Travel Routes or Geography Facts? An Evaluation of Voice Authentication User Interfaces

Alina Hang[1], Alexander De Luca[1], Katharina Frison[1], Emanuel von Zezschwitz[1], Massimo Tedesco[2], Marcel Kockmann[2] and Heinrich Hussmann[1]

[1] University of Munich (LMU), Media Informatics Group, [2] VoiceTrust, Germany

{alina.hang, alexander.de.luca, emanuel.von.zezschwitz, hussmann}@ifi.lmu.de, frison@cip.ifi.lmu.de, {massimo.tedesco, marcel.kockmann}@voicetrust.com

Abstract. Fallback authentication based on voice recognition provides several benefits to users. Since it is a biometric method, there are no passwords that have to be remembered. Additionally, the technique can be used remotely without the user having to be physically present. We performed stakeholder interviews and we iteratively designed and evaluated different voice authentication user interfaces with a focus on ease-of-use. The main goal was to keep embarrassment low and to provide an interaction as natural as possible. Our results show that small changes in the interface can significantly influence the users' opinions about the system.

Keywords: voice user interfaces; voice authentication

1 Introduction

The number of passwords that users have to remember is steadily increasing. The odds of not being able to recall one's passwords are high and often result in bad habits. In particular, if users frequently have to change their passwords, they counteract these problems by writing them down or reusing the same password or variations of it for all accounts [1]. In the last resort, when even those questionable solutions fail (e.g. the user forgot where she wrote down the password), fallback authentication systems are required that enable users to regain access to their accounts. Well-designed fallback systems are important to encounter the mentioned habits.

Most companies rely on helpdesks, which require administrative personnel and that are expensive in costs (between $10 and $15 per transaction [4]). Automatic password reset systems reduce costs and disburden helpdesk staff. There exist a variety of solutions. Wood [10] categorizes them as something you have [7], something you know [9] or something you are. Systems in the latter category usually rely on biometrics. They require tokens of authentication that are tightly coupled to a person.

Voice authentication systems take advantage of the complex composition of the human speech which creates unique voice characteristics that, in turn, depend on the size of the oral cavity, the throat, nose, mouth and the arrangement of muscles in the

according areas [5]. In combination with voice user interfaces, where human and computers communicate over speech, voice authentication systems are a good basis for automated password resets. Firstly, they are capable of recognizing if the speaking person is authorized to perform a password reset. Secondly, the system can guide the user through the process in a communicative manner, making the need of a helpdesk person obsolete. Finally, they do not require the person to be physically present as voice can be transmitted over distance communication channels like a telephone. Being completely automatic, usability aspects are highly important for such a method.

Naturalness and embarrassment are essential keys for the success of voice user interfaces. Embarrassment refers to the feeling of being uncomfortable caused by unnatural dialogs with a voice user interface. A conversation can be considered as natural if the same wording could result from talking with a human partner in daily conversations.

In the late 90s, Delogu et al. [2] showed that users are willing to put up with lower recognition rates in exchange for higher naturalness. The same study made a comparison between speech recognition and dual-tone multi-frequency (DTMF). Users were more satisfied with the speech recognition technique, highlighting the benefits of plain voice user interfaces. A reason for this can be found in [6] which states that the use of DTMF increases the cognitive load of the users, since they not only have to focus on the instructions on the phone, but also on the buttons they have to press.

In this paper, we present the iterative design and evaluation of enhanced voice authentication user interfaces for password resets. In a pre-study, we found out that current deployments of voice authentication and voice user interfaces work very well from a technical point-of-view. Still, they can be enhanced to increase user satisfaction and ease-of-use. We present two adaptations of a commercially available system and their evaluation. Our results show that both adaptations increase naturalness and reduce the level of embarrassment. We additionally show that feedback in the form of audicons can be helpful and disturbing at the same time, depending on how it is used.

2 Pre-Studies

Most voice-recognition based authentication systems work in at least two phases [3]: During enrollment, the user's voice is recorded and a voiceprint is created to which the user's voice input is compared during authentication. We evaluated a text-dependent commercial system in which the user has to repeat selected word pairs during enrollment and a randomly selected subset of those during authentication. We conducted nine stakeholder interviews and a usability analysis with 18 participants.

2.1 Stakeholder Interviews

We performed interviews in two major companies who employ voice authentication systems for password resets. We decided to interview system administrators as they are hubs of user-reported issues. It has to be noted that this also means that they are only aware of issues that are actually reported and that they may be positively biased to a certain extent as these systems save them a lot of work.

We interviewed nine system administrators (5/4 of each company) with an average age of 38 years (30-43 years). The interviewees reported that some users do not like to perform the password reset in open plan offices, mostly due to embarrassment issues (e.g. the use of word pairs that do not occur in daily conversations). Some users also requested to improve the naturalness of the voice user interface and to shorten the overall password reset process. However, most users are satisfied with how well the voice authentication password reset works in general and how easy it is to learn.

2.2 Lab Study

We recruited 18 participants (7 female) with an average age of 23 years (19-30 years). We invited the participants to our lab to evaluate a currently available commercial system. They were placed in an office-like environment and we recorded all participants with a camera while they completed their tasks. In the first task, participants were asked to enroll for the voice authentication service. They had to call the service and follow the instructions for enrollment. After their voice was registered successfully, we asked them to perform the voice authentication that is needed for password reset (they did not have to actually reset their password).

The results of the usability analysis show that users were positively surprised about how well the voice authentication system worked. As shown in [8], some users are negatively biased against biometrics, which does not correspond to the actual capabilities of those systems. The usability study also confirmed the issues revealed during the stakeholder interviews. Lack of naturalness, embarrassment and temporal demand were the main concerns. The detailed results are outlined in section 5.

3 System Adaptation

Based on the preceding interviews and the lab study, we adapted the commercial system in two different ways. We addressed aspects like the number of voice inputs, feedback, naturalness and embarrassment.

Number of voice inputs: Speech recognition is more reliable if higher numbers of speech samples are available. If the number is too low, users might not be recognized during authentication, while a high number leads to annoyance and frustration due to the longer duration. Thus, a tradeoff between usability and security needs to be taken into account. We reduced the number of voice inputs from four to three, after the consultation of the voice biometrics provider in order to maintain security.

Feedback: We added two types of feedback (in the form of audicons). One feedback was given after successful voice input, the other occurred when the user reached the next level of the authentication process. The goal was to decrease users' insecurity when longer pauses occurred (this is due to the so-called endpointing, during which the end of a voice input is identified, when the pause reaches a certain threshold).

Naturalness: In order to reduce the perceived duration of the password reset, we redesigned the overall conversation style. Therefore, we recorded a new dialog with a new speaker. Particular attention was paid to the right intonation of phrases, i.e. by adding linguistic stylistics. For example, the dialog appears more natural when using

pronouns. Instead of saying "*The enrollment will take five minutes*", the system says "*It will take about five minutes*". Based on the context, the user will still be able to understand the meaning. We also introduced discourse particles into the dialog. These are words with no real semantics, which help to structure the statement (e.g.: "*Well, that's fine*") and helped us to make the overall conversation more natural.

Embarrassment: Most text-dependent authentication systems ask for randomly selected sequences of words to be repeated. Without context, this feels unnatural and leads to embarrassing situations in public or semi-public spaces (imagine someone repeating the word "*cuckoo*" many times incoherently). In the enhanced systems, the users recite fragments embedded in phrases. Furthermore, we focused on more common word pairs used in daily situations instead of word pairs that are rarely used.

3.1 Adaptation 1: Geography Facts

The main goal of this adaptation was to reduce perceived duration of the password reset. Users are not only authenticated by their voice, but they also learn geography facts. This places the authentication process in a more playful setting, leaving the users with the feeling of not wasting their time. Additionally, we used fragments that may also occur in daily conversations, making the overall dialog more natural.

Altogether, we used three different phrases: "*The Danube flows into the Black Sea*"; "*New Delhi is the capital of India*"; and "*Munich is the provincial capital of Bavaria*". Many other facts are possible with this approach as well. This is important to avoid replay attacks (where attackers record the phrase and replay). Figure 1 (left) depicts an example of the dialog for "geography facts".

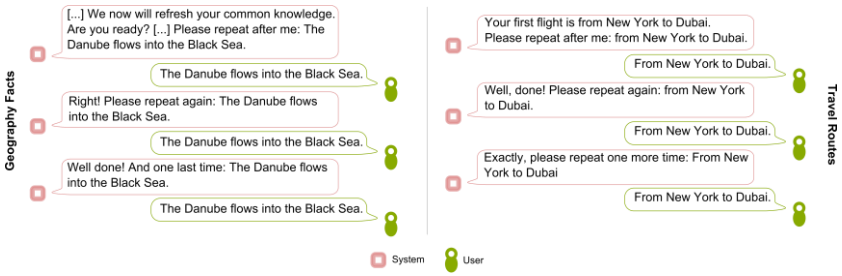

Fig. 1. Dialogs for the systems "geography facts" (left) and "travel routes" (right).

3.2 Adaptation 2: Travel Routes

The main goal of this adaptation was to reduce the level of embarrassment during the dialog with the voice user interface. We placed the communication into a travel context, where users imagine planning a travel route. Again, users had to recite conversational fragments like "*from Munich to Berlin*", "*from New York to Dubai*" or "*from Nuremberg to Hamburg*". This approach allows many different combinations. With only six cities, there are already 6x5 different combinations possible. Figure 1 (right) shows an excerpt from the dialog for "travel routes".

4 User Study

We performed an additional lab study with the adapted commercial system.

4.1 Study Design

We used a repeated measures factorial design. The dependent variables were *system* ("geography facts"; "travel routes") and *feedback* (with and without feedback) resulting in 2^2=4 combinations. The order was counterbalanced to minimize order effects. For all combinations, we analyzed the subjective assessments of users. In particular, we were interested in how well the participants accepted the adapted systems. We also analyzed the influence of feedback on user satisfaction and the assessments of participants who had already participated in the pre-study.

4.2 Study Procedure

We conducted the study in an office-like setting to allow the participants to imagine the environment in which password resets are often done. The participants were informed about the general study procedure and were asked to provide demographic information. After that, the participants had to complete two tasks.

The first task consisted of the enrollment phase. In the second task, the participants tested "geographic facts" and "travel routes" with and without feedback. Additional questionnaires were handed out a) in case the participants tested a system that included feedback and b) when they tested a system for the second time. At the end, they were asked to answer a final questionnaire covering all systems (with and without feedback), including number of voice inputs, naturalness, speech, etcetera.

We recruited 24 participants (8 female). Their average age was 25 years (from 21 to 31). There were two groups of participants. 12 participants who already took part in the first lab study (group A) and 12 new participants (group B). This was desirable as it allowed for comparing the assessments of the different user groups.

5 Results

Duration: As shown in figure 2 (right), reducing the number of voice inputs had a positive influence on the perception of the overall duration. This is not surprising (since the actual duration was shortened). Nevertheless, it is interesting to see whether participants were satisfied with the number of voice inputs or whether they considered it as inappropriate. The differences between group A of the main study and the pre-study are statistically significant as shown by a Wilcoxon test (Z=-2,209, p=0.027). While 42% of group A rated the number of voice inputs inappropriate for the pre-study, only one person maintained this opinion. All other participants (75% of group A) rated the duration as appropriate; almost half of which even found it as very appropriate. Regarding the main study, 67% of all participants considered the number of voice inputs as appropriate. The results are shown in figure 2 (left).

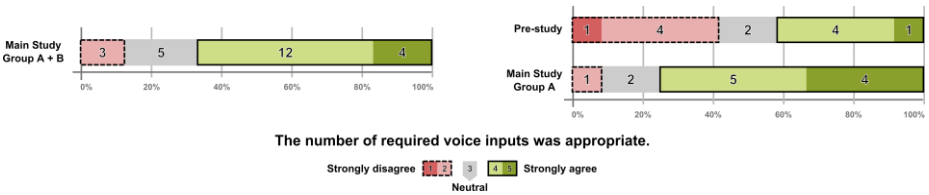

Fig. 2. User ratings of appropriateness of duration in the main study (left). Comparison of the results of group A to the results of the pre-study (right).

Feedback: There were two types of feedback (in the form of audicons), one after successful voice input, the other one when the user reached the next step. While more than 70% of participants considered feedback after successful voice input as important, more than 50% rated the latter form of feedback as obsolete (see figure 3).

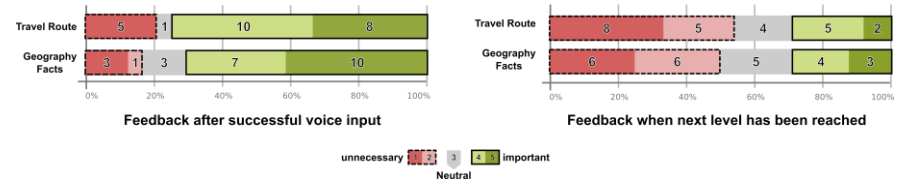

Fig. 3. Importance of the different types of feedback for "travel routes" and "geography facts".

Naturalness: Increased naturalness was one of the main aspects that our participants mentioned in the pre-study. Figure 4 (left) shows the assessment of all participants with respect to naturalness. 75% rated the naturalness as positive. On average, group A perceived the systems as more natural than group B. Regarding the language as well as the voice of the speaker, 92% of all participants found it pleasant, 64% of which even rated it as very pleasant. An overview can be seen in figure 4 (right).

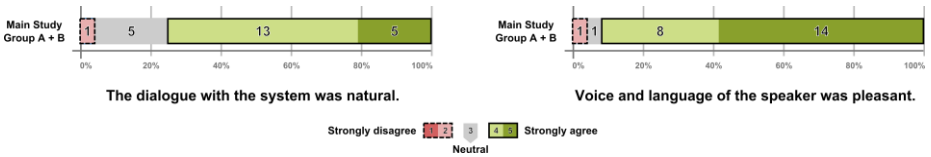

Fig. 4. Naturalness (left) and pleasantness of voice as well as language of the speaker (right).

Embarrassment: Amusement occurs with different notions: It can be positive, increasing pleasure, while a negative touch influences the level of embarrassment. This means that amusement, pleasure and embarrassment are tightly coupled.

One of the main goals of the adapted systems was to enhance naturalness and thus, decrease embarrassment. Figure 5 shows the embarrassment (left) and amusement (right) ratings of group A for each system in comparison to the pre-study. It can be seen that "travel routes" is rated least embarrassing and least amusing, while "geography facts" is also less embarrassing, but with a higher rating for amusement in comparison to "travel routes".

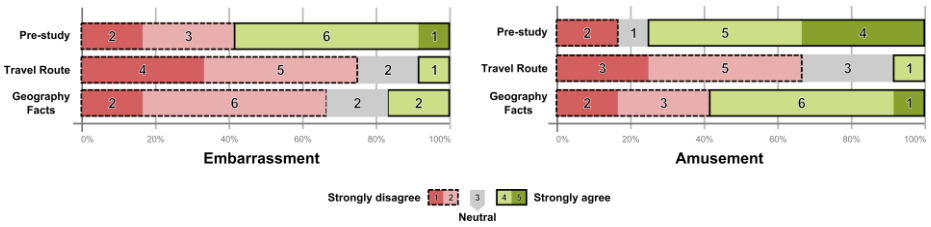

Fig. 5. Level of embarrassment (left) and level of amusement (right) for each system in comparison to the results of the pre-study.

Geography Facts vs. Travel Routes: Finally, we asked the participants to compare the two adapted systems. For this, we developed a small webpage showing a simple square. Each corner of the square represented one system – "travel routes" and "geography facts" with feedback in the left corners; "travel routes" and "geography facts" without feedback on the right corners. We asked the participants to express their preferences by clicking inside the square. The results are shown in figure 6.

15 participants tended to systems with feedback and 9 participants preferred systems without feedback. Comparing the two systems, 10 participants preferred "geography facts" while 12 participants tended towards "travel routes". Interestingly, three participants selected almost the same area, independently from each other.

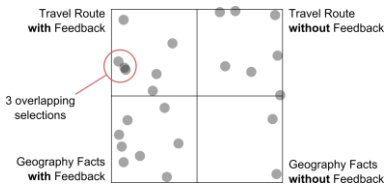

Fig. 6. User tendencies for "travel routes" / "geography facts" with and without feedback.

6 Discussion and Conclusion

The results of the main study nicely show how slightly tweaking a few parameters (e.g. lower number of voice inputs) had a high influence on perceived embarrassment and naturalness when using voice user interfaces.

In particular, "travel routes" was the system that participants assessed as the least embarrassing, e.g. due to the use of neutral word pairs embedded in short phrases. In turn, it was also considered as less amusing than the old system. This is not necessarily a bad thing, since one has to consider the interplay between amusement and embarrassment. While the old system of the pre-study was rated as the most amusing, the embarrassment was almost as high. In this case, amusement has a negative notion and is not desirable. In turn, most participants found the system "geography facts" as amusing and only few rated it as embarrassing. The result highlights the entertaining factor of the system (and also the positive touch that the amusement holds). An interesting question related to this is the influence of entertainment on security awareness,

meaning whether the users are still aware of the seriousness of the situation (e.g. the access to sensitive information) or whether this is clouded by the entertaining factor.

Having a closer look at the preferences, we can see that there are few participants with a strong preference (selections at the one of the four corners) for one of the four combinations. While only few participants were indecisive about their preference, most of them have a tendency towards one of the two systems. For example, one person may like the "travel routes" more, but is not averse to "geography facts" either. Therefore, one could think of enhancing one system with touches of the other (e.g. alternating between "geography facts" and "travel routes" for each level).

Regarding the use of feedback in the form of audicons, figure 6 shows that the opinions are diverse. While some liked having feedback, others preferred less feedback or even almost no feedback at all. Thus, feedback is mainly advisable at the end of a successful voice input, but should be avoided when the next step during the authentication process is reached. One may also think of allowing users to make personal feedback settings during enrollment.

There is still room for additional optimizations. For instance, a participant suggested adding a start signal for speaking to enhance feedback and to reduce insecurities during smaller pauses (which is due the so-called endpointing mentioned earlier). This is an interesting aspect and should be considered in future work. Additionally, a closer look has to be taken at the tradeoff between usability and security.

References

1. Adams, A., Sasse, M.A.: Users Are Not the Enemy. Communications of the ACM. Vol. 42 (12), 40-46 (1999).
2. Delogu, C., Di Carlo, A., Rotundi, P., Sartori, D.: Usability Evaluation of IVR Systems with DTMF and ASR. In Proc. of ICSLP'98, (1998).
3. Dugelay, J.L., Junqua, J.C., Kotropoulos, C., Kuhn, R., Perronnin, F., Pitas, I.: Recent Advances in Biometric Person Authentication. In IEEE Transactions ICASSP'02. Vol. 4, 4060-4063 (2002).
4. Forrester Consulting. Enhancing Authentication to Secure the Open Enterprise. Whitepaper commissioned by VerySign (Symantec), (2010).
5. Jain, A.K., Ross, A., Prabhakar, S.: An Introduction to Biometric Recognition. IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics. Vol. 14 (1), (2004).
6. Lerer, A., Ward, M., Amarasinghe, S.: Evaluation of IVR Data Collection UIs for Untrained Rural Users. In Proc. of DEV'10, 2:1-2:8 (2010).
7. Mannan, M., Barrera, D., Brown, C., Lie, D., van Oorschot, P.: Mercury: Recovering Forgotten Passwords Using Personal Devices. Financial Cryptography and Data Security, 315-330 (2012).
8. Pons, A.P., Polak, P.: Understanding User Perspectives on Biometric Technology. Communications of the ACM. Vol 51 (9), 115-118 (2008).
9. Rabkin, A.: Personal Knowledge Questions for Fallback Authentication: Security Questions in the Era of Facebook. In Proc. of SOUPS'08, 13-23 (2008).
10. Wood, H.M. The Use of Passwords for Controlling Access to Remote Computer Systems and Services. In Proc. of AFIP'77, 27-33 (1977).