis

As mentioned before, we want to evaluate the impact of textual content and email subject lines on the performance of email campaigns by analyzing the emotions and opinions they convey. This approach was used in [4] with email subject lines, and their findings validated some hypotheses on how emotions influence email perception. However, these results are difficult to transfer to our context for several reasons, including language. Indeed, the authors investigated the Enron dataset [5], a large set of emails in English from 150 employees, mainly executives, of the Enron company. This type of resource does not seem to exist in French, and many non-English speaking studies have to build their datasets specifically for their tasks [6], [7].

Another major difference with the cited study is that we consider both the subject and the textual content of the email.

C. Emotion Detection

In recent years, emotion detection in text has become increasingly popular due to its wide range of applications. It can be viewed as an extension to a more diverse emotional spectrum of research on sentiment analysis which focuses on positive and negative emotions. While many studies have proposed their own approaches [8], one of the most common is to use word lexicons labeled with categories of emotions. These categories are often the six basic emotions proposed by P. Ekman [9]: joy, fear, disgust, sadness, anger, and surprise.

This type of resource exists in French. A. Abadoui et al. proposed the FEEL lexicon [10], composed of French words or expressions represented by zero-one vectors of size 7. Six entries indicate whether the word carries one of the six basic emotions, and one represents the polarity associated with the word. This lexicon was constructed automatically, from the English NRC-Emolex lexicon, by crossing the results of several automatic translators. A professional translator subsequently enriched the lexicon and validated the results. The final lexicon consists of 14,127 different lemmatized terms, including 11,979 simple words and 2,148 compound words. Each lemmatized form gets the emotions contained in all of its inflected forms.

III. Dataset presentation

Our dataset is composed of newsletters from various organizations such as companies and associations. These organizations use the same customer relationship management (CRM) system and design their emailing campaigns using the same framework. The main objective of these organizations is to inform their subscribers about events or new opportunities. Our dataset does not include campaigns that target purchase actions such as online shopping.

A newsletter's performance can be measured by tracking the included links, which provide the number of unique opens and the number of unique clicks generated by the reader for each newsletter. These are good indicators of the performance of an email campaign and are commonly used in email analysis [11], [12]. One can view the open rate as a measure of the email's attractiveness and the click rate as the engagement generated by the newsletter.

After cleaning up the data provided by the CRM servers and removing test emails and duplicates, we ended up with 973 newsletters, each sent to multiple subscribers, with their performance information such as click rates and open rates. The number of emails per customer is not balanced, as illustrated in Figure 1. While this could represent a bias, we assume it does not impact our analysis. Indeed, we focus on features that can be considered independent of the email's author.

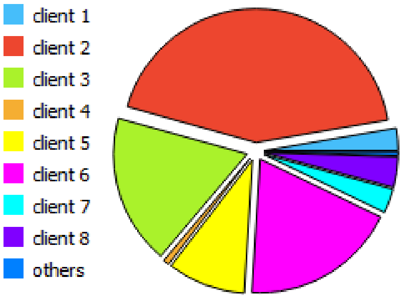


Fig. 1. Distribution of newsletters per client

IV. Data analysis

A. Features extraction

To process the data, we collected descriptive, emotional and sentiment information about the newsletters. The descriptive features were obtained directly from the data host and consist of email subject line length, size of the *.eml* file sent, and unique open and click rates.

We used standard NLP techniques to assess the emotion and sentiment features of the newsletters. First, we segmented the textual content of each newsletter into sentences from which we extracted all words, excluding the French stop words.

For emotion analysis, we assigned to each word an emotion vector according to the FEEL lexicon [10]. If a word is an inflected form, we consider the vector associated with the lemma as the aggregation of all emotions contained in its inflected forms. Then, we computed the emotion vector of each sentence as the average of the emotion vectors of its constituent words. These vectors, constructed from the FEEL lexicon, represent the six basic emotions described by P. Ekman

For sentiment analysis, we evaluated the subjectivity and polarity of the newsletters using the free NLP tool *Python TextBlob for Natural Language Processing*. The textblob library, detailled by Klein and Loper [13], uses a built-in model to compute subjectivity and polarity scores of sentences. A subjectivity value close to 0 indicates objective text, while a value close to 1 indicates highly subjective text. Polarity values range from -1 to 1 and reflect the negativity or positivity of a sentence, respectively.

At this point, each sentence is represented by a vector with eight entries: six emotion scores from the FEEL lexicon and two scores from the textblob analysis. For the full content of the email, we aggregated the information from all sentences by taking the average of all sentence vectors.

In addition to analyzing the newsletter content, we were interested in the emotions conveyed by the email subject line, its polarity, and subjectivity. We, therefore, performed the same NLP processing as for the textual content of the email by considering the subject line as a single sentence. We observed that the emotion scores were almost all null, which led us to consider only the polarity and subjectivity.

In the end, we represent the newsletters by vectors with ten entries that express the emotions and opinions conveyed by the content and topic of the email. Other studies [11] have combined some of the features we consider with non-emotional features, and we question whether emotion and sentiment features are as discriminative as non-emotional features in predicting performance.

B. Statistical results

We explored in our dataset the relationships between the emotion and sentiment features of the newsletters and their performance indicators, namely open and click rates.

TABLE I
PEARSON CORRELATIONS BETWEEN THE CHARACTERISTICS OF THE NEWSLETTERS AND THEIR PERFORMANCE INDICATORS

Features	**Open rate**	**Click rate**
File size (FS)	**-0.14*****	**0.25*****
Subject line length (SL)	**-0.13*****	**0.18*****
Subject line polarity (SP)	-0.07**	-0.03^{ns}
Subject line subjectivity (SS)	-0.01^{ns}	-0.07*
Content Polarity (CP)	-	0.09**
Content Subjectivity (CS)	-	-0.07*
Content Joy (J)	-	-0.10**
Content Fear (F)	-	**-0.11*****
Content Sadness (S)	-	**-0.23*****
Content Anger (A)	-	0.06^{ns}
Content Surprise (Su)	-	**-0.11*****
Content Disgust (D)	-	-0.07*

*p-value $<$ .05, **p-value $<$ .01, ***p-value $<$ .001,^{ns} not significant

The results are presented in Table I in terms of Pearson correlation. It appears that classical descriptors such as subject line length or file size significantly correlate with performance. Indeed, longer subject lines or heavier emails are associated with fewer opens but more clicks if the email is opened.

More interestingly, all emotions conveyed by the email content are negatively correlated with the click rate, regardless of the type of emotion. Sadness is the emotion most negatively associated with the click rate: the more sad the content of the email, the fewer clicks are measured. On the other hand, Table II sheds light on the relationships between the features of the newsletters. One can see that polarity and subjectivity are positively associated both in the text's content and in the email's subject line. Content polarity is positively associated with all emotions except fear and disgust. The significance of the correlations is even greater between content subjectivity and emotions except for disgust. Finally, subject line subjectivity is positively associated with all emotions except disgust, while its polarity is only associated with joy.

On the other hand, emotions are, for the most part, positively associated with each other. If we focus on the highly significant correlations, in bold in the table, we can see that surprise is positively correlated with all emotions except joy and that disgust is associated with rather negative emotions (fear, sadness, anger, and surprise). It also appears that fear and sadness are particular emotions by their strong association and their high correlation with all the emotions.

TABLE II
PEARSON CORRELATIONS BETWEEN THE FEATURES OF THE NEWSLETTERS

	SP	SS	CP	CS	J	F	Sa	A	Su	D
SP	1	**0.49*****	0.06^{ns}	-0.02^{ns}	**0.14*****	0.02^{ns}	0.06^{ns}	-0.03^{ns}	0.04^{ns}	0.02^{ns}
SS	-	1	0.04*	0.07^{ns}	**0.12*****	**0.14*****	**0.11*****	-0.02^{ns}	**0.15*****	0.07*
CP	-	-	1	**0.4*****	0.1**	0.02^{ns}	**0.12*****	0.1**	**0.2*****	0.02^{ns}
CS	-	-	-	1	0.1**	**0.14*****	**0.21*****	**0.12*****	**0.2*****	-0.03^{ns}
J	-	-	-	-	1	**0.19*****	**0.14*****	0.01^{ns}	0.04^{ns}	0.07*
F	-	-	-	-	-	1	**0.61*****	**0.37*****	**0.32*****	**0.37*****
Sa	-	-	-	-	-	-	1	**0.21*****	**0.25*****	**0.34*****
A	-	-	-	-	-	-	-	1	0.08**	**0.27*****
Su	-	-	-	-	-	-	-	-	1	**0.17*****
D	-	-	-	-	-	-	-	-	-	1

*p-value $<$ .05, **p-value $<$ .01, ***p-value $<$ .001,^{ns} not significant
SP: subject line polarity, SS: subject line subjectivity, CP: content polarity, CS: content subjectivity
J: joy, F: fear, Sa: sadness, A: anger, SU: surprise, D: Disgust

Emotion and sentiment features may not be the best predictors of newsletter performance, but we propose to evaluate their effectiveness in predicting click rate in the following.

V. UNSUPERVISED CLUSTERING

A. Multidimensional representation

We first explored our data graphically to see if there is a global structure that we could exploit. Since the newsletters are represented in a 10-dimensional space, we used a dimensionality reduction technique, namely the t-SNE (t-distributed Stochastic Neighbor Embedding). This method is mainly used to project high-dimensional data into low-dimensional spaces (2D or 3D) while preserving local distances between data points.

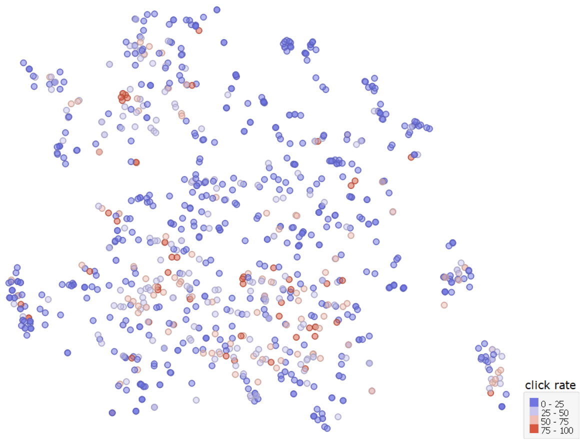


Fig. 2. t-SNE projection of our dataset

Figure 2 gives a visualization of our data in a 2D map. The color associated with the newsletters ranges from blue for "low-performing or bad" newsletters to red for "high-performing or good" newsletters. There is no clear separation between "good" and "bad" newsletters. However, it appears that the "good" newsletters are more grouped while the "bad" ones are more scattered across the map.

We investigate this hypothesis in the next section, using unsupervised clustering.

B. K-Means approach

K-means clustering is a vector quantization method that aims at partitioning a dataset into k clusters by assigning each observation to the cluster with the closest center (or centroid).

The effectiveness of K-means depends highly on the chosen number of clusters, and we usually do not have prior knowledge of the number of clusters that correspond to the most relevant clustering. A measure of the effectiveness of a cluster is the silhouette coefficient as introduced by Kaufman and Rousseeuw [14]. It measures how similar an object is to its cluster (cohesion) compared to other clusters (separation). The value of the silhouette ranges between -1 and 1, where a high value indicates that the object is well matched to its cluster and poorly matched to neighboring clusters. We compute the silhouette coefficients of all points and average them to obtain a global silhouette score. The clustering configuration with the best global silhouette score is the most relevant.

For a clustering in k clusters, the cohesion of a data point i assigned to a cluster I_k is defined as:

$$a(i) = \frac{1}{|I_k| - 1} \sum_{j \in I_k, j \neq i} d(x^i, x^j) \tag{1}$$

where $d(x^i, x^j)$ stands for the distance between the representative vectors x^i and x^j. We chose the cosine distance based on the angle between vectors for its efficiency in clustering textual data. The separation of point i is its average distance to all points in the closest cluster to its cluster I_k:

$$b(i) = \min{}_{k' \neq k} \frac{1}{|I_{k'}| - 1} \sum_{i' \in I_{k'}} d(x^i, x^{i'}) \tag{2}$$

The silhouette coefficient of point i is then computed as:

$$\begin{aligned} s_{silhouette}(i) &= \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \quad && if : |C_i| > 1, \\ s_{silhouette}(i) &= 0 && if : |C_i| = 1 \end{aligned} \tag{3}$$

We aim to cluster the newsletters using only their emotion and sentiment features. However, some of these features are significantly correlated, as shown in Table II. We then used PCA to denoise the data and construct a representation free of redundant information.

We, therefore, had two hyperparameters to determine: the appropriate number of principal components and the optimal number of clusters. To this end, we used two criteria: the ratio of variance explained by the PCA components and the clustering silhouette score. Table III gives, for each number of principal components, its explained variance rate, its associated optimal clustering and the corresponding silhouette score.

It appears that the partition into two clusters is the best clustering configuration for most of the PCA representations.

We decided to consider eight principal components in the subsequent analysis because they account for more than 91% of the variance in our data set.

TABLE III
OPTIMAL CLUSTERINGS ASSOCIATED WITH DIFFERENT REPRESENTATIONS OF THE DATA

PCA[a]	Explained variance	Number of clusters[b]	silhouette score
1	24%	2	0.577
2	40%	2	0.501
3	53%	4	0.411
4	63%	2	0.358
5	72%	2	0.274
6	79%	2	0.269
7	86%	3	0.250
8	91%	2	0.258
9	96%	4	0.392
10	100%	4	0.366

[a] Number of PCA components
[b] The optimal number of clusters is chosen to maximize the silhouette score

We compared the performance of the two resulting clusters in terms of click rates, but the results were not conclusive. The distributions of the click-through rates in the two clusters are presented in Figure 3.

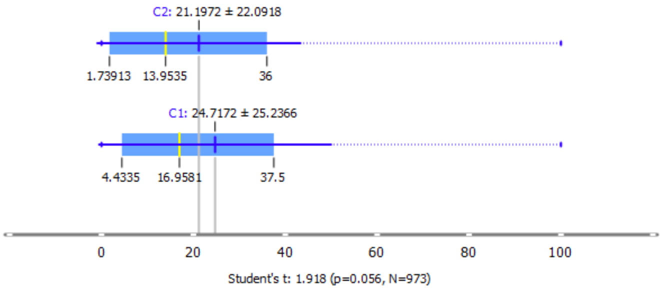


Fig. 3. Click rates in the two clusters obtained by k-means with 8 principal components

As a result, the emotion and sentiment features alone did not allow us to discriminate between "good" and "bad" newsletters. Nevertheless, in the next section, we use another approach based on supervised classification.

VI. Classes of performance prediction

A. Classes of performance

To implement supervised classification, we need to label our data set. Following the considerations in the previous section, we decided to create two performance classes around the median click rate. One class contains the 50% of newsletters that generate the fewest clicks, and the other class contains the highest click rates. We refer to them as the "poor or lower-performing" class and the "good or higher-performing" class.

Figure 4 gives the distribution of data silhouette scores by performance class. Here, the embeddings cover all emotion and sentiment features.

We observe that newsletters with lower click rates are more dispersed around their class center than better performing newsletters. This trend is even more marked when we do not

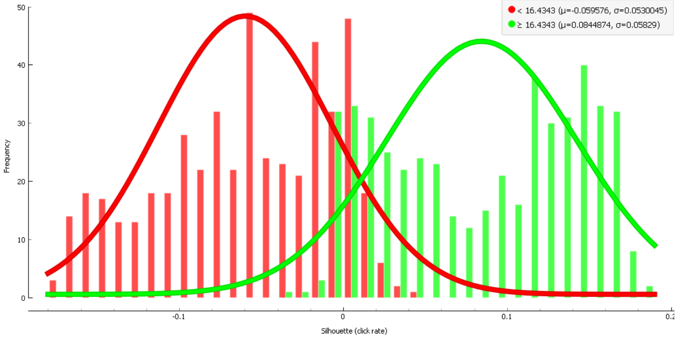


Fig. 4. Distribution of silhouette scores in the "bad" class (red) and the "good" class (green), with subject line features

take into account the subject line's subjectivity and polarity, as shown in Figure 5.

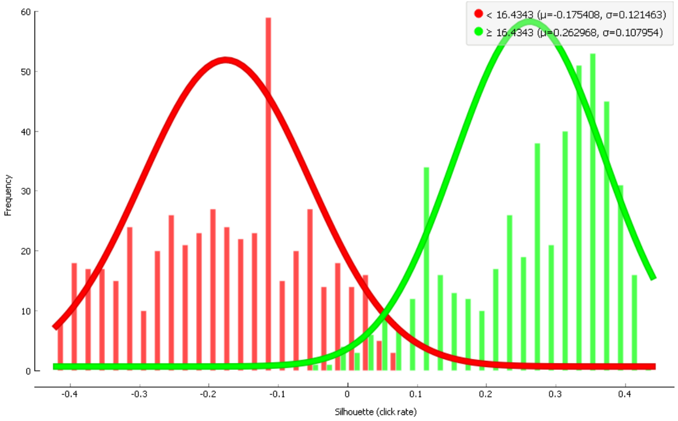


Fig. 5. Distribution of silhouette scores in the "bad" class (red) and the "good" class (green), without subject line features

B. Supervised classification

We are interested in evaluating the predictive power of the ten emotion and sentiment features of the newsletters, categorized into bad and good newsletters as defined above.

For this purpose, we performed classifications of our dataset with various machine learning methods [15]. Table IV presents their performance measures estimated with 10-fold cross-validation procedure.

It appears that the best classifiers are AdaBoost, Neural Network, and Random Forest. We also notice that the performance scores are very slightly lower without the subject line information.

To measure the contribution of each feature in the predictive model, we tested "leave-one-out" and "one-at-once" procedures with the best classifier, Adaboost. In leave-one-out experiments, we considered all but one feature, and in one-at-once experiments, we considered one feature at a time. The F1-scores presented in Table V are to be compared with the F1-score of the full model, constructed with all predictors, which is 0.723 (see Table III).

TABLE IV
PERFORMANCE SCORES OF THE CLASSIFIERS, WITH AND WITHOUT SUBJECT LINE INFORMATION

Classifier	F1 Score	Precision	Recall
With subject line information			
AdaBoost	**0.723**	**0.724**	**0.724**
Neural Network	0.712	0.712	0.712
Random Forest	0.711	0.711	0.711
kNN	0.681	0.688	0.683
Naive Bayes	0.666	0.666	0.666
SVM	0.607	0.617	0.612
Logistic Regression	0.585	0.594	0.590
Constant	0.500	0.500	0.500

Without subject line information			
Model	**F1 Score**	**Precision**	**Recall**
AdaBoost	**0.722**	**0.723**	**0.723**
Neural Network	0.714	0.715	0.715
Random Forest	0.710	0.710	0.710
kNN	0.679	0.683	0.680
Naive Bayes	0.666	0.666	0.666
SVM	0.628	0.640	0.633
Logistic Regression	0.621	0.643	0.630
Constant	0.500	0.500	0.500

TABLE V
ADABOOST PERFORMANCE SCORES WITH A SINGLE FEATURE OR ALL BUT ONE FEATURE

Feature	F1-score with a single feature	F1-score with all but one feature
Subject line polarity	0.498	0.720
Subject line subjectivity	0.503	0.721
Content Polarity	0.614	0.719
Content Subjectivity	0.570	0.725
Content Joy	0.624	0.723
Content Fear	0.604	0.722
Content Sadness	**0.633**	**0.711**
Content Anger	0.618	0.713
Content Surprise	0.614	0.721
Content Disgust	0.626	0.721

These results confirm the impact of emotions and sentiment on newsletter click rates, and as observed in Section IV, sadness is the emotion with the most impact. We can also see that text content subjectivity has a negative effect on prediction. We should investigate these observations further to improve our embeddings in future work.

VII. CONCLUSION

In this paper, we explored to what extent emotion and sentiment detection can help predict the performance of an email campaign. Literature in the marketing field suggests that email communication generally results in a misunderstanding of the emotions being conveyed and, due to negativity and neutrality effects, these emotions are often misinterpreted as neutral or negative by the recipient. When the recipient is a potential customer or subscriber, this negative effect can lead to unwanted behavior, measured with objective metrics such as open rate or click rate.

We presented a dataset composed of French emailing campaigns and represented them with emotion and sentiment embeddings. Our study shows that almost all emotions are negatively correlated with newsletter performance, especially sadness.These results are consistent with the marketing literature, which suggests that negative emotions, such as sadness, are well identified by the recipient, while positive emotions or opinions in a text (represented by the subjectivity score) are poorly understood.

In addition, we observed that the best-performing newsletters have more homogeneous emotion and sentiment features than the less-performing newsletters. This finding needs further investigation to build a guide for writing effective newsletter.

Finally, we used emotion and sentiment embeddings to predict performance classes of our newsletters. The presented approach is perfectible, but it already constitutes a good baseline for our future work on emotion detection in French emails. Areas of improvement concern, in particular, the hyper-parameters of the classifiers and the embeddings. Moreover, we will soon provide the scientific community with our dataset to enrich the French resources and allow interested researchers to reproduce and improve our work.

References

[1] K. Byron, “Carrying too heavy a load? the communication and miscommunication of emotion by email,” *The Academy of Management Review*, vol. 33, no. 2, pp. 309–327, 2008, Accessed on Aug. 18, 2021. [Online]. Available: http://www.jstor.org/stable/20159399

[2] J.-E. Kim and K. Johnson, “The Impact of Moral Emotions on Cause-Related Marketing Campaigns: A Cross-Cultural Examination,” *Journal of Business Ethics*, vol. 112, no. 1, pp. 79–90, January 2013, Accessed on Aug. 18, 2021. [Online]. Available: https://ideas.repec.org/a/kap/jbuset/v112y2013i1p79-90.html

[3] M. S.-F. Virginie Rodriguez, “Le contenu des communications relationnelles par email des enseignes : Quelle perception par le consommateur ? [Content of retailers' relational e-mails: what is the consumer's perception ?],” in *20th International Marketing Trends Conference*, Venise, Italy, Jan. 2021, Accessed on Aug. 29, 2021.

[4] R. Miller and E. Charles, “A psychological based analysis of marketing email subject lines,” in *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2016, pp. 58–65.

[5] B. Klimt and Y. Yang, “The enron corpus: A new dataset for email classification research,” in *Machine Learning: ECML 2004*, J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 217–226.

[6] R. Kalitvianski, “Traitements formels et sémantiques des échanges et des documents textuels liés à des activités collaboratives[Formal and semantic processing of exchanges and textual documents related to collaborative activities.],” Theses, Université Grenoble Alpes, Mar. 2018, Accessed on Sep. 3, 2021. [Online]. Available: https://tel.archives-ouvertes.fr/tel-01893348

[7] H. Guenoune, K. Cousot, M. Lafourcade, M. Mekaoui, and C. Lopez, “A dataset for anaphora analysis in French emails,” in *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*. Barcelona, Spain (online): Association for Computational Linguistics, Dec. 2020, pp. 165–175, Accessed on Aug. 27, 2021. [Online]. Available: https://aclanthology.org/2020.crac-1.17

[8] A. Seyeditabari, N. Tabari, and W. Zadrozny, “Emotion Detection in Text: a Review,” p. arXiv:1806.00674, Jun. 2018, Accessed on Sep. 1, 2021.

[9] P. Ekman, *Basic Emotions*. John Wiley & Sons, Ltd, 1999, ch. 3, pp. 45–60, Accessed on Aug. 21, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013494.ch3

[10] A. Abdaoui, J. Azé, S. Bringay, and P. Poncelet, “FEEL: a French Expanded Emotion Lexicon,” *Language Resources and Evaluation*, vol. 51, no. 3, pp. 833–855, Sep. 2017, Accessed on Sep. 3, 2021. [Online]. Available: https://hal-lirmm.ccsd.cnrs.fr/lirmm-01348016

[11] A. Kumar, “An empirical examination of the effects of design elements of email newsletters on consumers' email responses and their purchase,” *Journal of Retailing and Consumer Services*, vol. 58, p. 102349, 2021, Accessed on Aug. 29, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S0969698920313576

[12] A. Bonfrer and X. Drèze, “Real-time evaluation of e-mail campaign performance,” *Marketing Science*, vol. 28, no. 2, p. 251–263, 2009, Accessed on Aug. 22, 2021. [Online]. Available: https://doi.org/10.1287/mksc.1080.0393

[13] U. Yaqub, S. A. Chun, V. Atluri, and J. Vaidya, “Analysis of political discourse on twitter in the context of the 2016 us presidential elections,” *Government Information Quarterly*, vol. 34, no. 4, pp. 613–626, 2017, Accessed on Aug. 31, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S0740624X17301910

[14] L. Kaufman and P. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons., 1990, john Wiley & Sons, New York.

[15] A. Mueller and S. Guido, *Machine learning avec Python*. O'Reilly Media, Inc., 2018.

User Sentiments Towards Smart Grid Flexibility

A survey of early adopters' attitude towards allowing third parties to control electricity use in households

Marius Rohde Johannessen
School of Business
University of South-Eastern Norway
Horten, Norway
email: marius.johannessen@usn.no

Lasse Berntzen
School of Business
University of South-Eastern Norway
Horten, Norway
email: lasse.berntzen@usn.no

Boban Vesin
School of Business
University of South-Eastern Norway
Horten, Norway
email: boban.vesin@usn.no

Qian Meng
School of Business
University of South-Eastern Norway
Horten, Norway
email: qian.meng@usn.no

Thomas Brekke
School of Business
University of South-Eastern Norway
Horten, Norway
email: thomas.brekke@usn.no

Inessa Laur
School of Business
University of South-Eastern Norway
Horten, Norway
email: inessa.laur@usn.no

***Abstract*—In this paper, we present the findings from a pilot survey on attitudes towards incentives for allowing third parties to control electricity use in households as part of the change to smart, green, and sustainable power grids. The survey was aimed at early adopters of smart home technology and shows that for this group, there is significant resistance towards allowing a third party to control household electricity use, at least unless the monetary incentive is high. However, early adopters are positive towards using smart home technology to lower their electricity bill if they stay in control.**

Keywords-smart grid; user sentiment; adoption; smart home; incentives; Smart-MLA.

I. Introduction

The energy market in Europe is in a state of change. The transition from fossil fuels to renewable energy, new industries such as hydrogen, electric vehicle batteries, data centers, and green shift in existing industries require electricity. The change demands an increased focus on effectiveness and smart utilization of energy resources and the power grid. In 2019 alone, Europe installed 30 GW of renewable energy production, cutting emissions by 12% compared to the year before [1]. Fighting climate change means the matter is becoming more urgent, and the European Union has set a target of 55% emissions cuts by 2030 [2], which means European energy markets need to speed up their work on smart grids, as the transition to renewables means less oil and gas and more electricity [1].

Smart grids, allowing bidirectional power flow, two-way communication, and control functions, are essential to this transition to handle the increased need [3]. Smart grids provide consumers with the information and tools needed to adjust their energy usage and may contribute to savings for the consumers and reduced needs for electricity in households [4]. However, changing consumers' behavior requires action, and there are several possible strategies, such as policy change, working to change consumer perceptions and attitudes, and material incentives [5].

The transition to renewable smart grids might increase the need for flexibility from consumers, as renewables such as wind and solar do not produce the same amount of electricity throughout the day. The transition makes grid balancing more of a challenge. The solution is either to have backup power (battery storage, coal, gas, hydropower) to meet excess demand or control demand at peak hours [6]. While Norway already produces most of its electricity using renewable hydropower, it is still affected through participation in the European markets. It can play a role in balancing the grid as hydropower can be switched on and off using water as a "battery" [1].

Vrain and Wilson [7] show significant potential for energy saving and CO_2 cuts through smart home technology. Still, Hargreaves and co-authors point out some challenges for adoption: Smart home technology is seen as complicated, time-consuming, and disruptive [8]. Sanguinetti, Karlin, and Ford point out that cost and savings are essential for adopting smart home technology [9]. Hence, there is a need for research on incentives for the adoption and efficient use of smart home technology.

The main scope of the ERA-NET project Smart-MLA (Multi-Layer Aggregator) [10] is to develop cloud-based multi-layer aggregator ICT solutions to facilitate optimum Demand Response (DR) and grid flexibility to energy systems to utilize up to 100% renewable energy. The project includes research on smart grid flexibility and possible barriers to adoption.

Thus, the objective of this paper is to examine incentives for smart home technology, as seen by early adopters. We focus on early adopters because this is the consumer group currently purchasing smart home technology [11], and we want to hear the opinions of actual users. Further, we are focusing on incentives that allow third parties to control household consumption, as this might be necessary at certain times to balance a renewable-driven power grid [6] properly.

The rest of the paper is organized as follows: The following section contains a literature review. Section III discusses the research approach, followed by Section IV presenting the findings. The last section contains the conclusion and ideas for further research.

II. Literature Review

This section provides a literature review, first on smart grids and smart homes, then on adoption, use, and diffusion.

A. Smart grids and smart homes

As defined in the introduction, smart grids are about control, balance, and increased efficiency through communication, allowing users to save energy [4]. Smart grids need to respond to varying supply and demand [12] and rely on smart meters providing real-time consumption data and the possibility to regulate power consumption. Smart grids also include communication technologies such as 4G/5G and smart home protocols (Zigbee, Z-wave, Bluetooth, etc.) for data exchange.

Smart home technology mixes artificial intelligence, communication, monitoring, and control of household appliances [13]. A smart home consists of the external network linking home and grid, a household hub for connecting components, and the individual smart/controllable devices in the house (sensors, thermostats, heating, ventilation, air conditioning, lighting, etc.) [14]. The combination of smart homes and grids allows for dynamic pricing and load-shifting programs for managing demand and supply of electricity [15]. The International Panel on Climate Change (IPCC) estimates a 70% decrease in energy demand from lighting alone if people optimize lighting at home [16].

There are also different approaches to control. Some vendors leave it up to the user to set up automation, turn appliances on and off, etc. In contrast, others apply advanced algorithms attempting to optimize power consumption within the boundaries set by the user – such as needing a full charge on your electric vehicle by 8 am or not allowing the temperature to sink below a certain threshold [15]. A third option is to allow a third party to control some electricity use, or a combination of the above such as outlined by the Smart-MLA project (see Section I).

B. Adoption, ease of use, and diffusion

One challenge with home automation and grid optimization lies in this tension between control, what we are willing to sacrifice, and use complexity. There is some emerging research into this area of home automation usability, such as the paper by Stojkoska and Trivodaliec, which proposes a framework for smart home management [17].

Other studies point to specific challenges in various user contexts. Coughlin and co-authors, for example, have examined the older population's user experience with health-related smart home technology and found that older people tend to see the benefits of technology but still find it challenging to use. There is no comprehensive or integrated market for these things, meaning users have to work with many different user interfaces [18]. Yang, Lee, and Zo also find challenges for user acceptance related to mobility, security, privacy, and trust, suggesting unmet design needs in these systems [19].

Nikou [20] has researched the adoption of smart home technology and found support for an extended technology acceptance model: Perceived usefulness and ease of use were important determinants for adoption, as were compatibility with existing hardware. The cost of systems had a significant negative effect, and men and women have different attitudes towards smart home technology. Shin, Park, and Lee [10] found that the younger age group was more likely to be concerned with usability, while those over 40 were slightly more concerned with usefulness. Those with higher education were, in general, more positive towards smart home technology.

Sanguinetti, Karlin, and Ford applied diffusion of innovation theory to examine smart home energy management adoption and found four clusters of consumer segments: Those unfamiliar with the technology, those who were unpersuaded or persuaded, and finally, owners. Those who owned or planned to purchase smart home technology were, in general, more positive towards and informed about technology. They also had higher incomes and were more likely to own their own home. Those who were less positive pointed to barriers such as the difficulty of setup/use and concerns with the cost of purchase [9].

III. Research Approach

The study was conducted as a pilot survey study [21]. The study was conducted in Norway, so respondents replied with the Norwegian context in mind, which means high consumption due to long and cold winters; users being used to low-moderate prices; and seeing electricity as a shared social good rather than a market commodity, even though the energy sector has been deregulated since the Energy Act of June 1990.

In "The Lean Startup," Ries advocates testing ideas with early adopters [22]. The sample is not representative of the population but is focused on early adopters only since they will provide more valuable responses in the context of this paper.

As we were interested in the attitudes of early adopters, we reached out to two online discussion forums (for smart home automation and electric vehicle enthusiasts) and four Facebook groups (for electric vehicle enthusiasts, two different smart home groups, and a group for electricity pricing). As participation was by self-selection within these groups, we do not claim the findings are representative. However, they still present the sentiment potential early adopters show towards giving up flexibility to gain advantages (rewards or lower bills). The survey was left open for five days, and in this period, we received 209 answers and several comments to the post where we invited people to participate.

In the survey, we asked about the demographic background, existing smart home technology in use, and acceptable incentives for allowing outside control of appliances, using a four-point Likert scale. In addition, we had an open-ended question where respondents could elaborate on their answers, which 52 of the respondents chose to do.

As this is an exploratory pilot survey, we chose not to apply a specific model such as the Technology Acceptance Model (TAM). However, we did include some questions from TAM and related models. At this stage, we are more interested in descriptive statistics of the incentives required for consumers to allow outside control of their electricity use. A more structured model-based survey approach, based on this pilot's answers, is the next step in our research.

IV. Findings

This section discusses findings related to demographic characteristics of respondents and their attitude towards technology, and what they think of incentives and motivation to provide flexibility.

A. *Demographic characteristics of early adopters*

A vast majority of our respondents were male – 95 %. This is perhaps somewhat skewed due to the self-selection of respondents, but other studies of adoption show similar results. Men are more likely to adopt smart home technology, meaning current marketing only reaches half the population. Age-wise, our respondents are mainly in the 30-60 age group, with equal distribution for each decade. This is not surprising as most of them own houses (75%, vs. 12% for apartments and 13% for other housing types). Most Norwegians own their home and typically buy their first home when they get their first job and settle down with a partner in their middle- to late twenties or early thirties.

Further, we see that smart home early adopters are relatively affluent, but not extensively so. According to Statistics Norway, the median income for all households in Norway is € 68,600, for households with no children € 86,000, and € 117,000 for couples with children aged 0-17. In our survey, only 6% have a household income below € 60,000, and 77% earn more than € 100,000. This would put most respondents in a comfortable financial position, with two adults in the household having well-paid jobs, indicating that investment in smart home technology is a surplus phenomenon.

B. *Attitudes towards technology, existing smart technology*

Here, we asked respondents about their attitudes towards technology and technology adoption to examine if they had early adopters' characteristics. Table 1 shows that a vast majority of respondents are positive towards technology and that friends and family will consult them in technical matters. This indicates that we were indeed able to capture the sentiment of early adopters.

TABLE I. Adoption of technology. Responses in percent

"I adopt new technology…"	Fully agree	Somewhat agree	Somewhat disagree	Disagree
quickly	58.4	38.8	2.4	0.5
if it is easy to use	59.1	33.2	6.7	1.0
if it is useful to me	79.9	20.1	0.0	0.0
if the price is right	71.8	25.4	2.9	0.0
Friends and family ask my advice about technology	57.2	36.1	5.3	1.4

We also asked the respondents about their preferred smart home setup (Fig. 1). As early adopters, more than 70% prefer to tinker with advanced settings or custom build their own system.

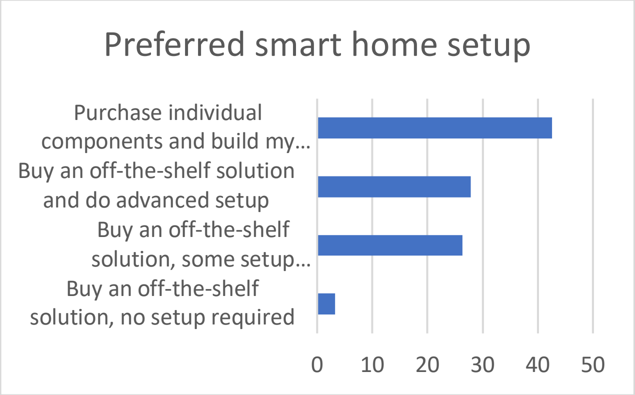


Figure 1. Preferred smart home setup

In Fig. 2, we show the smart home products owned by respondents. We see that the most common are off-the-shelf technology such as electric vehicle charging, smart plugs, and thermostats. 15.5% have installed solar panels, which is quite a bit higher than the national average. 14% report other technology such as Heating, Ventilation, Air-Condition (HVAC), heat pump, sunscreens, alarm systems, and door locks. Two of the respondents have installed battery packs for energy storage.

Only 2.4% report having solar capture technology (storing solar energy as warm water), even though solar capture makes sense in the cold Norwegian climate.

C. *Incentives and motivation*

Here, we asked specifically about incentives for allowing the Distribution System Operator (DSO) or other external parties to regulate different areas of people's homes. The responses are listed in Table 2. In short, we see that significant incentives are needed, and the respondents are generally negative towards allowing others to control their homes' electricity use.

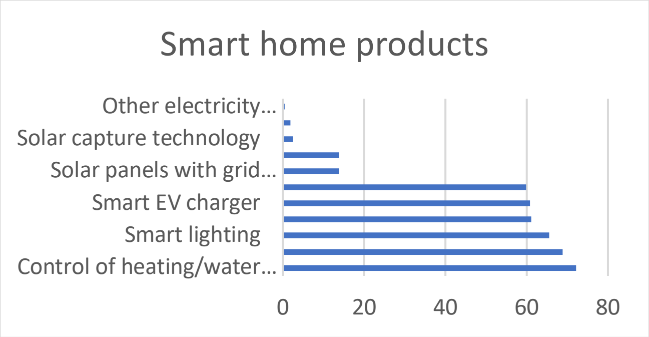


Figure 2. Smart home products owned

They far prefer being in control themselves and setting up automation within the boundaries they find acceptable, such as the electric vehicle having a full charge by a specific time, lowering temperatures when rooms are not in use, etc. There is some interest in allowing outside parties to control electric vehicle charging or the house's water heater, but only if the savings exceed €150 a/year in each case.

TABLE II. INCENTIVES FOR ALLOWING CONTROL TO THIRD-PARTIES

I am willing to let outsiders	Annual savings					
	Less than €30	€30-79	€80-119	€120-149	€150 or more	Not at all
Use my EV's battery to balance the grid	2.9	4.3	10.1	5.3	**39.9**	37.5
Control charging of my EV	10.6	9.7	14.0	6.3	**31.4**	28.0
Control heating in rarely used rooms	11.1	7.7	13.0	5.3	16.8	**46.2**
Control heating in frequently used rooms	4.8	4.8	7.7	3.8	19.2	**59.6**
Control my water heater	12	4.3	16.7	5.3	26.3	**35.4**

Further, 97% report that good statistics and visualizations of energy use and savings are important or somewhat important for their motivation to use smart home technology (Fig. 3).

We also asked about other incentives for energy saving in general. It seems that while early adopters are reluctant to release control to others, they are concerned with societal issues. Keeping costs down for everyone via energy-saving and better utilization of the national grid and contributing to phasing out fossil fuel energy in Europe is seen as important or somewhat important for 70 – 90% of the respondents.

D. *Qualitative concerns*

Summing up the findings, we see that while users are happy to contribute to a more sustainable future and invest in smart home technology to cut costs, monetary incentives need to be significant for users to allow outside control of their home's energy use. The free text answers supplement the survey questions, with 52 of 207 respondents choosing to comment. The following categories emerge from the free text-answers and are candidates for future research:

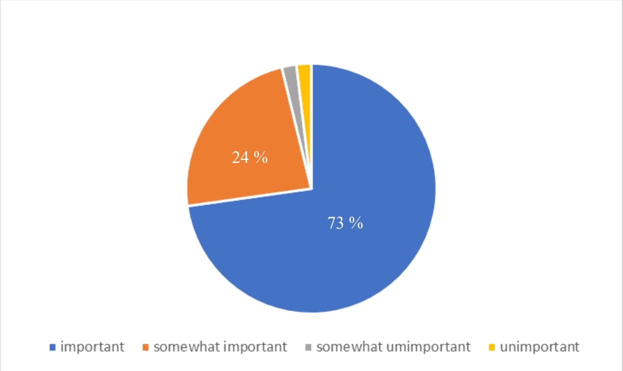


Figure 3. The importance of statistics and visualization

1) The cost of grid access and use is a barrier

In Norway, electricity customers pay a fixed rate plus a certain amount per kWh to access and use the power grid. The sums are set via a complex set of regulations and meant for grid maintenance and updates. This means that the total electricity bill comprises the cost of electricity and the grid access tariffs (plus taxes). There are currently proposals to change this tariff to save on investments in the grid. Several respondents claim that the way this cost is structured, as well as suggestions for tariff changes such as paying for delivering excess solar power to the grid or raising the tariff based on maximum electricity use, take away the monetary incentives for investing in smart home equipment:

> *"The grid tariff in its current and planned form is the major obstacle to a more aggressive approach to cutting electricity consumption."*

> *"I can easily upgrade my home, so I can charge my two electric vehicles with a total of 14 kW during the two hours at night when the grid is least used, but I have no incentives for that [with a tariff based on maximum kWh used]."*

> *"Now we are threatened we might have to pay the DSO for the electricity we supply to the grid from our solar panels."*

2) Money first, ideology second

We also see several comments showing that monetary incentives and the total cost of electricity weighs heavier than

ISBN: 978-1-61208-896-9

social or ideological reasons for power saving and investment, and there are some calls for increased support for the installation and upgrade of solar panels, heat pumps to replace regular heaters, etc. Several mention home installation of solar as an option, but one that is currently too expensive. This is in line with findings from other surveys, such as the Norwegian electric vehicle user survey [23], showing that clean air and fighting climate change are important reasons for adoption, but the monetary savings from electric vehicles are still the most important reason why people buy electric cars.

> *"The choices in the survey are way too low. I'd need to save a lot more than 150 Euros if I were to release control to the DSO."*

> *"I'd need to save a lot more than what the questions in this survey suggest for the investment in smart home equipment to pay off."*

> *"There should be better incentives for a gradual replacement of old technology…such as better grid tariffs."*

Storage capacity is also mentioned as important:

> *"Cheaper solar and maybe battery storage would make this a priority, but without the possibility to store generated electricity for later use yourself, it is too costly… Or maybe the DSO "borrowing" your solar energy [when generated in the summer] and delivering it back to you for free later [in winter, when needs and prices are higher] could work"*

Others point out that that smart home technology is too expensive for some (as reflected in the income question in the survey):

> *"Those with a lot of money can afford to do all kinds of things and are rewarded with money for doing it, but others can't afford to invest in power-saving technology. So this pricing of maximum effect used will hit the poor hardest."*

3) Users are happy to invest in power saving smart homes but prefer to be in control

This is by far the topic most commented on, which is not surprising given that we asked about incentives for allowing others to control the use of electricity in people's homes. The conclusion seems quite clear, both from the free text answers and the survey: Most users are interested in lowering their electricity use and keeping costs down, but they are not comfortable allowing the DSO or other parties to control this. They list several reasons:

Lack of trust is a recurring issue. In Norway, cheap electricity used to be seen as a common good, where prices were kept low so people could stay warm in the cold winter. After deregulation in the 90's/00's and the establishment of the Nordpool electricity market, however, prices have fluctuated a lot more, and the media covers every price raise. This seems to have led to a lack of trust in the market and Transmission System Operators (TSOs) in particular:

> *"Energy companies will never be allowed to control anything in my house. They have shown time and time again they can't be trusted, with their hidden terms and conditions."*

> *"The DSOs…have neglected investing in the grid for the past 25 years while paying out hundreds of millions to shareholders. It's time they step up, without shoving the [financial] burden on to consumers."*

> *"I don't trust them. What if something goes wrong?... and if the system is able to cut costs, that won't get back to consumers."*

Privacy and security issues are also mentioned as reasons for not allowing outside control.

> *"privacy issues…if something is to be controlled, or data stored, who has access and for what purpose? How are data kept?"*

> *"I don't want anything in the cloud or stored on external servers. (there is no Cloud - it's just someone else's computer)."*

Technology not perceived as mature. Some raise concerns that the technology just isn't ready yet, or not stable enough.

> *"What happens when the DSO system suddenly crashes, and you have no electricity in your car, no hot water, no heating?"*

> *"I have tried to turn control over to a third party but found the technology was just not mature yet."*

> *"I have the hub from [producer name], and while it is ok to use, it is a bit complicated. I think regular users with little interest in technology would struggle with setting up conditions and rules".*

4) Social aspects.

Finally, we see some comments regarding social and societal aspects. One respondent says, "this is mostly for people with interest in technology. There's no way I can get my family on board with these things" – a statement supported by the fact that 95% of the survey respondents are male. Others are concerned with sustainability and are positive towards efforts that visualize their carbon footprint:

> *"It is just as important to inform and visualize the greater good, for example, by creating a community for those who allow the DSO to take control and show what this effort does in terms of energy-saving."*

> *"I would like to see my carbon footprint and how [smart technology] contributes to a more green and sustainable consumption of electricity."*

V. Conclusion and Future Research

In this paper, we have reported the findings of a survey on incentives for allowing the DSO or other third parties to control household electricity use through smart grid/smart home technology. The survey was aimed at early adopters of technology, as this is the group who so far seems to have invested the most in this kind of technology, and also are more reflecting on issues related to electricity use.

The responses show there is a lot of interest in this issue, with more than 200 replies in just a few days, and over 50 free text comments elaborating on the answers, as well as comments directly in the forums and Facebook posts used to recruit respondents. The main finding is that smart home users are interested in saving money by controlling household energy use, but they are unwilling to allow third parties to take control. Monetary incentives seem to be the most important, with most saying they need to save more than € 150 a/year for each of the categories listed in the survey.

For practitioners, our survey shows that to make the grid smarter and control household consumption in peak hours, the consumer needs to be rewarded enough to offset the resulting lack of flexibility. Trust seems to be a barrier, so there is a need to address this by clearly showing how and how much households benefit. Finally, we see a great deal of interest in this area. It seems many consumers (at least in the demographics who responded to the survey) are willing and eager to save on their electricity bills through smart home technology.

For researchers, the free text answers and comments reveal some emerging themes, which should be topics of future research on smart grids and user acceptance:

- The cost of grid access, use, and smart home technology is a barrier to investment
- Ideology and sustainability are important, but money comes first
- Users are happy to invest in power-saving smart homes but prefer to be in control
- The technology is not yet perceived as mature.
- Social aspects, including sustainability and gender differences, are important

The lack of trust and reluctance to surrender control and flexibility to the DSO could perhaps be offset by more localized initiatives, such as the neighborhood approach proposed by the Smart-MLA project, where an aggregator acts as a broker between consumers and DSO. Figuring out how to organize this is also a topic for future research.

Finally, the responses are skewed towards males with a relatively high income and deliberately aimed at early adopters. Future research should aim to examine the views of the wider population, including the late majority attitudes towards ease of use. While our early adopter sample prefers to build their smart home systems or tinker with complex settings and adjustments, user research has shown this is not the case for most users.

Acknowledgement

This work was supported under the ERA-Net Smart Grids Plus scheme Grant number 89029 and funded through the Research Council of Norway Grant number 295750 with the project title "Multi-layer aggregator solutions to facilitate optimum demand response and grid flexibility."

References

[1] The Norwegian Water Resources and Energy Directorate (NVE), "Langsiktig kraftmarkedsanalyse 2020-2040," 2020.

[2] European Commission, "Forging a climate-resilient Europe - the new EU Strategy on Adaptation to Climate Change," Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the regions, COM(2021)82 final, 2021 .

[3] R. Bayindir, I. Colak, G. Fulli, and K. Demirtas, "Smart grid technologies and applications," Renewable and Sustainable Energy Reviews, vol. 66, pp. 499–516, 2016.

[4] U. Shahzad, "Significance of Smart Grids in Electric Power Systems: A Brief Overview," Journal of Electrical, Electronics, Control and Computer Science, 6(19), pp. 7-12, 2020.

[5] P. C. Stern, "Information, Incentives, and Proenvironmental Consumer Behavior," Journal of Consumer Policy, 22, pp. 461-478, 1999.

[6] L. Bird, M. Milligan and D. Lew, "Integrating Variable Renewable Energy: Challenges and Solutions," National Renewable Energy Laboratory (NREL), 2013.

[7] E. Vrain and C. Wilson, "Social networks and communication behaviour underlying smart home adoption in the UK," Environmental Innovation and Societal Transitions, 38, pp. 82–97, 2021.

[8] T. Hargreaves, C. Wilson, and R. Hauxwell-Baldwin, "Learning to live in a smart home," Building Research & Information, 46(1), pp. 127–139, 2018.

[9] A. Sanguinetti, B. Karlin, and R. Ford, "Understanding the path to smart home adoption: Segmenting and describing consumers across the innovation-decision process," Energy Research & Social Science, 46,. pp. 274–283, 2018.

[10] Smart-MLA, "Project Fact Sheet," [Online] Avaliable from http://smart-mla.stimasoft.com/wpcontent/uploads/2020/02/ERANetSES_SMART_MLA.docx, 2021.08.14.

[11] J. Shin, Y. Park, and D. Lee, "Who will be smart home users? An analysis of adoption and diffusion of smart homes," Technological Forecasting and Social Change, 134, pp. 246–253, 2018.

[12] J. N. Bharothu, M. Sridhar, and R. S. Rao, "A literature survey report on Smart Grid technologies," in 2014 International Conference on Smart Electric Grid (ISEG), 2015.

[13] M. R. Alam, M. B. I. Reaz, and M. A. M. Ali, "A review of smart homes - Past, present, and future," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(6), pp. 1190–1203, 2012.

[14] M. A. Al-QUtayri and J. S. Jeedella, "Integrated Wireless Technologies for Smart Homes Applications," In: Smart Home Systems, InTech, 2010.

[15] G. Lobaccaro, S. Carlucci, and E. Löfström, "A Review of Systems and Technologies for Smart Homes and Smart Grids," Energies, 9(5), 348, 2016.

[16] IPCC, "Intergovernmental Panel on Climate Change," in Climate Change 2014: Mitigation of Climate Change, New York: Cambridge University Press, 2014.

[17] B. L. R. Stojkoska and K. V. Trivodaliev, "A review of Internet of Things for smart home: Challenges and solutions," Journal of Cleaner Production, 140(3), pp. 1454–1464, 2017.

[18] J. F. Coughlin, L. A. D'Ambrosio, B. Reimer, and M. R. Pratt, "Older adult perceptions of smart home technologies: Implications for research, policy & market innovations in healthcare," in Proceedings 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1810–1815, 2007.

[19] H. Yang, H. Lee, and H. Zo, "User acceptance of smart home services: An extension of the theory of planned behavior," Industrial Management and Data Systems, 117(1), pp. 68-89, 2017

[20] S. Nikou, "Factors driving the adoption of smart home technology: An empirical assessment," Telematics and Informatics, 45, 101283, 2019.

[21] B. Pikkemaat and M. Peters, "Towards the measurement of innovation- a pilot study in the small and medium sized hotel industry," In Innovation in Hospitality and Tourism, Routledge, pp. 89–112, 2006.

[22] E. Ries, "The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses," Vikin, 2011.

[23] Elbilforeningen [Online] Available from https://elbil.no/elbilisten/, 2021.08.21