7 - 12 SEPTEMBER 2014, KRAKÓW
FORUM ACUSTICUM

Investigating Multimodal Feedback in a Mobile Interaction Paradigm

Julia Seebode, Robert Schleicher, Sebastian Möller
Technische Universität Berlin, Telekom Innovation Laboratories, Berlin, Germany.

Summary
It is well-known that adding feedback to user interfaces is beneficial. But what modalities are best suited for different events and usage situations? Are there advantages or disadvantages of different modalities that can balance each other in multimodal combinations? In this paper we describe a user study that focuses on the interplay of different types of feedback on a smartphone. We used a simple test application where participants received the information about a new picture message and had to type in a 6-digit PIN to unlock the smartphone screen and display the image. In each trial, the application provided one out of four feedback modalities: visual, audio-visual, tactile-visual, and multimodal (as a combination of the three) with different kinds of feedback within these modalities. We collected subjective ratings about the interaction and performance data such as the number of corrections, errors and the duration of entering a PIN. Results indicate that adding another modality to visual feedback is preferred and leads to a higher performance compared to unimodal visual feedback. However, the advantages of adding a modality to visual feedback depend on the actual kind of feedback messages, as additional intrusive sounds were rated similar to visual feedback alone. Moreover, visual-tactile feedback was preferred and gained highest user ratings, followed by multimodal feedback including subtle feedback soundsand vibration patterns. We conclude that especially for sounds used as feedback messages, the actual kind of message and its design has a big impact on user perception and that the combination of sound and vibration leads to a different user perception compared to sound and vibration alone.

1. Introduction

Modern smartphones are equipped with a touch screen, allowing users to interact with a variety of applications running on such devices. While it is possible to take advantage of nearly the whole surface of the phone for displaying the output as well as adjusting the input options according to the application in use - for example by adding a virtual keyboard - new challenges have to be met. The physical keyboard of older phones offered both a tactile and an (unobtrusive) audio feedback when a key was pressed. On a touch screen, the fingertip or a stylus has to be used. In both cases no 'natural' feedback is emitted, but most devices provide the possibility to return visual, audio and even vibrotactile feedback for any user input. Whereas the necessity of some kind of feedback seems not to be questioned in general, we are interested in which kind of feedback exactly is beneficial. Thus, different modalities and combinations hereof have been investigated and compared in the presented study.

In Poupyrev et al. [1] pen input has been augmented with tactile feedback resulting in a higher user performance in a drawing task. Similar results have been found by Brewster et al. [2] for finger-based input. Hoggan et al. [3] compared devices with a physical keyboard, a standard touch screen, and a touch screen with tactile feedback added in both static and mobile environments. They found that the addition of tactile feedback to the touch screen significantly improved finger-based text entry, bringing it close to the performance of a real physical keyboard. Koskinen et al. [4] focused on the pleasantness of tactile feedback; their results also support the advantage of tactile feedback compared to no feedback.

Burke et al. [5] conducted a meta-analysis of 43 studies comparing visual-auditory and visual-tactile feedback to visual feedback alone. Tasks analyzed in those studies varied with regard to type (alert, warning, or interruption; target acquisition, communication; navigation and driving / vehicle operation), complexity (single task or multiple tasks), and workload. The authors found that adding a modality to visual feedback improves performance overall.

Further research by Hoggan and colleagues [6, 7] focused on environmental effects on preference and performance and on the congruence of audio and tactile feedback with the visual appearance of buttons. They found significant decreases in performance for audio feedback at environmental noise levels of 94 dB and above as well as for tactile feedback at environmental vibration levels of 9.18 g/s [6]. They also showed that by choosing congruent sets of audio/tactile feedback for visual touch screen buttons ('soft keys'), users' preconceptions of how the button should feel and sound are met and that the perceived quality of the buttons is improved [7].

Altinsoy et al. [8] investigated the design and interaction issues of auditory and tactile stimuli for touch-sensitive displays and the combined influence of auditory and tactile information (i.e. vibration) on the system quality. They found an advantage of tactile or auditory feedback in perceptual quality and error rate for a number-dialing task compared to no feedback, and even higher ratings for a combination of both modalities.

Even though these studies suggest multimodal feedback to be advantageous compared to visual feedback alone, they give less information of the actual feedback design. Moreover, they disregard aspects of affect and its relation to the general feedback impression as well as user performance. Feedback messages usually are not in the users' focus, but are processed unconsciously. Hence, their first reception concerns mostly their basic emotional connotation, i.e. 'positive or negative' with little conscious reflection [9]. Therefore, besides the functional adequacy also affective aspects are regarded as important for the design of feedback. Hence, this paper addresses the following research questions: What modalities are best suited for different events in a mobile usage situation? Are there advantages or disadvantages of different modalities that can balance each other in multimodal combinations? How should feedback messages be designed to be preferred by users and connoted with a positive affect? To answer these questions, an interaction study was conducted that is described in the following sections.

2. Methodology

To have an easy and controllable interaction, a simple application was created. Use case of this application was to have a locked smartphone screen that can be unlocked by typing in a correct 6-digit PIN. The participants were instructed to wait for a notification on the phone while walking. They received a message informing them about a new picture. To download and see the picture, the currently locked screen had to be unlocked. This is a common means to secure smartphones from malicious use even if it has various disadvantages [10]. Neutral pictures from Berlin sights were used to avoid affective stimulation and thus, bias of the participants.

The study was designed with a secondary task, which is very common for the use of modern smartphones: waiting for a messages and typing on the smartphone whilst walking. As location for the walking task, a long floor in our lab building was chosen. With this, a reasonably steady and controllable situation with a constant temperature, lighting and a quiet office building atmosphere could be guaranteed. Unforeseen events like groups of people in the corridor or loud noises were written down in the test protocol separately for each test session. The walking path of 55 cm width was marked on the floor with masking tape. It was realized as a closed ellipse of 57 m length in total with eight barriers of 24.5 cm length, also marked with masking tape. The participants were instructed not to leave the marked corridor and not to step on the barriers. Every breach of this instruction was also noted in the protocol.

2.1. Feedback Types

The following events during the interaction with the app had feedback messages of various types. The descriptions here are given with the baseline visual feedback, which was complemented by auditory or/and tactile feedback in the other trials:

- **Incoming message:** "A new picture message was received" as a textual pop-up.
- **Button press:** While typing a PIN each number button provided feedback whilst pressed.
- **Correct PIN:** Feedback when PIN was correct.
- **Incorrect PIN:** Feedback when PIN was wrong.
- **Picture download completed:** After 5 seconds download time a notification about the completed download appeared and the image could be displayed.

As the aim of this Interaction Study was to directly compare the different possible feedback modalities and alternative feedback types, a multidimensional within-subjects design was chosen. As a baseline, each of the aforementioned events had visual feedback in each condition. Hence, in the following only the additional modalities are mentioned for the sake of brevity. Auditory, tactile, and audio-tactile feedback was added in the other conditions. For the audio and tactile feedback, two types were tested. Consequently, with audio-tactile feedback, four combinations were generated that are called *multimodal* feedback types in the following.

Table I shows the different combinations of modalities and types, resulting in nine different feedback types. Each participant tested each feedback condition in counterbalanced order. The messages were selected after prior studies [11, 12] with regard to their functional applicability ratings for the events described above and their affective impression.

Table I. Different combinations of modalities and types.

Type Name	Modalities	Type Description
V	visual	baseline condition, texts & pop-up messages
T1	tactile (+ visual)	subtle vibration patterns
T2	tactile (+ visual)	stronger vibration patterns
A1	auditory (+ visual)	subtle earcons
A2	auditory (+ visual)	intrusive auditory icons
M1	auditory, tactile (+ visual)	A1 + T2
M2	auditory, tactile (+ visual)	A1 + T1
M3	auditory, tactile (+ visual)	A2 + T2
M4	auditory, tactile (+ visual)	A2 + T1

2.2. Material and Participants

2.2.1. Questionnaires

In an introductory questionnaire, demographic data such as age, gender and prior experience with smartphones was gathered. To assess the affective impression of feedback in this interaction setup the self-assessment manikin (SAM) [13] in a nine-step paper version was used. It represents the three affective dimensions of valence, arousal, and dominance using pictogram-based scales. Participants were asked to answer this short questionnaire after each trial with one feedback type. Additionally, the participants rated each feedback type in general on a five-step scale varying between *very bad* and *very good*.

2.2.2. Performance Data

Performance data and additional information about the test conditions were assessed in two different ways. The test protocol was written by hand during the test for each trial and consisted of:

- the number of rounds on the course, passes over the side marks of the corridor and steps on barriers
- events of slowing down or standing still
- whether it was calm or noisy
- unforeseen events (other people blocking the way, system crashes, or unanticipated behavior of a participant).

The second means to collect performance data was the automatically written log on the smartphone. It contained timestamps for every system event and user action as well as the expected and the entered PINs. With these log files the following performance parameters were determined for each participant per trial:

- the number of errors and corrections
- the time for each complete trial
- the average time needed to enter a PIN in each trial
- the average time to wait until a new message arrived (varying between one and nine seconds, measured mean: 2.77 seconds (SD = 0.5))

2.2.3. Participants

70 participants took part in the Interaction Study. All of them were students aged between 20 and 40 years (M= 25.2 SD = 3.6), 44 of them were male. Of all participants, 87 % owned a smarthpone (mostly Android devices with N = 36, following iOS with N = 19). Non-smartphone users had at least a conventional mobile phone. The majority of 58.6 % of the participants used a method to secure their phone against malicious use, most of them with a PIN (N = 19), an alphanumerical password (N = 8), a re-drawn pattern (N = 8) or something else (N = 6).

2.3. Procedure

Upon arrival, the participants were instructed and informed about the general goals of the study. Then, they filled in a short introductory questionnaire. Next, the participants were asked to try one example PIN with visual feedback while walking along the path.

Then, the actual experiment started, which consisted of four blocks each offering a different feedback modality. The order of modality blocks was varied between all participants using a Graeco-Latin square. Each block was divided into a different number of trials depending on the according modality (i.e. one with visual feedback, two with audio and tactile feedback each, and four with multimodal feedback), resulting in nine trials in total. In each trial the participants had to enter a sequence of five different randomly generated 6-digit PINs that were displayed on the smartphone above the number pad for two seconds.

The procedure of receiving one picture message started with the simple screen showing the current time and date. After a random time varying between one and nine seconds, the message of a new picture was displayed, followed by the PIN entry. The participants could correct the input of the current PIN, when they were unsure about it. After supplying a complete PIN and pressing enter, there was no possibility to re-enter it. Then, they received a response from the smartphone indicating the correctness of the PIN with the according feedback event. In case of a wrong PIN entry, the same PIN had to be re-entered and was shown again in the upper part of the screen for two seconds. After entering the same PIN wrong five times, the current picture was skipped and the LockScreen was shown again with a new message. In case of the correctly entered PIN, the picture was downloaded. After five seconds download time, the notification about a successful download was given and

Table II. Effect sizes and degrees of freedom for the SAM and general feedback ratings.

Variable	df	F	part. η^2
SAM Valence	5.85	5.46	0.074
SAM Arousal	5.21	3.77	0.052
SAM Dominance	5.47	4.66	0.064
General Impression	4.84	13.42	0.165

the according picture could be displayed. This procedure was done five times in each of the nine trials of the four modality blocks.

After each trial of five PINs with one feedback type, the participants were asked to answer the SAM questionnaire [13] and to rate the feedback type in general. Each session lasted approximately 60 minutes.

3. Results

First, this section describes the results regarding the affective impression and the general feedback ratings. Afterwards, the performance results are presented.

3.1. Affective and General Feedback Impression

A repeated-measures ANOVA with *feedback type* as factor for the three SAM variables and the general feedback rating shows significant differences for all four variables. Table II shows the according univariate test statistics. All differences are significant on a level of $p <= 0.01$. However, the effect sizes are low for the SAM ratings but higher for the general feedback question.

Post-hoc comparisons (Bonferroni-corrected) reveal many significant differences between the nine feedback types. The according comparisons are given in Table III for the general feedback question. It can be seen that all feedback types except A2, M3 and M4 are rated significantly better than unimodal visual feedback. This trend is also depicted in Figure 1, which shows the mean general ratings for each feedback type in comparison.

Figure 1 also shows that the unimodal visual feedback type gets the lowest rating in general. The overall best feedback is given with the tactile modality, independent from actual the feedback type. Also the combination of tactile feedback and earcons (M1 and M2) is rated very good. A trend that was also observed in a prior study [11] is that feedback in terms of auditory icons is rated worse than earcons. Thus, it can be said that unimodal visual feedback is rated worse than most of the other feedback types. Only all the feedback types including auditory icons (A2, M3 and M4) are rated equally low.

Figure 2 shows the mean valence ratings for each feedback type. It can be seen, that the trial with unimodal visual feedback gains the lowest valence rating. The combination of tactile feedback and earcons is

Table III. Mean differences with standard errors for significant different feedback types on the general feedback impression.

Compared Feedback Types	Mean Difference	Standard Error	p
V with T1	- 1.087	0.156	0.000
V with T2	- 1.043	0.183	0.000
V with A1	- 0.710	0.169	0.003
V with M1	- 1.043	0.184	0.000
V with M2	- 1.058	0.165	0.000
T1 with A2	0.826	0.187	0.001
T1 with M3	0.754	0.202	0.014
T1 with M4	0.696	0.184	0.012
T2 with A2	0.783	0.183	0.002
T2 with M3	0.710	0.206	0.034
T2 with M4	0.652	0.196	0.049
A1 with A2	0.449	0.131	0.038
A2 with M1	- 0.783	0.160	0.000
A2 with M2	- 0.797	0.166	0.000
M1 with M3	0.710	0.161	0.001
M1 with M4	0.652	0.150	0.002
M2 with M3	0.725	0.176	0.004
M2 with M4	0.667	0.153	0.002

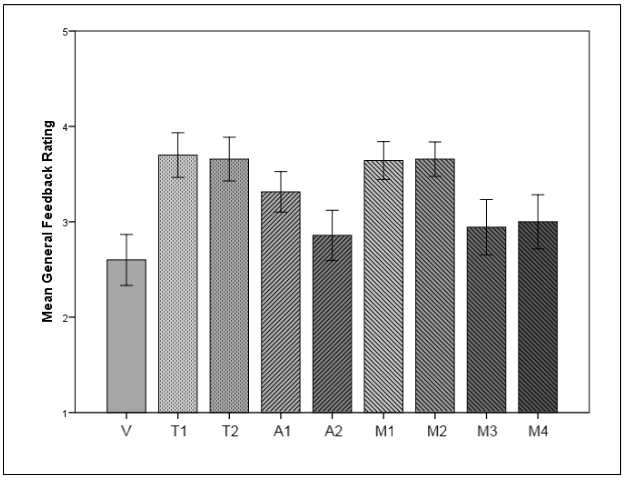

Figure 1. Mean ratings of the general feedback impression for each type.

rated most pleasant, as the feedback types M1 and M2 are rated highest on the valence scale. The combination of tactons and auditory icons is less pleasant. The dominance ratings are highly correlated with the valence ratings ($p <= 0.01$, Pearson's $r = 0.54$). Therefore, similar results are not presented with an extra figure here. Valence and arousal ratings also show a significant correlation in this study but with a very low correlation coefficient ($p <= 0.01$, $r = - 0.14$).

Figure 3 shows that the mean arousal ratings for each feedback type are fairly low for all nine trials. The visual, tactile and the first two multimodal feedback types are rated least arousing. In contrast, the multimodal feedback types with auditory icons (M3 and M4) and the second audio feedback type (A2 with auditory icons) are rated most arousing.

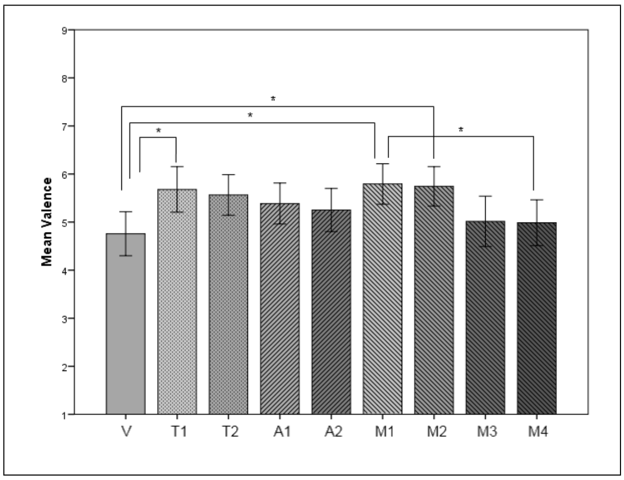

Figure 2. Mean valence ratings for each trial with a different feedback type, lines between bars marked with * show significant differences on a level of p <= 0.05.

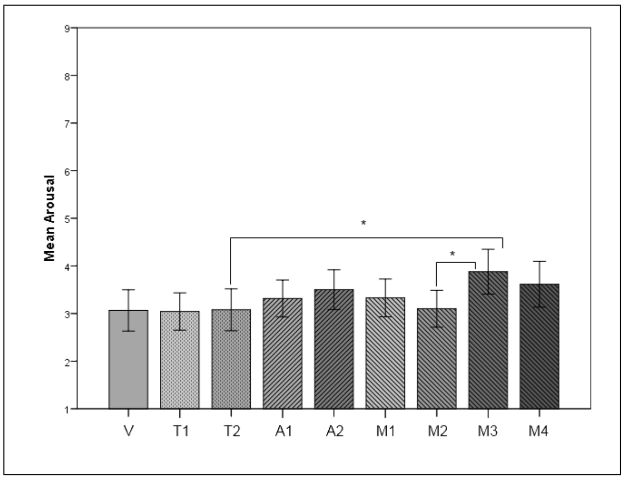

Figure 3. Mean arousal ratings for each trial with a different feedback type, lines between bars marked with * show significant differences on a level of p <= 0.05.

3.2. Course Performance

For the participants, the marked course was easy to follow as is shown by the course performance data. The number of rounds on the course was similar for all trials and all participants with a mean number of rounds varying between 1.99 rounds in the fourth multimodal trial M4 (with SD = 0.47) and 2.12 rounds in the first tactile trial T1 (SD = 0.84). Also the number of passes over the side marks of the corridor as well as the number of steps on the marked barriers were very low for all trials:

- **passes over the side marks:** $SUM_{Min} = 6$, $M_{Min} = 0.09$, SD = 0.28 in trial M4 and $SUM_{Max} = 17$, $M_{Max} = 0.09$, SD = 0.84 in trial T1
- **steps on the barriers:** $SUM_{Min} = 10$, $M_{Min} = 0.14$, SD = 0.43 in T2 and $SUM_{Max} = 24$, $M_{Max} = 0.34$, SD = 0.68 in trial M1

The same holds for stops or events of slowing down, which were mainly caused by other people walking trough the corridor: $SUM_{Min} = 20$, $M_{Min} = 0.39$, SD = 0.84 in trial M4 and $SUM_{Max} = 36$, $M_{Max} = 0.51$, SD = 1.14 in trial M1. Hence, it is not possible to state a relation between feedback type and course performance.

3.3. Errors, Corrections and Durations

In general, the mean number of errors was very low as shown in Table IV. Similarly, participants used the option to correct a PIN only rarely. Nevertheless, the sums of errors and corrections are notable and reveal some variation for the different feedback trials. With visual feedback alone most errors were made. According to a pairwise non-parametrical Wilcoxon test, the differences are significant between trial V and T2 (U = -2.38, p <= 0.05) as well as between V and M2 (U = -2.82, p <= 0.01) based on positive ranks. For the corrections, no significant differences were found, and no evident pattern emerged.

The durations of a complete trial do not differ significantly, shown by a repeated-measures ANOVA with *trial* as main factor (F(8,54) = 1.91, p = 0.07).

On the contrary, for the mean time needed to enter a PIN in each trial, a repeated-measures ANOVA reveals significant differences (F(8,58) = 4.24, p <= 0.01). Post-hoc tests (Bonferroni-corrected) show differences between trial V and M4 ($Mean_{diff} = 0.59$, p <= 0.01), trial M1 and M2 ($Mean_{diff} = 0.39$, p <= 0.05), as well as between trial M1 and M4 ($Mean_{diff} = 0.53$, p <= 0.01). Hence, the longest time needed for the average PIN entry is observed with visual feedback, the shortest time is observed with multimodal feedback including auditory icons and vibrations (M4).

4. Discussion

This study investigated nine different feedback types in a mobile interaction setup. A simple application was used to demonstrate the different feedback types to the users. Following related research, the findings regarding user ratings show that unimodal visual feedback gets the lowest ratings. Contrary, the feedback types including tactile feedback are rated most pleasant and best in general. However, the disadvantages of unimodal visual feedback are not compensated by every combination with another feedback modality. Depending on the modality and the combined type of feedback, some types ware rated similar to the visual condition. Especially, feedback types including auditory icons were rated poor. Thus, it can be said, that unimodal visual feedback is rated worse than most of the other feedback types and the disadvantages of some kinds of auditory feedback seem to be compensated by the addition of tactile feedback. It has to be noted, that the auditory icons used in this interaction study were very salient feedback stimuli, especially

Table IV. Mean, standard deviation and sum of errors and corrections, and mean and standard deviation of trial and PIN entry durations (measured in sec) for all nine feedback types.

	V	T1	T2	A1	A2	M1	M2	M3	M4
Mean Errors	1.24	0.90	0.84	0.90	1.01	1.03	0.72	0.87	0.75
SD Errors	1.72	1.21	1.37	1.25	1.72	1.36	1.22	1.08	1.22
Sum Errors	83	62	58	63	71	72	50	60	52
Mean Corrections	0.75	1.03	0.83	0.84	0.99	0.63	0.81	0.77	0.72
SD Corrections	1.45	2.58	1.19	1.81	1.53	1.11	1.97	1.14	1.01
Sum Corrections	50	71	57	59	69	44	56	53	50
Mean Trial Duration	121.25	122.76	114.41	122.23	123.24	120.13	112.63	122.61	116.59
SD Trial Duration	31.75	47.65	23.37	34.31	30.14	32.88	23.95	35.60	23.51
Mean PIN Duration	6.57	6.28	6.25	6.48	6.36	6.47	6.23	6.22	6.13
SD PIN Duration	1.54	1.27	1.52	1.68	1.60	1.74	1.41	1.49	1.35

compared to the much more subtle earcons. Nevertheless, these were the sounds selected after being rated most suitable for the different events in a previous study [11]. This shows the big difference between the sole and single perception of a stimulus and the effect in an interaction with repeated occurrence.

The performance data measured in this study did not reveal much difference between the different feedback types. Especially for the walking task performance, no conclusion could be drawn to find a relation between performance and feedback type. Because of the easy tasks, during the interaction with the system also a low number of errors and corrections was measured. Nevertheless, the overall sums of errors and corrections are notable and display some variation for the different feedback types. Most errors were made in the trial with visual feedback alone, where participants also needed the most time to enter a single PIN.

To sum up, this study showed once more that adding a modality to unimodal visual feedback leads to better user ratings and to a better user performance. The best rated feedback was observed to be tactile feedback, followed by multimodal feedback containing subtle sounds and vibrations of any type.

5. Conclusion

This paper presented an interaction study with a simple application on a smartphone. The results once more prove that an additional modality complementing unimodal visual feedback can increase user ratings as well as user performance. However, due to the easy tasks only low differences in performance were observed. Regarding user perception there were more distinct differences. Tactile feedback was rated most pleasant by the participants and led to best user performance. Nevertheless, also subtle auditory feedback and the combination of both can be concluded to be likable and helpful. Auditory feedback in terms of auditory icons can not be recommended after this study, as this feedback type was rated worst even in combination with tactile feedback.

References

[1] I. Poupyrev, M. Okabe, and S. Maruyama. Haptic feedback for pen computing: directions and strategies. In Proc. CHI 2004, 1309-1312.

[2] S. Brewster, F. Chohan, and L. Brown. Tactile feedback for mobile interactions. In Proc. CHI 2007, 159-162.

[3] E. Hoggan, S. Brewster, and J. Johnston. Investigating the effectiveness of tactile feedback for mobile touchscreens. In Proc. CHI 2008, 1573-1582.

[4] E. Koskinen, T. Kaaresoja, and P. Laitinen. Feel-good touch: Finding the most pleasant tactile feedback for a mobile touch screen button. In Proc. ICMI 2008, 297-304.

[5] J. L. Burke, M. S. Prewett, A. A. Gray, L. Yang, F. R. B. Stilson, M. D. Coovert, and E. Redden. Comparing the effects of visual-auditory and visual-tactile feedback on user performance: a meta-analysis. In Proc. ICMI 2006, 108-117.

[6] E. Hoggan, A. Crossan, S. Brewster, and T. Kaaresoja. Audio or tactile feedback: which modality when? Proc. CHI 2009, 2253-2256.

[7] E. Hoggan, T. Kaaresoja, P. Laitinen, and S. Brewster. Crossmodal congruence: the look, feel and sound of touchscreen widgets. In Proc. ICMI 2008, 157-164.

[8] M.E. Altinsoy and S. Merchel. Audiotactile feedback design for touch screens. In Haptic and Audio Interaction Design, LNCS **5763**. Springer, 2009, 136-144.

[9] P.J. Lang. Emotion and Motivation: Toward Consensus Definitions and a Common Research Purpose. Emotion Review **2(3)** 2010, 229-233.

[10] N. Ben-Asher, N. Kirschnick, H. Sieger, J. Meyer, A. Ben-Oved, and S. Möller. On the need for different security methods on mobile phones. In Proc. MobileHCI 2011, 465-473.

[11] J. Seebode, R. Schleicher, and S. Möller. Affective Quality of Audio Feedback in Different Contexts. In Proc. MUM 2012.

[12] J. Seebode, R. Schleicher, I. Wechsung, and S. Möller. Investigating the affective impression of tactile feedback on mobile devices. In Proc. BritishHCI 2013.

[13] M.M. Bradley and P. J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. In J. Behav. Ther. and Exp. Psychiar. **25(1)**. 1994, 49 -59.