The influence of dialogue flow on stress levels when booking healthcare appointments with AI

Milos Kravcik[1], Elisabeth Reiswich[2], Iwan Lappo-Danilewski[2],
David Buschhüter[1], Patrick Jähnichen[2]
[1] DFKI, Alt-Moabit 91c, 10558 Berlin, Germany
[2] Doctolib GmbH, Mehringdamm 51, 10961 Berlin, Germany

Abstract. Advancements in artificial intelligence (AI) have opened new possibilities for user adaptation, providing greater accessibility to healthcare and mental health support. Stress is an inevitable phenomenon in modern society and has become a critical factor for well-being in both the professional and personal spheres. This paper examines the application of human-centred AI in an appointment-booking scenario, with a focus on stress detection in patients. The study utilises pre-trained machine learning classifiers for user adaptation and examines the effect of various prosody rates on stress, considering measurements to identify physiological biomarkers associated with psychological stress. Test scenarios involving appointment booking, prescription and referral requests via the digital phone assistant were explored. The speech rates of the AI assistant were randomised and varied from slow to normal to fast. Our analysis reveals promising results in distinguishing between static and dynamic systems, using a sample size of n = 12 in a user study. Stress responses measured with Empatica EmbracePlus suggest that dynamic systems are preferred over static ones. This finding could be replicated using self-reports from participants in the study. This work contributes to the growing body of research on digital health tools for healthcare assistance, highlighting the need for interdisciplinary collaboration to advance the field responsibly. With stress-related disorders and AI usage rising globally, understanding the interaction between stress and automated dialogue flow could provide helpful strategies to improve the user experience, which could then be scaled up to other health and work environments.

Keywords: User Adaptation, Stress Response, Biosensors, Human-Centered AI, Health Care, ML, HCI.

1 Introduction

Developing an AI model that can adapt the dialogue behaviour of telephone-based voice assistants (e.g. tone, choice of words and prosody) to the needs and meta-communicative signals of callers, as well as relevant contextual information, is challenging. This would include operator-specific conditions, previous conversations, and cross-channel interaction history. The intention is to increase acceptance and reduce personnel workload. In our context, optimising dialogue flow aimed to achieve an appropriate, dynamic formulation that considers input parameters and appropriate, natural prosody.

Overall, this was a multidimensional, highly complex optimisation problem with a correspondingly large parameter space. Another technical objective was to make the models and findings usable for real-time dialogue optimisation.

In adaptive systems, user characteristics can be categorised as either static (e.g. personality, cultural background) or dynamic (e.g. context, mood, emotions) for customisation purposes. The Big Five [1] are main personality dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism), which are assessed using questionnaires or personality tests. We had to consider the relevant static and dynamic user and context characteristics, and define useful metrics for dialogue behaviour.

Intelligent conversational assistants, such as chatbots, are widely used today. These chatbots utilise various types of AI, including natural language understanding, machine learning and deep learning. Chatbots have already been successfully deployed in various domains, including customer service. Their ability to interact via natural language provides an intuitive user experience, which is particularly beneficial for older people who are less familiar with the internet. It has emerged that a key requirement for chatbots is the ability to correctly interpret users' messages and intentions, and adapt to different users to avoid frustration.

2 Related Work

Digital assistants can help people in various situations, related to healthcare. A smart, pervasive chatbot supports victims or witnesses of emergencies, helping to stabilise their condition until assistance arrives [2]. Another example is a conversational bot that uses AI to bridge the gap between demand for and supply of human healthcare providers, thereby improving patients' access to healthcare knowledge [3].

Despite the growing technical maturity and availability of commercial dialogue systems, user acceptance remains an area for improvement. A German study [4] showed that, while these systems can reduce hotline staff workload, they cannot fully replace personal conversations, particularly when dealing with complex issues that require empathy. Recent advances in natural language processing have increased interest in using Large Language Models (LLMs) to improve dialogue systems. Studies such as [5] have explored the ability of Very Large Language Models (VLLMs) to generate coherent, contextually relevant responses, while [6] investigated the use of LLMs such as OpenAI’s GPT or Google’s BERT as user simulators for system training and evaluation. Recent research [7] suggests that LLMs can serve as substitutes for human judges in dialogue evaluation, excelling in assessing coherence, relevance and overall quality. However, challenges remain, particularly with regard to specificity and diversity.

3 Our Approach

We planned to use a survey and data analysis to identify the relevant characteristics and personas (user groups) for adaptation and personalisation. For instance, language, speed, word choice and sentence length could be adapted for different age groups and language proficiency levels. The Flesch index, for instance, measures the readability of

a text. Modern information technologies, such as LLMs, can simplify and rationalise communication processes at human-machine interfaces on a massive scale. In healthcare, particularly in patient interactions, the acceptance of automated voice assistants by users remains a challenge.

To address these issues, we conducted an experiment using a novel user satisfaction model [8]. This model uses simple yet insightful dialogue features, collected during real interactions, to quantitatively predict interaction dropouts. Furthermore, we investigated ways to prevent dropouts and boost user satisfaction by tailoring user-driven utterances based on the language model.

4 Sensor-based Affect Recognition

The aim of our study was to use physiological sensors to identify potential communication issues in conversations between patients (callers) and a digital phone assistant implemented at a test practice. This raises various questions: For instance, can observable emotions correlate with communication or dialogue problems? Are there patterns in sensor data that correspond to patterns in dialogue? Do language parameters (e.g. speaking rate, gender, prosody, vocabulary and sentence complexity) influence stress levels? We have formulated two hypotheses: 1. Language parameters (e.g. speech rate) influence the caller's stress/emotional state. 2. The LLM-based (dynamic) system has a different effect on the caller's stress level compared to the original (static) system.

The Empatica EmbracePlus sensors, which can measure electrodermal activity (EDA), digital skin temperature (TEMP), acceleration (ACC), blood volume pulse (BVP) and heart rate, were used. In an experiment involving 12 participants, each person made four telephone calls to a digital telephone assistant simulating calls to a doctor. The calls were categorised by task: 1. To make an appointment (static system), 2. To obtain a prescription (static system), 3. To obtain a referral (static system), 4. To make an appointment (dynamic system).

In the static system, the tasks were randomised to avoid influencing the perception of dialogue experience through differences in the tasks. Our focus was on parameters related to stress. When analysing the data, XGBoost (Extreme Gradient Boosting) was used as the classification method. To provide labels for physiological stress, we pretrained a classifier based on the WESAD Dataset [9]. Stress detection was based on binary classification using the EDA+TEMP modalities. Two diagrams (Fig. 1) enable us to compare the four tasks (upper figure) with the stress indicator (lower figure). It can be interpreted that the person experiences a higher level of stress when communicating normally and slowly with a static system and favours a dynamic system instead.

The results of our stress frame experiment indicate individual preferences for speech rate and a general preference for the dynamic system. We compared the results in different ways. The most interesting is the comparison between the static system with normal speaking rate (in different tasks) and the dynamic system. The specified stress level is much less frequent in the dynamic system than in the static one. Of course, we are aware of the limitation of our experiment in terms of the number of participants. These observations need to be confirmed in larger studies.

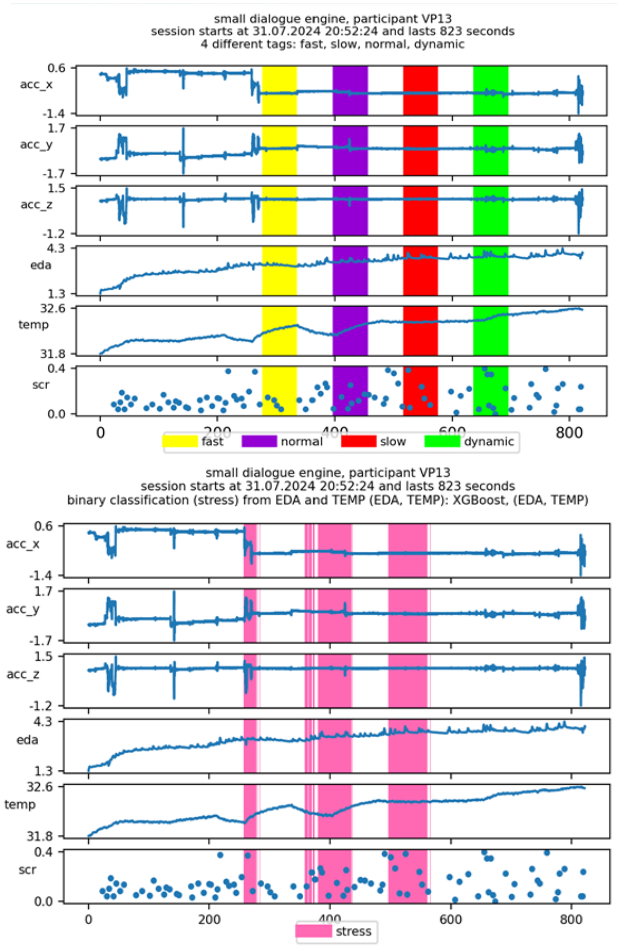

Fig. 1. Stress detection based on binary classification using EDA+TEMP modalities.

5 Survey

A pre-conducted survey of randomly selected doctor practices called into question the influence of the Big Five personality traits, basic emotions (happiness, grief, anger, fear, surprise and disgust) and user stress on the use of the digital phone assistant. Participants estimated that the influence of neuroticism and all basic emotions except surprise and disgust, as well as stress, was 50–60%. The role of stress in communication with the digital phone assistant led to the development of the stress framework. Additionally, the ATI (affinity for technology interaction) score was obtained to determine whether individuals with a greater affinity for technology experience less stress when using the phone assistant. Participants reported enjoying using the digital system, with

80% expressing a preference for it. Notably, users with a greater affinity for technology exhibited reduced stress in the digital scenario.

6 Conclusion

To investigate the acceptance of various alternatives in automated patient communication, we conducted a sensor-based affect recognition study and a survey. The preliminary results of our stress-frame experiment suggest that people have individual preferences for speaking rate and a general preference for the dynamic system (LLM-based). Our findings also indicate that insights regarding physiological stress could be applied to other contexts. For example, speech feedback in hectic traffic situations could enhance the user experience.

References

1. McCrae, R.R. and Costa, P.T., Jr., (1999). A five-factor theory of personality. *In* L.A. Pervin and O.P. John (eds.), *Handbook of personality: Theory and research*. 2nd ed. New York: Guilford Press, pp.139-153.
2. Ouerhani, N., Maalel, A., & Ben Ghézela, H. (2020). SPeCECA: a smart pervasive chatbot for emergency case assistance based on cloud computing. *Cluster Computing*, *23*(4), 2471-2482.
3. Bharti, U., Bajaj, D., Batra, H., Lalit, S., Lalit, S., & Gangwani, A. (2020, June). Medbot: Conversational artificial intelligence powered chatbot for delivering tele-health after covid-19. In *2020 5th international conference on communication and electronics systems (ICCES)* (pp. 870-875). IEEE.
4. Voelskow, V., Meßner, C., Kurth, T., Busam, A., Glatz, T., & Ebert, N. (2023). Prospective mixed-methods study evaluating the potential of a voicebot (CovBot) to relieve German health authorities during the COVID-19 infodemic. *Digital Health*, *9*, 20552076231180677.
5. Huynh, J., Jiao, C., Gupta, P., Mehri, S., Bajaj, P., Chaudhary, V., & Eskenazi, M. (2023). Understanding the effectiveness of very large language models on dialog evaluation. *arXiv preprint arXiv:2301.12004*.
6. Hu, Z., Feng, Y., Luu, A. T., Hooi, B., & Lipani, A. (2023, October). Unlocking the potential of user feedback: Leveraging large language model as user simulators to enhance dialogue system. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (pp. 3953-3957).
7. Zhang, C., D'Haro, L. F., Chen, Y., Zhang, M., & Li, H. (2024, March). A Comprehensive Analysis of the Effectiveness of Large Language Models as Automatic Dialogue Evaluators. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 17, pp. 19515-19524).
8. Soliman, H., Kravcik, M., Basvoju, N., & Jaehnichen, P. (2024, June). Using Large Language Models for Adaptive Dialogue Management in Digital Telephone Assistants. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (pp. 399-405).
9. Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Van Laerhoven, K. (2018, October). Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction* (pp. 400-408).