Speech Matters – Psychological Aspects of Artificial versus Anthropomorphic System Voices in User-Companion Interaction

Swantje Ferchow(✉), Matthias Haase, Julia Krüger, Matthias Vogel, Mathias Wahl, and Jörg Frommer

Department of Psychosomatic Medicine and Psychotherapy, Medical Faculty, Otto-von-Guericke University Magdeburg, Magdeburg, Germany
{swantje.ferchow, matthias.haase, julia.krueger, matthias.vogel, mathias.wahl, joerg.frommer}@med.ovgu.de

Abstract. The design of this forthcoming study was created to investigate the influences of different system-voices on users while they interact with a simulated Companion-system. By using a Wizard of Oz experiment, we want to find out what kind of voice output (artificial vs. anthropomorphic) is better suited for keeping up users' cooperation with a system while solving a task. The goal of this study is to gain a deeper understanding of influences of the speech-output in User-Companion Interaction. Users' perceived trustworthiness towards the system, their experienced affective states and individual user characteristics as important mediators are the main focus of the present study.

Keywords: Companion-system · Wizard of Oz experiment · System voice · Anthropomorphism · User characteristics

1 Introduction

For speech-based dialog systems without visual representation, the system's voice is the only feature a user can relate to. Therefore, it has to be perceived as trustworthy and empathic. This especially applies to Companion-systems, which are "cognitive technical systems with their functionality completely individually adapted to each user […which] interact with [him/her] as competent and cooperative service partners" [1]. Companion-systems should be able to support every user in different situations and in all kinds of emotional states – positive as well as negative [1].

In general, anthropomorphic and/or naturally sounding voices are used in most areas of Human-Computer Interaction (HCI), like navigation systems, smart home environments or voice user interfaces (VUIs) in smartphones [2]. There is indeed evidence for the human tendency to use schemes from human-human interaction for the communication with computer systems or virtual agents, regardless of the level of anthropomorphism of their voice [3]. However, empirical findings, which support the hypothesis that human-like, anthropomorphic voices most likely support human-computer cooperation as well as users' perceived trust, are rare in comparison to

A. Marcus (Ed.): DUXU 2016, Part III, LNCS 9748, pp. 319–327, 2016.
DOI: 10.1007/978-3-319-40406-6_30

artificial voices [4]. In order to provide a deeper comprehension of users' individual experiences with different kinds of voices, an established experimental design [5] was adapted with the focus on users' subjective perceptions of two diverse voices in a task-related dialog with a simulated Companion-system.

2 Background

During the past decades, the recognition of the importance of users' emotions increased significantly in the field of HCI, which lead to the emergence of the research area of Affective Computing [6]. Up to now, it hardly seems imaginable to do research without the consideration of users' affective states, especially as far as User-Companion Interaction (UCI) is concerned [7]. Actual user affect influences most factors of users' perception of and experience with systems, e.g. performance, cognition, concentration, or memory [8]. Therefore, systems must be able to avoid negative affective states for creating and perpetuating cooperation as well as trust. For this current research, this shall be realized by the voice solely.

The question arose if the variation of the speech-output at all is able to fulfill this requirement. To answer this, several studies which investigated the impact of systems' voices on users' perception and user behavior were surveyed. Here, important effects were detected regarding system voices and their influences on users. The variation of the voices' gender, speed, volume, manner etc. e.g. [9–11] has different impacts. For example, a female voice helps to communicate emotional content, whereas male voices tend to sound competent and convey task-related information [9]. The manner can help to increase interaction success, e.g. when motivational feedback is provided [10, 11]. In comparison of a human-sounding voice with a computerized one, users significantly prefer the human-sounding voice; even learn faster while solving a task [10]. These findings prove that a system voice indeed is able to affect the interaction. Furthermore, users' personality characteristics strongly influence the perception of speed and volume of system voices. For example, introvert users prefer low speed and volume; extrovert users sympathize with louder and even exaggerated tones [12]. Therefore, user characteristics also have be taken into account when examining the effect of different voices on users.

Before an explanation of the intended research goal, it still needs to be clarified when an interaction between a Companion-system and a user can be labeled as successful. For this purpose, Frommer et al. [13] developed a Wizard of Oz (WOz) experiment where users had to interact with a simulated Companion-system (description follows below) while solving a task. An artificial, computerized voice was chosen for guiding users through this experiment. During the interaction, challenges occurred at specific stages, which demanded the adaption of current task-solving strategies from the users [5].

Quantitative as well as qualitative methods were used to analyze users' perceptions, their interaction behavior as well as user characteristics. This research process established the basis for the forthcoming study introduced here. Individual user characteristics (e.g. personality traits, socio-biographic variables or technical experience) influence actual user behavior directly and have to be taken into account while analyzing data of users of technical devices [14]. User characteristics were shown to influence users' (task-)performance, especially during situations that were perceived as challenging. Participants with greater performance "were younger, more experienced with computers, showed lower amounts of neuroticism and higher amounts of agreeableness (NEO-FFI) on average." [15]. Furthermore, the analysis of semi-structured interview material showed the importance of the subjective experience of users while interacting with the artificial speech-based system. It became obvious that users tended to anthropomorphize the system, even if it's just a voice and a screen [16]. But this is not necessarily linked to comfortable feelings in the interaction. In fact, the artificial voice is associated with feelings like anxiety or scariness and with the tendency to distance from the system by reducing initiative. Hence, wishes for a change of the artificial voice into a more human-like one occurred, maybe as a result of imagining a deeper and more trustworthy relationship including a more comfortable interaction atmosphere with such a system voice [16, 17]. Furthermore, participants used more negative attributes for the description of the voice than neutral or positive attributes [18].

To deal with these findings, the aforementioned experiment was modified and the application of two different voices was chosen: an anthropomorphic voice compared with an artificial one. The psychological research goal is to find out which effects the voice has on users' perception of its support while solving the task. This study shall also evaluate which kind of voice is most likely to evoke positive affect and greater perceived trust in users. Furthermore, we want to survey the influence of user characteristics on the voice preference.

3 Methods

The aforementioned WOz experiment represents a suitable approach for our research. Before we explicate the modifications and hypotheses, we will give a short description of this experiment and the LAST MINUTE corpus, which is the result of previous research.

3.1 Wizard of Oz Experiment LAST MINUTE

The WOz experiment and the resulting LAST MINUTE corpus were developed as a research tool to investigate subjects during an interaction with a speech-based interactive dialog system, including a problem-solving task with planning, re-planning and strategy change [15]. All tasks had to be solved by users with the help of a solely speech-controlled computer system. In accordance with the central design feature of WOz experiments, this system was controlled by hidden human operators. The subjects believed they communicate with an autonomous computer-system. A male sounding, clearly computerized voice (MARY TTS, mary.dfki.de) was chosen to reinforce the feeling of interacting with a computer system [5].

According to Frommer et al. [13] as well as Rösner et al. [5], the experiment was executed as described in the following: At first, the system introduced itself and asked some personal questions, the so called personalization module. The system explained users that this information is needed for individual adaptation. After that, the actual last minute module [5] began with the explanation of the task. Subjects had to pack a suitcase for a suggested summer vacation for fourteen days in a predefined time. They were informed that detailed weather information will be gathered and provided later. Participants could choose items out of twelve categories (e.g. tops, shoes, accessories), which were presented in a predefined order on a screen in front of them. This stage is called “baseline” (BSL). Within the interaction course, particular restrictions, namely challenges, occur. The first of these challenges is called “weight limit barrier” (WLB). Here, users were informed that their suitcase is confined by the airlines’ weight limit. New items could be added only when others were unpacked before. As a result, participants had to adapt to this unexpected condition and to cope with their possibly emerging stress. After they passed more than half of all categories the final information regarding the destination was revealed. The vacation resort was located in the southern hemisphere where the seasons are switched. Now, subjects had to pack for cold climate, which means they had to change their strategy. This challenge is called “weather information barrier” (WIB). Apart from time and weight restrictions, this rendered the packing process even more complicated. In this situation, about half of participants got an empathic intervention inviting them to express their actual feelings. The remaining time could be used for correction, and is called “revision stage” (RES). In the end, participants had the chance to explicate how satisfied they were with the content of their suitcase [5, 13].

3.2 Modification of the Wizard of Oz Experiment LAST MINUTE

For the purpose of our prospective research we modified the established WOz experiment to focus on users’ perceptions of the system voice. Particular attention is paid to users’ individual ratings of the system and its voice as well as possible changes in users’ affective states during the course of interaction.

With respect to prior results, we modified the personalization module to avoid primary uncertainty regarding the system and the interaction [16] and to strengthen the sympathy of users towards the system in the beginning. The intervention was removed because of its indistinct effects [19].

There will be two experimental groups: One half of the participants will interact with the artificial voice which was already used in the prior experiments; the other half will interact with an anthropomorphic voice (IVONA TTS, www.ivona.com). We paid attention to use male voices to avoid the aforementioned gender effects. The setting stays equal for both groups. We expanded the experiment with two rating phases to gather information of actual user conditions and to detect significant changes during the interaction. Altogether, we survey users’ conditions and perceptions in three particular experimental phases as described below (also Fig. 1).

The first rating occurs before the start of the experiment. Here, we survey general information about the user, like socio-biographic variables and experience with technical devices. Furthermore, we measure users’ task-related motivation (Achievement Motives Scale, AMS) [20] and their actual affective state (Positive and Negative Affect Schedule, PANAS) [21]. This rating phase represents the *Baseline* (see Fig. 1) for further points of measurement. The actual experiment begins after this phase.

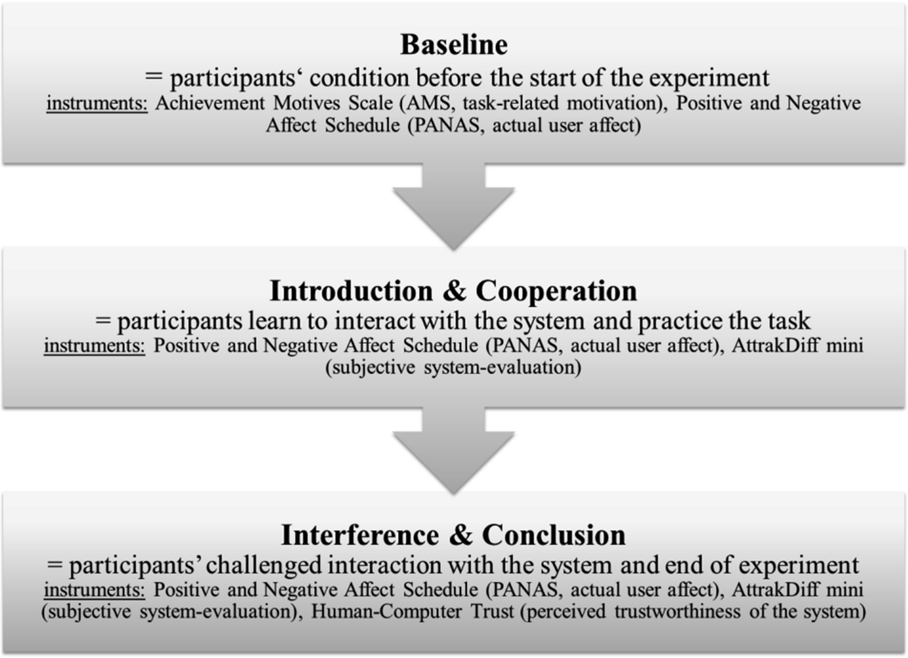
Baseline
= participants‘ condition before the start of the experiment
instruments: Achievement Motives Scale (AMS, task-related motivation), Positive and Negative Affect Schedule (PANAS, actual user affect)
Introduction & Cooperation
= participants learn to interact with the system and practice the task
instruments: Positive and Negative Affect Schedule (PANAS, actual user affect), AttrakDiff mini (subjective system-evaluation)
Interference & Conclusion
= participants' challenged interaction with the system and end of experiment
instruments: Positive and Negative Affect Schedule (PANAS, actual user affect), AttrakDiff mini (subjective system-evaluation), Human-Computer Trust (perceived trustworthiness of the system)

Fig. 1. The three rating phases during the modified WOz experiment

Participants pass through the personalization module to get to know the system (*Introduction*), and immediately start with the last minute module. Here, users are enabled to practice the task while packing the first three categories (tops, jackets & coats, trousers & skirts) (*Cooperation*). After finishing the third category, the second system rating occurs. Here, we measure the actual affective state (PANAS) again and also the subjective system-evaluation by using a shortened version of the AttrakDiff (AttrakDiff mini) [22], which quantifies hedonic and pragmatic product quality. This rating happens aside the experimental screen, which ensures objective appraisal by participants without the effect of politeness towards the system [23]. After that, the last minute module continues. During this last phase, participants have to face all challenges (WLB, WIB and RES, see Sect. 3.1) (*Interference*) and to finish the task (*Conclusion*). The third and last rating occurs after the system-initiated goodbye. The applied questionnaires gather information about users' present affective state (PANAS), final subjective system-evaluation (AttrakDiff mini) and the perceived trustworthiness of the system (Human-Computer Trust) [24].

Therefore, we have three particular rating phases during the experiment: the Baseline, the Introduction & Cooperation as well as the Interference & Conclusion (as shown in Fig. 1). By doing so, we want to evaluate possible changes in users' emotional states during these phases. Furthermore, we want to survey differences in the subjective system-rating and perceived trustworthiness of both groups. A comparison of all ratings between the experimental groups may offer a profound basis for reaching the intended research goals.

With respect to users' individuality, all participants will answer open questions regarding their subjective experience of the system's voice, including possible influences on their feelings and behavior during the experiment as well as possible ideas regarding a change of the voice. Furthermore, some questions refer to users' ascriptions to the system [16, 25] as well as users' experiences of the relationship between themselves and the system.

We will also gather information about specific user characteristics in a second, separate session. Standardized psychological questionnaires are used to gather information about users' affinity towards technology, emotion regulation, personality dimensions, coping with stress, self-efficacy, locus of control in the usage of technical devices as well as the psychological concept of the individual need to evaluate. This information may help to classify the different reactions and perceptions into distinct groups of users.

3.3 Hypotheses

This design aims at the evaluation of the perception of trust and cooperation between user and Companion-system by means of a variation in system's speech-output. More precisely, we survey the impact of an anthropomorphic system-voice compared to an artificial system-voice on users' actual affect, system evaluation as well as the development of trust. Furthermore, this design serves to detect possible correlations between the perception of the system voice and specific user characteristics, e.g. gender, personality dimensions or affinity towards technology.

With regard to the previous explanations, we suppose that the anthropomorphic voice has a more positive influence during the interaction with the simulated Companion-system, compared to the artificial voice. The anthropomorphic voice may increase users' perceived trustworthiness. Furthermore, the possibly cooperative relationship between user and simulated Companion-system will be influenced during the phase of Interference & Conclusion to an unacquainted extent.

Hence, several hypotheses were formulated:

1. Regarding the phase of Introduction & Cooperation, we expect more positive affect (PANAS) and a higher system rating (AttrakDiff mini) of those users who interact with the anthropomorphic voice in comparison to the other group.
2. Both experimental groups have to face the barriers during the phase of Interference & Conclusion and will show lower system-rating (AttrakDiff mini), compared to the phase of Introduction & Cooperation.

Especially the change between cooperative interaction (second rating phase) to possibly interfered (or even failed) interaction (third rating phase) seems interesting. But here, we can just formulate explorative questions:

3. Will significant differences occur between the two groups regarding their perceived trustworthiness and subjective system-evaluation for the third rating phase?
4. Will significant differences occur between the two groups regarding users' affective state for the second and third rating phase?

5. Are there significant influences of user characteristics on the following goal criteria: perceived trustworthiness, subjective system-evaluation and users' emotional state?

The human-like voice may rather evoke the assumption of competence in users which possibly can or cannot be satisfied during the interaction. Trust issues and higher levels of negative affect may be the result. As mentioned before, the perception of system-voices is strongly influenced by user characteristics. Even if we suppose that the anthropomorphic voice may evoke more positive affect in general, individual preferences and perceptions have to be taken into account, too. Therefore, it seems possible for some users that they perceive the artificial voice as less competent, and thus may more likely forgive mistakes.

Besides these assumptions, the design of the study shall help to get a profound understanding of the effects of several user characteristics (e.g. personality dimensions, coping with stress, motivation, self-efficacy) on the preference of a specific system voice.

4 Outlook

The experiment takes place in a research lab of the Otto von Guericke University Magdeburg. A small sample of six participants already passed a test phase. The first (not systematically analyzed) results show that they indeed experience different, albeit marginal affective states during the interaction. Of course, we will need a greater sample size to support our hypotheses. We plan experimental group sizes of about 30 participants for statistical evaluation of all measurements. In order to reduce influences based on participants' age or gender, both experimental groups will be homogeneous regarding these characteristics (only students aged 18 to 28, gender balanced).

The inclusion of actual affective user states and a profound understanding of users' subjective perceptions of UCI are required for the development of Companion-systems, which shall be experienced as supportive, empathic and trustworthy partners by their individual users.

Acknowledgements. The presented study is performed in the framework of the Transregional Collaborative Research Centre SFB/TRR 62 "A Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). The responsibility for the content of this paper lies with the authors. We are grateful for cooperation in this study regarding technical implementation, modification and computer science aspects to the research group of Dietmar Rösner, including Rico Andrich and Rafael Friesen. Furthermore, we want to thank Ralph Heinemann and Michael Tornow for the additional assistance regarding technical implementation of the experimental setting.

References

1. Wendemuth, A., Biundo, S.: A companion technology for cognitive technical systems. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) COST 2102. LNCS, vol. 7403, pp. 89–103. Springer, Heidelberg (2012)

2. Karitnig, A.: Analyse von künstlichen und natürlichen Sprachausgabesystemen im Smart-Home-Bereich. In: Hitz, M. Leitner, G., Kruschitz, C. (eds.) HASE 2010 – HCI Aspects of Smart Environments, pp. 29–38. Klagenfurt (2010). http://www.uni-klu.ac.at/tewi/downloads/HASE10_Conference_Proceedings.pdf
3. Suzuki, N., Katagiri, Y.: Prosodic alignment in human-computer-interaction. Connect. Sci. **19**(2), 131–141 (2003)
4. Waytz, A., Heafner, J., Epley, N.: The mind in the machine: anthropomorphism increases trust in an autonomous vehicle. J. Exp. Soc. Psychol. **52**, 113–117 (2014)
5. Rösner, D., Frommer, J., Friesen, R., Haase, M., Lange, J., Otto, M.: LAST MINUTE: a multimodal corpus of speech-based user-companion interactions. In: Calzolari, N. (Chair), Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), p. 96. European Language Resources Association (ELRA), Istanbul, Turkey (2012)
6. Picard, R.W.: Affective Computing. MIT Press, Cambridge (1997)
7. Wolff, S., Kohrs, C., Scheich, H., Brechmann, A.: Temporal contingency and prosodic modulation of feedback in human-computer interaction: effects on brain activation and performance in cognitive tasks. In: Heiß, H.-U., Pepper, P., Schlingloff, H., Schneider, J. (eds.) Informatik 2011, Berlin, GI-Edition. LNI, vol. 192, p. 238. Koellen, Bonn (2011)
8. Hudlicka, E.: To feel or not to feel: the role of affect in human-computer interaction. Int. J. Hum Comput Stud. **59**, 1–32 (2003)
9. Nass, C., Moon, Y.: Machine and mindlessness: social responses to computers. J. Soc. Issues **56**(1), 81–103 (2000)
10. Wolff, S., Brechmann, A.: Carrot and stick 2.0: the benefits of natural and motivational prosody in computer-assisted learning. Comput. Hum. Behav. **43**, 76–84 (2015)
11. Partala, T., Surakka, V.: The effects of affective interventions in human-computer interaction. Interact. Comput. **16**, 295–309 (2004)
12. Nass, C., Lee, K.M.: Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. J. Exp. Psychol. **7**(3), 171–181 (2001)
13. Frommer, J., Rösner, D., Haase, M., Lange, J., Friesen, R., Otto, M.: Project A3 prevention of negative courses of dialogues: wizard of Oz experiment operator's manual. Working Paper of the Collaborative Research Project/Transregio 62 "A Companion Technology for Cognitive Technical Systems". Pabst Science Publication, Lengerich (2012)
14. Haase, M., Lange, J., Frommer, J.: Eigenschaften von Nutzern in der Mensch-Computer-Interaktion. In: Peters, S. (ed.) Die Technisierung des Menschlichen und die Humanisierung der Maschine: Interdisziplinäre Beiträge zur Interdependenz von Mensch und Technik. Mitteldeutscher Verlag, Halle (Saale) (2015)
15. Rösner, D., Haase, M., Bauer, T., Günther, S., Krüger, J., Frommer, J.: Desiderata for the design of companion systems. KI - Künstliche Intelligenz **30**(1), 53–61 (2016)
16. Krüger, J., Wahl, M., Frommer, J.: making the system a relational partner: users' ascriptions in individualization-focused interactions with companion-systems. In: Berntzen, L., Böhm, S. (eds.) Proceedings of the Eighth International Conference on Advances in Human Oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC 2015), pp. 47–53. IARIA (2015). http://www.iaria.org/conferences2015/CfPCENTRIC15.pdf
17. Frommer, J., Rösner, D., Andrich, R., Friesen, R., Günther, S., Haase, M., Krüger, J.: LAST MINUTE: an empirical experiment in user companion interaction and its evaluation. In: Companion-Technology: A Paradigm Shift in Human-Technology Interaction. Springer, Heidelberg (in press)

18. Lexow, A., Andrich, R., Rösner, D.: LAST MINUTE: User perception of the computer voice. In: Biundo-Stephan, S., Rukzio, E., Wendemuth, A. (eds.) Proceedings of the 1st International Symposium on Companion-Technology (ISCT 2015), Ulm, pp. 137–142 (2015). http://vts.uni-ulm.de/doc.asp?id=9771
19. Wahl, M., Krüger, J., Frommer, J.: From anger to relief: five ideal types of users experiencing an affective intervention in HCI. In: Berntzen, L., Böhm, S. (eds.) Proceedings of the Eighth International Conference on Advances in Human Oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC 2015), pp. 55–61. IARIA (2015). http://www.iaria.org/conferences2015/CfPCENTRIC15.pdf
20. Lang, J.W.B., Fries, S.: A revised 10-item version of the achievement motives scale: psychometric properties in German-speaking samples. Eur. J. Psychol. Assess. **22**(3), 216–224 (2006)
21. Krohne, H.W., Egloff, B., Kohlmann, C.-W., Tausch, A.: Untersuchungen mit einer deutschen Version der "Positive and Negative Affect Schedule" (PANAS). Diagnostica **42**, 139–156 (1996)
22. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Szwillus, G., Ziegler, J. (Hgg.) Mensch & Computer 2003 (Berichte des German Chapter of the ACM), Bd. 57, S. 187–196. Vieweg + Teubner Verlag, Wiesbaden (2003)
23. Reeves, B., Nass, C.: The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press/CSLI, New York (1996)
24. Madsen, M., Gregor, S.: Measuring human-computer trust. In: Gable, G., Vitale, M. (eds.) 11th Australasian Conference on Information Systems, vol. 53, pp. 6–8 (2000)
25. Krüger, J., Wahl, M., Frommer, J.: Users' relational ascriptions in user-companion interaction. In: 18th International Conference on Human-Computer Interaction, 17–22 July, Toronto, Canada. LNCS. Springer, Heidelberg (accepted)