Evaluating an adaptive dialog system for the public

Benjamin Weiss, Simon Willkomm, Sebastian Möller

Quality & Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany
Benjamin.Weiss@tu-berlin.de

Abstract

An embodied conversational agent was developed implementing four strategies to adapt to the user: User tracking and recognition with a camera, remembering interests in topics, remembering preferences concerning the level of detail of information, and changes in confirmation strategy. The agent was integrated into a system providing public information on ICT related projects of research and development for visitors of the laboratories. In an interactive experiment, the adaptive version was compared to a non-adaptive version. The logging data, but not the questionnaire data, shows significant differences, indicating a benefit of the adaptive version in terms of efficiency in interaction. The logging data and results from a final interview are discussed in relation to other work on this subject, concluding on the difficulties to provide not only more efficient interaction, but also higher User Experience.

Index Terms: evaluation, spoken dialog system, logging data

1. Introduction

When designing interactive systems, variability in user requirements, preferences, and expectations is increasingly taken into account. Two approaches are common to build flexible systems, namely systems that either can be configured by hand or that automatically adapt to the individual user. However, to actually configure a system a user has to have at least some expertise and motivation, which can not be taken for granted in public infrequent application scenarios. In contrast to this, adaptive systems are in principle suited to serve all types of users, as the adaptation should require no additional effort for dealing with a flexible system [1].

This assumption, however, only holds for the case of adaptations properly conducted. Meaningful automatic changes in the behavior of an interactive system can initiate positive User Experience and results in the user impression of a valuable and intelligent system. But if the user does not experience meaningful changes in the system's behavior – i.e. no contextual reason for it or no benefit in it – this might result in irritation, increased cognitive load, and negative User Experience. Additionally, for the specific area of the use in public spaces the time of exposure of the user to the system might be too short for either the user to realize the adaptation of the system, or for the system to obtain data to decide on preferences of one user.

In contrast to sophisticated machine learning approaches like reinforcement learning, which could in principle also be used for individual adaptation [2], we present an easy applicable rule-based approach to design an adaptive interactive system for the public. The aim is to evaluate, if a public spoken dialog system (SDS) can benefit from automatic adaptations, examining various dimensions of subjective quantitative data as well as logging data from an interaction experiment.

2. Related work

Adaptive SDS subsume a variety of possible sources of context information as well as resulting changes in the system behavior. For the domain of telephone-based information retrieval (such as train schedule), user models has been built to solve challenges to automatic speech recognition (ASR) based on training data [1]. The system may change its dialog strategy from user initiative to system directed and from no confirmation to implicit or even explicit confirmation dependent on ASR errors. Currently, such kind of adaptivity seems to represent the state-of-the-art of dialog design when taking common textbooks [3, 4] or the current VoiceXML standard into account. In difference to manually implementing such an adaptation strategy, however, is that in [1] a machine learning algorithm was used to derive the mechanism triggering the changes in dialog strategy. Their subsequent evaluation comparing two systems with 12 users (within-design) revealed a significant benefit for the adaptive version concerning task success, but only a trend for the questionnaire assessing usability and for metrics derived from logging data.

While using interaction data to sort users into three levels of expertise, the voice operated adaptive email system presented in [5] provides different levels of information for each user group. The data used for classifying user expertise include, e.g., number of help requests or time to trigger a no-input, and is also sensitive to information the system has already given to a specific user in previous turns. An evaluation of this system was carried out with respect to age as well as three levels of expertise of the 24 subjects, but did not include a baseline system without adaptivity for comparison.

SDSs designed for the public do have additional issues to deal with. For example, a system placed in a museum to inform about computers and technology, was revised to interact not only with staff, but also visitors, especially children [6]. Apart from an emphasis on very good microphone set-up required, the main issue was a robust speech understanding for the variability of users. Additionally, a system designed for use in museum environment needs to be engaging and actually teach the visitors or increase their interest, which differentiates it from a dominant class of task-related systems. Therefore, like the system presented in this paper, the SDS was embodied with an audio-visual speech synthesis (TTS) using visual information (articulation, gaze and gesture). A summative evaluation using observation, interviews and questionnaires confirmed reaching the goal of an engaging system that increases knowledge and interest in the presented topics.

3. A public information system for visitors

VirtualK is an an embodied information system which informs visitors from varying backgrounds (students, colleagues from academia and industry, Berlin visitors of the annual "long night of the science") about ongoing and past research and development projects. It finally aims at motivating the users to try out demonstrators on their own. Our SDS is based on the Thinking Head system [7]. It is a modular, event-based framework, and exhibits an audiovisual TTS. The TTS "OpenMary" [8] and a German ASR (Sphinx) using push-to-talk (see Figure 2) were integrated. The dialog was defined in VoiceXML running with Optimtalk [9]. The description of an older version was presented in [10].

The audiovisual TTS with a male visualization was chosen to attract attention of visitors in order to initiate interaction and increase User Experience when interacting with spoken language as complementary to the text and pictures presented with the project poster descriptions (see Figure 1). VirtualK gives visual conversational feedback, i.e. a nod signals the processing of a user utterance and if the user utterance is not recognized, VirtualK will close his eyes and stop/pause the conversation. Content-wise, VirtualK is able to provide more project related information than the demonstrators and posters.

A webcam is used to detect a user within the interaction sphere of the system, and VirtualK will open his eyes and initiate the dialog with general information, and by asking the user about the interest in one of four research fields (video, audio, smartphone apps, or mobile interfaces). The system provides project-related information either by project name or by suggesting a project based on the preferred topic (audio: music, communication; video: quality, mobile TV; apps: phone control and leisure time; mobile interfaces: security, cross-service). If a face is not recognized for 20 video frames, VirtualK will again close his eyes. Figure 3 shows a simplified structure of the dialogue, the points are representing the decision nodes: If no project is explicitly chosen by a user, it is chosen by the system from the user defined topic.

For each project, there are two levels of information (and if available, using a demonstrator is offered): General description and additional information. After each block of information presented, the system asks whether it should proceed or not.

The authors of [11] differentiate between adaptations centered on the user, the SDS and the environment, namely dialog-behavioral and emotional adaptation, speech adaptation, or event, device and task adaptation, respectively. Adopting their terminology, we consider dialog-behavioral, event and task adaptations as result of our requirement analysis for the public information system VirtualK. The main aim implementing these types of adaptations in our system is to enable a more efficient dialog by removing unnecessary turns and thus enhancing the User Experience.

There are four kinds of adaptation strategies implemented, separating this from a non-adaptive version:

User recognition: With a web-cam a photo is shot for later recognition of the face of the user (based on OpenCV [12]), so users can interrupt/pause the dialog, even try out an demonstrator, and come back to continue the interaction. The user is informed about this and asked for her/his name to address the user properly after a break. A user model is provided to allow to pause user tasks and enable user adaptations. This is a key feature for multi-party public interfaces, aiming to reduce redundancy and providing positive User Experience by exploiting face recognition technology at hand.

Remembering interests: After finishing the presentation of a project, the non-adaptive version of VirtualK asks the user again which field and topic he/she is interested in. The adaptive version memorizes the previously chosen field and topic of the presented project and gives the user the choice to hear information about another project of the same field and topic, to change the field or to change the topic. If the user wants the system to propose a project more often than specifying his interests to narrow down the choices, VirtualK memorizes this and will directly propose a project after the user returns from a demonstrator. With this user adaptive dialog strategy, a more natural and efficient interaction is anticipated in order to adapt to strategic preferences of the user.

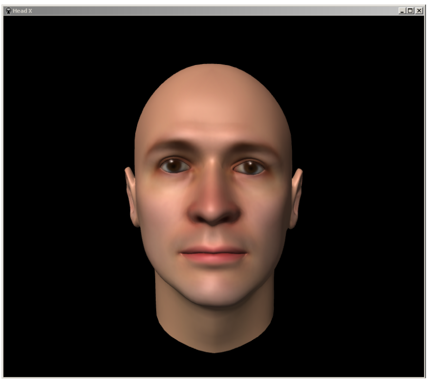
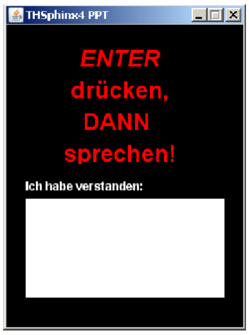


Figure 1: *Graphical representation of VirtualK.*



THSphinx4 PPT

ENTER drücken, DANN sprechen!

Ich habe verstanden:

Figure 2: *The Push-To-Talk button.*

Confirmation strategy: The non-adaptive version of VirtualK always confirms the voice entry from the user with an "echo" confirmation, repeating what it understood. The default setting of the adaptive version is also set to "echo". After a negation of this "echo" from the user or a no-match the confirmation strategy is changed to "explicit". If an utterance was understood, the confirmation is set from "explicit" to "echo" or from "echo" to "no confirmation". The implemented changes for the adaptive version are chosen to be rapid in order to provide one feature to be experienced often, as the adaptations of the previous version have not been very obvious to the test users [10].

Level of project details: Whereas the non-adaptive version always asks the user to proceed to the next level of detail when presenting project information, the adaptive version logs user answers and automatically sticks to the user choices, if there is one more choice for additional information or not. This feature is similarly motivated as the "remembering interests" feature above is.

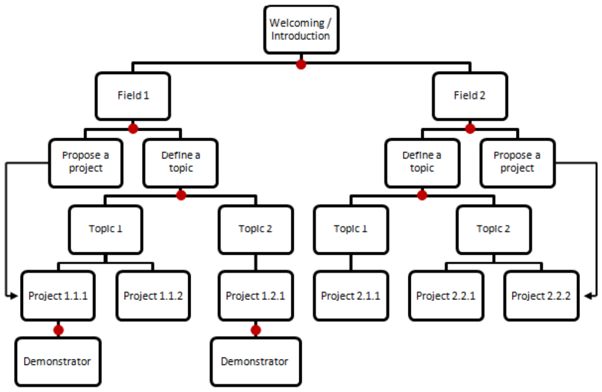


Figure 3: *The simplified dialog structure. Red dots indicate a user decision node.*

4. Evaluation setup

For the purpose of evaluation, the face recognition was set to a maximum by deleting previous users at the start of each experimental trial. For continuous duty, we lack information of the number of visitors a day, but it is expected to "forget" users after about four hours in order to successfully discriminate users. Also, the four research field were split into two categories, each comprising about half of the projects and demonstrators available to avoid boredom when trying out the system repeatedly.

A total of 30 test subjects took part in the experiment, gender balanced (14 female, 16 male), aged between 20 and 43 (average 26.4). All were paid for their contribution.

The experimental design was a 2x2x2 mixed factorial design with *version* (adaptive, non-adaptive) as the within-subjects variable, *order* of the project category (first Category A, second Category B (AB) and vice versa (BA)), and order of the version (first adaptive version, second non-adaptive version (an) and vice versa (na)) as the between-subjects variables.

All users successively interacted with both versions of VirtualK. They were asked to inform themselves about three to four projects and try out at least one demonstrator. Each individual experimental session took about one hour with roughly 15 min. for each interaction.

After each trial the test subjects answered a questionnaire comprising aspects of User Experience [13, 14] to subjectively test for a benefit of the adaptations. Finally an interview was conducted.

5. Results

Logging data of the individual interactive sessions is analyzed in order to obtain information on whether the adaptive version of VirtualK could improve efficiency of the interaction and engagement of the users or not. As the system occasionally missed recognizing a face, mostly due to issues with the lighting and background, and thus also stopped logging during the interaction, data from only 21 participants was analyzed (12 men, 9 women).

Using Wilcoxon rank sum test (equivalent to the Mann-Whitney test) there are no differences between both versions concerning the number of no-inputs ($V = 106.5, p = 0.66$) and no-matches ($V = 74.5$, $p = 0.42$), which corresponds to the non-significant result for the users' impression of ASR performance, asked with a 5-point Likert scale ($V = 75.5, p = 0.69$).

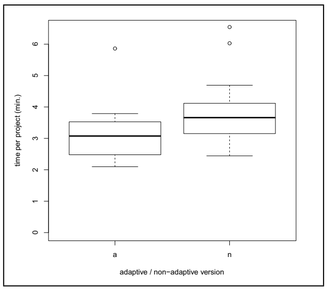


Figure 4: *The interaction time per project.*

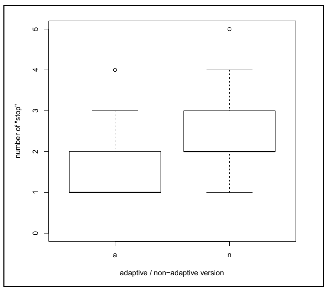


Figure 5: *The number of "stop" commands.*

However, there are two significant results for the logging data: The elapsed time per project is shorter for the adaptive version ($V = 45$, $p = 0.012$, see Figure 4), although the number of projects asked for is not significantly different ($V = 96.5$, $p = 0.64$). The origin seems to be the shorter overall amount of time spend interacting with the adaptive system ($V = 45$, $p = 0.12$).

Also, the number of "stops" of more detailed information of one project is lower during interaction with the adaptive version ($V = 17$, $p = 0.039$, see Figure 5). The related parameter, the number of "go on" with more detailed information, is higher for the adaptive version, but not significantly so ($V = 100$, $p = 0.10$, see Figure 6).

For the questionnaire data there are no significant differences between adaptive and non-adaptive version.

In a final interview after both interactions the participants were asked several questions about their awareness of system differences and preferences. Eleven out of 30 users stated differences between both trials. When explicitly stating the four kinds of adapting, all users could identify some of those (confer Table 1 for the details).

Table 1: *Number of test subjects remembering and identifying the four adaptation strategies.*

Type of adaptation	remembered	identified	no differences
user recognition	11	19	0
remember. interests	5	24	1
confirmation strat.	4	17	9
level of detail	0	10	20
any on the four	11	30	0

Finally, a five-point Likert skale was used to assess if the adaptations have been supportive or obstructive during the interaction (1 for very supportive, 5 for very obstructive). The

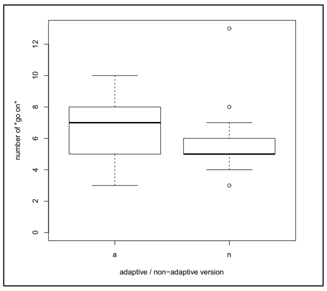


Figure 6: *The number of "go on" commands.*

mean over all 30 users is 1.8, which is significantly better than the scale's mean of 3, suggesting a supportive retrospective impression from the users ($V = 9.5$, $p < 0.001$). When asked specifically, 23 of the participants stated to prefer the adaptive version, four the non-adaptive version, and three would not decide on one. The reasons given by the users for preferring the adaptive one are the following (number of citations in brackets): User recognition (8), fewer repetitions (8), more human (7), remembering interests (6), confirmation strategy (5), better dialog structure (5), less time-consuming (5), more trustworthy (5), more personal (4), more intelligent (3), easier communication (3).

6. Discussion

The logging data indicates a successfully conducted experiment for two reasons: Users did not differ in the number of projects asked for during both trials, and the ASR performance did not show differences neither for the number of barge-in, nor for the number of no-inputs. The last conclusion is confirmed by the user ratings of the ASR quality in the questionnaire.

Concerning the number of projects, there was a small degree of freedom provided with the experimental instruction, that was obviously not used systematically as result from the adaptation: Both versions have a median of five, giving no indication of higher engagement for the adaptive version.

The time users interacted with VirtualK per project is shorter for the adaptive version, indicating an increase in efficiency. Another result is concerned with the degree of project details asked for by the user: The number of "stops" (i.e. requests to stop the project presentation and not to go on with more details) is lower for the adaptive version, suggesting a success of adapting to the user's preferences concerning the level of detail. The number of "go on", when the system stopped the project information is higher, but not significantly, indicating a slight bias towards too early stops.

Interestingly, only a third of the users consciously experienced the adaptation strategies. However, when asking specifically for the different kinds of adaptations, all could identify at least some of the differences between the non-adaptive and the adaptive version. The most salient ones were "user recognition" (i.e. being addressed by name and continuing the dialog after a visit to a demonstrator) and "remembering interests" (stated as topics and field). The least well identified (and actually not remembered) strategy is about the level of detail, even if its relevance is found in the logging data. About the reason for this result can only be speculated at this point: Maybe the resulting changes in dialog are just not salient enough from a perceptual point of view. Another possible interpretation is a minor relevance of the level of detail for the course of interaction, as it requires low effort for the control of the level of detail anyhow. In contrast to this, correct "user recognition" reduces potentially more irritating impediments to the dialog flow and is most likely responsible for the significantly reduced interaction time per project.

The number of identified adaptations is in line with the ranking of reasons cited to prefer the adaptive version: "Recognizing the user" and "fewer repetitions" are both ranked first, whereas the latter could be the consequence of the first (in terms of adaptation strategy). In sum, the named reasons are somewhat surprising, as they include not only efficiency-related aspects, but (in contrast to the lacking results for the questionnaire) also social aspects (e.g., more human). However, this specific result may be biased by the interview structure, as the adaptations implemented had to be presented just before this question.

7. Conclusion

In an experimental interaction the version of VirtualK with four adaptation strategies implemented revealed a higher efficiency concerning interaction time, compared to the non-adaptive version. There is also evidence for higher efficiency in terms of fewer commands to control the level of detail due to the adaptation.

Interestingly, only few users could remember the differences between both system, indicating that about 2/3 did not consciously recognize the adaptation strategies. When told, however, all users could identify at least one of the four. Consequently, there are no differences between both versions according to the questionnaire data. Finding an effect for logging data, but not questionnaires is similar to the results in [1] and also in line with a smaller evaluation of a previous version of VirtualK [10].

Apparently, adapting a spoken dialog system to the user results in a more efficient interaction, even for short conversations of unknown persons (i.e. new users) with little time of the system to decide on adaptation. Still, the overall User Experience assessed by means of a questionnaires does not reflect this benefit. Either the method to use questionnaires for assessing the experienced benefit of such adaptations is not valid in an laboratory experiment, or the observed increase in efficiency is not relevant to a user in such an infrequently (or unique) and short interaction. The former has to be tested by a field test.

In future versions, adaptations using information from more elaborate user models will be implemented and tested. Already, an age group and gender recognition is implemented based on visual and acoustic pattern matching. But this feature has not yet been evaluated, as the strategy to use such information to adapt wording and dialog strategy has still to be developed.

One major aim of studies like the one presented here is to obtain more insight into the relation between logging data and aspects of User Experience in order to complement comparative user tests with a non-adaptive version at all. Based on the result presented, the expected benefit especially for multi-party public interaction is questionable, though, indicating a surprising lack of relationship – albeit for the laboratory condition.

8. Acknowledgements

This work was financially supported by the European Institute of Technology (ICT-enabled Human Activity).

9. References

[1] D. J. Litman and S. Pan, “Empirically evaluating an adaptable spoken dialogue system,” *User Modeling and User-Adapted Interaction*, vol. 12, pp. 111–137, 2002.

[2] O. Lemon and O. Pietquin, Eds., *Data-driven methods for adaptive spoken dialogue systems*, ser. Computational Learning for Conversational Interfaces. New York: Springer, 2012.

[3] M. H. Cohen, J. P. Giangola, and J. Balogh, Eds., *Voice user interface design*. Boston: Addison-Wesley, 2004.

[4] M. F. McTear, Ed., *Spoken Dialogue Technology – toward the conversational user interface*. London: Springer, 2004.

[5] K. Jokinen, K. Kanto, A. Kerminen, and J. Rissanen, “Evaluation of adaptivity and user expertise in a speech-based e-mail system,” in *COLING Satellite Workshop Robust and Adaptive Information Processing for Mobile Speech Interfaces*, 2004.

[6] D. Traum, P. Aggarwal, R. Artstein, S. Foutz, J. Gerten, A. Katsamanis, A. Leuski, D. Noren, and W. Swartout, “Ada and grace: Direct interaction with museum visitors,” in *Intelligent Virtual Agents*, 2012, pp. 245–251.

[7] M. Luerssen and T. Lewis, “Head X: Tailorable audiovisual synthesis for ecas,” in *Interacting with Intelligent Virtual Characters Workshop (IIVC)*, 2009.

[8] 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, pp. 365–377, 2003.

[9] OptimSys s.r.o., “Voice browser.” [Online]. Available: http://www.optimsys.cz/en/products/optimtalk/

[10] B. Weiss and R. Tönges, “Automatic adaption of spoken dialog systems for public and working environments,” in *International Conference on Interfaces and Human Computer Interaction (IHCI), Lisbon*, 2012, pp. 284–288.

[11] T. Heinroth and W. Minker, Eds., *Introducing spoken dialogue systems into intelligent environments*. New York: Springer, 2013.

[12] Intel Corporation, “Open computer vision library.” [Online]. Available: http://sourceforge.net/projects/opencvlibrary/

[13] M. Hassenzahl and A. Monk, “The inference of perceived usability from beauty,” *Human-Computer Interaction*, vol. 25, no. 3, pp. 235–260, 2010.

[14] M. Heerink, B. Kröse, B. Wielinga, and V. Evers, “Assessing acceptance of assistive social agent technology by older adults: the almere model.” *International Journal of Social Robotics*, vol. 2, pp. 361–375, 2010.