EFFECT OF AESTHETICS ON AUDIO-ENHANCED GRAPHICAL BUTTONS

Sami Ronkainen, Jonna Häkkilä

Nokia Multimedia

P.O.Box 300, 90401 Oulu, Finland
{sami.ronkainen, jonna.hakkila}@nokia.com

Leena Pasanen

Department of Mathematical Sciences,
University of Oulu.
P.O.Box 3000, 90014 Oulun Yliopisto, Finland
lpasanen@mail.student.oulu.fi

ABSTRACT

In this paper, two different auditory feedback schemes related to graphical buttons are compared to each other and to a visual-only condition. The results show that aesthetically pleasing auditory design is clearly preferred among the users, and can lead to performance benefits over not only a design with no auditory enhancements, but also a design with aesthetically less pleasing auditory enhancements.

1. INTRODUCTION

A grapical push-button is an essential piece of today's graphical user interfaces. Whereas the use of graphical buttons (compared to physical ones) adds flexibility into user interface design and allows space savings, it also challenges the interaction with the device. Because of the lack of haptic feedback, it is beneficial aid the visual channel by offering also alternative feedback. There is a specific problem related to graphical buttons where the button action is only carried out when the pointer is still over the button when it is released. In that situation, especially with touchscreens it may happen that the pointer has in fact slipped off the button when it is lifted, causing the intended input event not to be carried out. This problem - a so-called slip-off error - is emphasized with handheld devices, where the small physical size of the screen often result small size and dense placing of buttons, along with usage situations where full visual attention may not be paid to the interaction. When used on the move, e.g. when walking or in a shaking environment such as a moving car, slip-off errors are likely to occur even more often than in static environments.

Adding auditory feedback provides one way to enhance the usability of graphical input objects[1], [2]. Especially, Brewster *et al* have researched sonically enhanced graphical buttons and compared them in combination with different visual settings [3], [4]. In the studies, the occurrence of a pointer or stylus slipping off a graphical button before releasing the mouse button or lifting the stylus off a touchscreen, is presented utilizing auditory output. The results showed improved efficiency with no indication of increased annoyance in audio-enhanced conditions over the visual ones.

When developing consumer products, the designer has to seriously consider the aesthetics of the design, as it is a major factor of attractivity affecting to the sales of the product. The possibility of annoying the user with sounds was a reason to add the annoyance factor to the Nasa Task Load Index (TLX) questionnaire [5] also in the previous studies. The issue of aesthetics and its effect on the prevalence of auditory displays in general has recently been raised up also by other researchers [6], [7]. Aesthetics plays a big role also in graphical user interfaces. For instance, its effect on the perceived usability of a system has been shown by Kurosu *et al* [8]. In fact, "aesthetic and minimalist design" has been selected as one of the ten usability heuristics by Nielsen [9]. Minimalism can also be an aesthetic value itself, as used e.g. in industrial design, architecture, music and other art forms.

In this paper, we have continued from the framework set by previous research, and using the previous studies as basis, extended the study to compare three feedback conditions in using graphical buttons. Two feedback conditions utilized audio feedback. One of them was designed using a minimalistic approach. The other was similar to the one used in a previous study [4] where pressing a button caused continuously playing auditory feedback. In the third condition, no auditory feedback was used. All feedback conditions also utilized visual feedback, the third condition being one where the visual was the only method of feedback.

The motivation for our research was to verify the amount of annoyance the two different auditory feedback conditions would cause, over a visual-only feedback condition. Furthermore, as the minimalistic design in one of the auditory feedback conditions could have lead into less recognizable sounds, degrading the performance of the users, we wanted to compare the users' performance with both auditory feedback conditions. Finally, we wanted to verify the performance benefits found in the previous studies in the audiovisual feedback conditions over the visual-only one.

2. SOUND DESIGN

Two auditory feedback schemes were created. One was taken almost "as is" from a previous study where its usefulness has been proven [4]. In this condition, continuous sound is played when a cursor is over a button or when the button is pressed. The pitch of the sound changes from C4(261Hz) to C5(523Hz) when the button is pressed. A slip-off error is indicated by that the continuous sound is stopped when the slip-off occurs. A successful selection is indicated by a combination of two short beeps (40ms with 40ms silence in between) of C7(2093Hz). The timbre used in all sounds of the test was the same as in the previous study, i.e. an electronic organ sound was used. In this paper, this sound design type is referred to as the "continuous feedback" type. A demo of the feedback can be found on the WWW page of the Glasgow Multimodal Interaction Group http://www.dcs.gla.ac.uk/~murray/audiowidgets/button1.1.shtml. The on-button sound included in the demo was not used in the original study [4], but we decided to use in our test too. In practice, though, users did not hear the on-button sound much as the test was conducted on a touchscreen where the users usually pressed the button immediately without keeping the stylus hovering over the button.

The second feedback type was designed to have shorter sounds and the design of the individual sounds was attempted to be kept very minimalistic in style. This feedback type will in this paper be called the "minimalistic feedback" type. The auditory feedback paradigm was similar to the "enhanced sound" set in another previous study [3]. In a similar manner as in the study, the sound set consisted of a pen-down sound, a pen-up sound, and a slip-off sound. The pen-up sound was played when a successful selection was made. A slip-off error was indicated with the slip-off sound. The sounds consisted of everyday sounds. The pen-down sound was a short (36ms) sample of a tin can surface being pressed to make a dent. The pen-up sound was an equally short sample of releasing the tin can so the dent straightened out again. The sounds were filtered and otherwise treated to make them even less obtrusive. The slip-off sound was a scratching sound, lasting 110ms.

The sound set, to which the minimalistic set was compared, was selected from the older study [4], consisting of longer sounds than in the more recent study [3]. This selection was made because either of the previous studied showed no increased annoyance factor when adding auditory enhancements in the graphical buttons. In our study, we wanted to verify this by allowing users to compare longer lasting sounds to shorter ones, and to a silent condition.

Auditory feedback was also noted to increase the users' performance in a selection task in the previous studies. The verification of this was also a goal in our study, along with a comparison of the two auditory feedback schemes.

3. USER TESTING

To test the designs, a user test involving twelve participants were conducted. The test included three different auditory feedback conditions: A) the minimalistic design sound set, B) the continuous feedback sound set, and C) no auditory feedback. In all feedback conditions, also the regular Windows visual feedback of button presses was presented along with the (possible) auditory feedback.

The feedback schemes were tested with a graphical touchscreen user interface. The NASA TLX questionnaire was used to measure subjects' opinions of the different feedback types. Finally, subjects filled out a written survey and selected their preferred feedback condition.

3.1. Test equipment

The tests were conducted using a Fujitsu-Siemens Lifebook B2131 touchscreen laptop computer. A piece of software was created to implement the test user interface and to log the variables measuring the performance of the users. Sounds were played through a pair of Sennheiser HD-25 SP closed-back headphones. Tests were conducted in a quiet office room.

3.2. Touchscreen UI test

The test setup was similar to one described in the earlier studies [3], [4]. In the test, the subject typed five-digit number sequences by pressing graphical buttons on touchscreen laptop with a stylus. The user interface used in the test is presented in Figure 1. A randomly generated code was presented in the bottom right corner of the screen. The user had to type the same code using the number buttons. The typed digits appeared to the box in top of the window. After each typed digit the user had to press an OK button, to maximise the user's focus shifts and movement of the hand holding the stylus. Next digit could be entered only after the OK button had been pressed. If a wrong digit was typed, it could be corrected using the Del button. Each of these buttons were treated with the A B or C feedback schemes. There was also a "Next" button the user pressed after completing entering a code. The "Next" button always had only a visual presentation and its usage was not included in the test data.

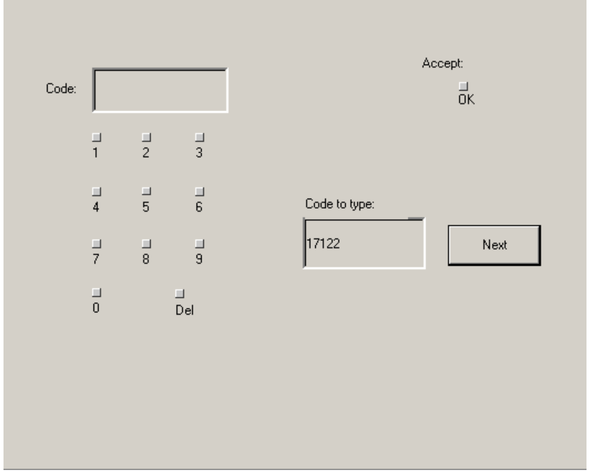
As the aim was to study the role of auditory feedback, the graphical buttons were designed small in size to make the accurate hitting more difficult. The size of the number buttons, the OK button and the Del button was about 2.5mm^2.

The different feedback conditions were fully counterbalanced for presentation order. The users were divided randomly into six groups (with two users in each). The feedback schemes A, B and C were presented in a different order to every group. The order of presentation of the feedback conditions to each group (g1 - g6) of two users is shown in Table 1.

g_1	ABC BCA CAB ACB CBA BAC
g_2	BCA CAB ACB CBA BAC ABC
g_3	CAB ACB CBA BAC ABC BCA
g_4	ACB CBA BAC ABC BCA CAB
g_5	CBA BAC ABC BCA CAB ACB
g_6	BAC ABC BCA CAB ACB CBA

Table 1. *Presentation order of feedback types*

The users entered codes for three minutes at a time with each feedback condition. Because of the counterbalancing of the presentation order, every user was subjected six times to each feedback condition. In other words, every user performed 18 three-minute conditions. The average number of codes entered by each participant was 208.



Code to type: 17122

Figure 1. *Test set-up.*

The variables measured in the test were the number of correctly entered codes (OK) in each three-minute period, the amount of errors (V) related either to slip-off situations or entering wrong digits for some other reason, the number of button presses needed for recovering from an error (VN) and the time needed for the recovery (VT). Furthermore, also the number of times (N) the user missed a button, hitting the background of the window instead, was logged. The variables (with the exception of N) were the same as in the previous studies [3], [4].

3.3. NASA TLX Questionnaire and User Preferences

In addition to the measured data, the subjects' opinions and preferences were asked. NASA TLX-questionnaire was used to chart the perceptions of annoyance, similarly as done in the previous studies [3], [4]. In a similar manner as in previous tests, the "Annoyance" factor was added to the regular TLX questionnaire. In the end of the test, users filled out a written survey, and named their preference of the three feedback conditions.

4. STATISTICAL ANALYSIS

The users' performance in the code entering task was studied utilizing generalized linear mixed models. A generalized linear mixed model is a model, where the effects of some variables explaining the response are assumed to be random. For each set of observed values of the response variables (OK, N, V, VT, VN), a model best fitting the set has been created. The most interesting characteristic defined by the model is the expectation value of the response, which is assumed to describe the real-world expectation value of the underlying phenomenon.

The expectation values of the responses depend on linear predictors, i.e. the terms, through which the explaining factors affect the response through a non-linear link function. In addition, the response can be assumed to have distributions differing from the normal distribution, where these distributions belong to the so-called exponential family of distributions. A more detailed explanation of generalized linear mixed models can be found e.g. in [10].

In here, either a logarithmic link function has been used when fitting the observations to a model of a response, or transformation log(Y) has been done to the observed values of the response before the analysis. Link functions have been used when the responses has been assumed to follow a Poisson distribution, and log(Y) transformations have been used when the responses have been assumed to follow a normal distribution. In either case, because of this, comparing the effects of different auditory feedbacks to the response is done by comparing the ratios of expectation values in each model with the different set of explaining factors (the most interesting one being the auditory feedback type).

The approximate 95% confidence interval for the response Y has been calculated with the equation

$$e^{(\hat{b} \pm t_{0.975}(df) SE(\hat{b}))} , \tag{1}$$

where

$\hat{b}$ is the estimate whose confidence interval is being calculated,

df is the number of degrees of freedom in the system,

$t_{0.975}(df)$ is the 0.975 quartile of t-distribution with *df* degrees of freedom, and

$SE(\hat{b})$ is the standard error of the estimate $\hat{b}$.

When analyzing the number of successfully entered codes (response OK) and number of errors (response V), a generalized linear mixed model was used so that the responses have been assumed to have a Poisson distribution with a logarithmic link function. The assumption of Poisson distribution was made, because the responses were count variables. The effect of test participants individual differences on the values of the responses has been assigned as a random variable. The parameters of the model have been estimated with ML method by using the NLMIXED procedure of the SAS statistics software tool.

Observations from responses N (measuring the number of missed hits) and VT (time to recover from error) were treated with the log(Y) transformation. Even though the response N is a count variable, the logarithm of N has been assumed to have normal distribution. This assumption was made as during the analysis it was found that normal distribution and Poisson distribution fit to the measurement data equally well. With response VT, with the observations having a large number of errors, the values of the responses were better estimable. Thus, the model was weighted with the number of errors during the analysis. The effect of test participants individual differences on the values of the responses has again been assigned as a random variable. The parameters of the model have been estimated with ML method by using the MIXED procedure of the SAS statistics software tool.

The response VN, measuring the mean number of button presses required for recovering from an erroneous press, has been assumed to have a Poisson distribution with a logarithmic link function. The assumption of Poisson distribution was done as the response was a count variable. In the analysis, the model was weighted with the number of errors. With the response VN it was discovered that with this data, the effect of an individual test participant was not statistically significant, and thus the random term could be left out. The parameters of the model were estimated by using the glm procedure of R statistics software tool.

4.1. Period's Effect to the Response

To estimate the effect of the auditory feedback as accurately as possible, the analysis has taken into account the effect of the first, second and third power of the period, as well as if the period's effect to the response was similar with every auditory feedback. The data did not give any proof of having a different period effect for different feedbacks, except of the response VN. Because of this, with the models of other responses than VN, the ratio of expectation values of the responses is constant with all periods - even though the values of the responses differ as the function of the period. For instance, the response OK (number of codes entered in three minutes) depended on the period so that the users' performance got better all the time towards the end of the test. However, this effect was the same on each auditory feedback type.

In the model of the response VN, the effect of the period with the auditory feedback C differs from the periods effect with feedbacks A and B, and thus the ratio between audio feedbacks did not remain constant when the period changes. However, the fact to be found out through the analysis was the ratio of the expectation values of the responses between different feedback types. As can be seen in Table 2, the ratio is below 1 (or, 100, as the table lists the values in percentages) in both when comparing VN response in feedback types A and C, and when doing the same for types B and C. Even though the ratio was dependent on the period, it nevertheless remained below 1 constantly, so in the final comparison the period effect does not appear. The values in Table 2 for the ratios for responses VN have been taken from the middle of the test.

5. TEST RESULTS

5.1. Touchscreen UI test

Because of the logarithmic link function and log(Y) transformations, all the comparisons of different variables must be done by comparing the ratios of the expectation values of the calculated responses.

In the following, all ratios are presented as percentages. In other words, if two different feedback conditions caused the same response in one of the studied variables (OK, N, V, VN, VT), the ratio is 100. The statistically significant findings are those where the approximate 95% confidence interval of the ratio is either above or below 100.

Ratio	A/B	A/C	B/C
OK	110	116	105
(95%CI)	99 - 123	104 - 129	94 - 118
N	112	121	108
(95%CI)	102 – 124	110 – 134	98 – 119
V	88	93	106
(95%CI)	80 – 97	85 – 102	97 – 116
VT	95	72	76
(95%CI)	88- 103	67 - 78	71 - 81
VN	104	68	65
(95%CI)	97 - 112	63 - 72	61 - 69

Table 2. *Averages and approximate 95% confidence intervals of the ratios of expectation values of different response functions*

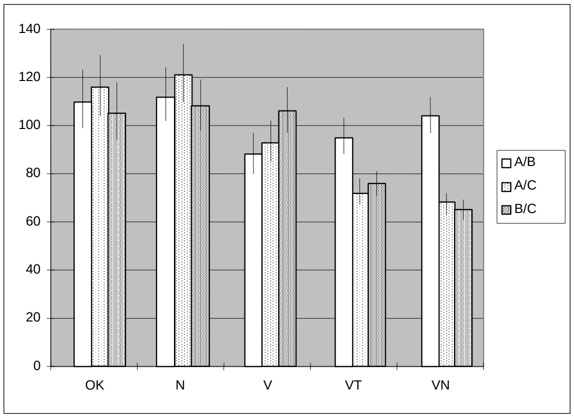


Figure 2. *Averages of the ratios of expectation values of different response functions (95% confidence intervals shown)*

The values for the ratios (along with their approximate 95% confidence intervals) are shown in Table 2. A graphical presentation of the same is shown in Figure 2. Notice that for the variable OK, higher values indicate better performance while for all the other variables, lower values indicate better performance.

The statistically significant results are found for variable OK (the number of correctly entered codes) in ratio A/C, for variable N (number of stylus hits missing the buttons) in ratios A/B and A/C, for variable V (number of errors made) in ratio A/B, for variables VT (time to recover from error) in ratio A/C and B/C and for variable VN (number of button presses to recover from error) in ratios A/C and B/C.

In other words, feedback type A (the minimalistic sound design) allowed users to enter more correct codes (OK) than type C (silent). The difference between type A and B is also very close to statistical significance, as the confidence interval for the A/B ratio is between 99 and 123. Between the B and C types, no such difference is indicated by these results.

As for the number of errors made (V), the A type caused significantly fewer errors to be made than the B type. No significant difference could be found between A and C or between B and C.

The time to recover from errors (VT) was significantly lower in the A feedback type than in C, and also in B type compared to C. No such difference between the A and B types was found.

A similar difference exists when looking at the number of button presses needed for recovering from an error (VN). Again, with the A or B types the number was significantly lower than in C type, and again the results do not indicate a difference between the A and B auditory feedback types.

As for the number of times (N) that the user missed a button, hitting the window background instead, the number was significantly higher in the A feedback type than in B or C. Again, no such difference was found between the B and C types.

5.2. NASA TLX Questionnaire

The average scores for each feedback type for each question in the questionnaire can be seen in Figure 3. Note that in all questions except 'Own performance', higher score means more workload.

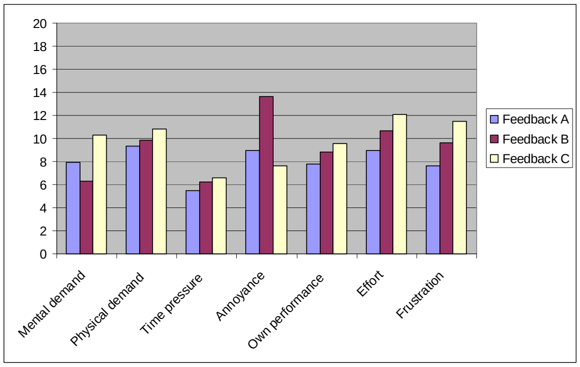


Figure 3. *Average TLX scores for each feedback type*

Wilcoxon's rank sum test was used to test whether the differences between answers to each feedback type were real. Data set A was compared against B and C, and B was compared against C.

The questions where Wilcoxon's rank sum test showed statistical significance ($p<0.05$) were as follows: Mental demand (A vs. C, B vs. C), Physical demand (only A vs C), Annoyance (A vs. B, B vs. C - but not A vs C), Effort (only A vs C) and Frustration (A vs. B, A vs. C - but not B vs C). Student's t-test (two-tailed, paired) was also performed for each seven sets of answers for each feedback type (A vs. B, A vs. C, B vs. C), even though the data in the sets were not tested for normal distribution. The test showed statistical significance ($p<0.05$) for all of the mentioned scores, backing up the indications of the Wilcoxon's rank sum test.

There was one borderline case as well. Wilcoxon's test showed statistical significance for Time pressure (A vs. B, $p=0.0469$) but t-test did not ($p=0.08$). Wilcoxon's test does not make very accurate comparisons when many users give equal ratings, as only those ratings that are not equal, take part in calculation of the test variable. In the Time pressure case, 6 users gave equal ratings to A and B conditions, so it is questionable whether the significance shown by Wilcoxon's test really means that the difference is real.

To summarize, mental demand with both auditory feedbacks (A and B) were rated significantly lower than without auditory feedback. Physical demand in the A auditory feedback condition was rated lower than in the silent condition, but for the B auditory feedback the same result was not found, nor when comparing the auditory feedback cases to each other. Annoyance was rated highest with B type auditory feedback compared to A type or the silent type. Effort was rated lower in feedback A than in C, but not in B vs. C or A vs. B. Frustration was rated lowest with the A type auditory feedback, compared to B type or the silent type.

5.3. User preferences

The user preferences were simply asked from all the participants, asking which feedback type they liked best. The results were almost unanimous: 10 participants preferred the A type, one preferred B and one preferred C. The free-text comments to the A type sounds varied from preferring natural sounds over synthetic ones, to several comments about the sounds being helpful, yet unobtrusive. The one user who preferred the B feedback type commented that audio feedback helped the interaction but the A type sounds were more boring than B type. The user preferring C type liked the fact that without auditory feedback one really had to concentrate on the task, which helped the performance.

6. DISCUSSION

In a similar manner as in previous study [4], the time to recover from errors, and the amount of button presses needed for it, were significantly lower when auditory feedback was utilized (A,B), compared to visual-only feedback (C). So, users were able to notice errors quicker when auditory feedback was utilized. The two auditory feedback schemes did not significantly differ from each other in this respect.

However, the number of correctly entered codes varied between different auditory feedback types. The minimalistic feedback type led into a significantly higher number than in the silent condition, and almost significantly higher number than in the continuous auditory feedback type. The continuous type of auditory feedback did not lead into a higher number of correctly entered codes, compared to the silent type.

The explanation may reside in the number of errors made and time to recover from them. The minimalistic feedback type led to roughly the same number of errors made than the silent condition, but error recovery was quicker. But in the case of the continuous feedback type, the number of errors was greater (albeit not significantly) than in the silent type. Even though the error recovery was quicker, the final amount of correctly entered codes was not significantly higher than in the silent condition.

The continuous feedback type therefore seems to have disturbed the users in their task, compared to the minimalistic type. The same disturbance was not apparent when comparing the minimalistic auditory feedback to the silent condition. It therefore seems that having too much (or perhaps unpleasant, as indicated by the free-text comments) auditory feedback can disturb the performance of the user. However, explaining the (possibly) underlying phenomenon will require further research. In a previous study [6], it was discussed that poor functional rating for a sound may also lead to a poor aesthetic rating. In our study, the performance of the users with the continuous sound feedback style was worse than with the minimalistic style. So the phenomenon can also be that some characteristic in the continuous sound feedback disturbed the users, leading them into regarding also the aesthetic values of the sounds less favourably.

The number of stylus hits on the window background was significantly higher in the A feedback condition than in B or C. The number in the B feedback was also higher than C, but the difference was not statistically significant. However, this may indicate that when working with no auditory feedback, the users were more careful and concentrated more on the task.

The increased concentration would be in accordance with the subjective workload shown in the TLX questionnaire. The score for Mental demand was significantly higher in the silent condition than either of the auditory feedback conditions. Also the score for Effort was significantly higher in the silent condition than in the minimalistic feedback condition. However, the score for Effort in the continuous feedback did not significantly differ from that of either of the other two conditions. Interestingly, also the score for Physical demand was similar to that of Effort - significantly lower in the A condition than C, but not when comparing the B condition to either A or C.

The score for Frustration was clearly lowest in the A feedback condition, compared to B or C. No such difference between B and C could be found.

An interesting finding that is contrary to the previous studies is that related to the score for Annoyance. In the continuous feedback condition, the score was significantly higher than in either of the other two conditions. So, the continuous feedback was clearly more annoying than the silent condition, which is contrary to the previous findings. The difference between the minimalistic and silent conditions was not significant, which is in line with the previous findings. So, our results do not back up the hypothesis that auditory feedback as such would be more annoying than visual-only feedback. Nevertheless, the design of the auditory feedback seems to play an important role.

This was also indicated by the free-text comments and the user preference scores. The minimalistic sound design was quite clearly the most preferred one among the feedback conditions, over both the continuous auditory feedback and silent types.

7. CONCLUSIONS

In this paper, we have presented a comparison of different auditory feedback schemes for audio-enhanced graphical buttons. Error recovery was quicker with either of the auditory feedback types, compared to the silent case, which verified a result presented also in earlier studies. A rather more interesting result was that the minimalistic sound design scheme led to better performance than the less minimalistic one, when comparing the number of errors made in the test. Also the users clearly preferred it over the other two - the silent case included. Earlier studies e.g. by Kurosu *et al* [8] have shown improved perceived usability with aesthetically more pleasing design. In our case, the difference was also objectively measured. Verifying that the result is really an effect of aesthetics and not some other differences in the designs, requires still further research. As discussed in an earlier study [6], the perceived aesthetics can also be affected by the context of the sounds with poorer absolute performance leading to less favourable subjective ratings.

The results show that aesthetically pleasing, short interaction sounds are not more annoying than having no sounds at all, and the users clearly perceived and appreciated the performance benefit the sounds offer. When compared to longer sounds, the shorter ones led to a moderate improvement in the performance of the users, and to a clear preference in their subjective ratings.

ACKNOWLEDGEMENTS

The authors wish to thank prof. Esa Läärä from the Department of Mathematical Sciences in the University of Oulu for his advices on the statistical analysis of the results.

8. REFERENCES

[1] W. W. Gaver, "The SonicFinder: An Interface that Uses Auditory Icons." *Hum.-Comp. Inter.* 4(1) (1989).

[2] M.C. Albers and E. Bergman, "The Audible Web: Auditory Enhancements For Mosaic." *CHI'95 Conference Companion*, ACM Press, Addison Wesley, 1995: pp318-319.

[3] S. A. Brewster, "Overcoming the Lack of Screen Space on Mobile Computers", *Personal and Ubiquitous Computing*, v.6 n.3, p.188-205, May 2002.

[4] S.A. Brewster, P.C. Wright, A.J. Dix and A.D.N. Edwards, "The sonic enhancement of graphical buttons" in *Proceedings of IFIP Interact'95*, Chapman& Hall, 1995, pp. 43-48.

[5] S. Hart and L. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research." In Hancock, P. and Meshkati, N. (eds.) *Human mental workload*, North Holland B.V., Amsterdam, 1988, pp 139-183.

[6] G.Leplâtre and I.McGregor, "How to Tackle Auditory Interface Aesthetics? Discussion and Case Study", *Proc. of 10th Meeting of the International Conference on Auditory Display (ICAD)*, Sydney, Australia, July 6-9 2004.

[7] C. Thornton, A. Kolb, F. Gemperle and T. Hirsch, "Audiocentric Interface Design: A Building Blocks Approach", *Proc. of 2003 International Conference on Auditory Display*, Boston, MA, USA, July 6-9 2003, pp. 180-183.

[8] M. Kurosu and K. Kashimura, "Apparent usability vs. Inherent usability", *CHI'95 Conference Companion*, ACM Press, Addison Wesley, 1995. pp. 292-293.

[9] J. Nielsen, "Heuristic evaluation". In J. Nielsen and R.L. Mack (Eds.), *Usability Inspection Methods*, John Wiley & Sons, New York, NY, 1994.

[10] C. E. McCulloch and S. R. Searle, *Generalized, Linear and Mixed Models*. John Wiley & Sons, inc., 2001.