HCI Methodology For Evaluating Musical Controllers: A Case Study

Chris Kiefer
Department of Informatics
University of Sussex,
Brighton, UK
C.Kiefer@sussex.ac.uk

Nick Collins
Department of Informatics
University of Sussex,
Brighton, UK
N.Collins@sussex.ac.uk

Geraldine Fitzpatrick
Interact Lab
Department of Informatics
University of Sussex,
Brighton, UK
G.A.Fitzpatrick@sussex.ac.uk

ABSTRACT

There is small but useful body of research concerning the evaluation of musical interfaces with HCI techniques. In this paper, we present a case study in implementing these techniques; we describe a usability experiment which evaluated the Nintendo Wiimote as a musical controller, and reflect on the effectiveness of our choice of HCI methodologies in this context. The study offered some valuable results, but our picture of the Wiimote was incomplete as we lacked data concerning the participants' instantaneous musical experience. Recent trends in HCI are leading researchers to tackle this problem of evaluating user experience; we review some of their work and suggest that with some adaptation it could provide useful new tools and methodologies for computer musicians.

Keywords

HCI Methodology, Wiimote, Evaluating Musical Interaction

1. INTRODUCTION

A deep understanding of a musical interface is a desirable thing to have. It can provide feedback which leads to an improved design and therefore a better creative system; it can show whether a design functions as it was designed to, and whether it functions in ways which may have been unexpected. The field of Human Computer Interaction [4] provides tools and methodologies for evaluating computer interfaces, but applying these to the specific area of computer music can be problematic. HCI methodology has evolved around a task-based paradigm and the stimulus-response interaction model of WIMP systems, as opposed to the richer and more complex interactions that occur between musicians and machines. Höök [5], discussing the relationship between HCI and installation art, suggests that art and HCI are not easily combined, and this may also be true in the multi-disciplinary field of computer music.

Wanderley and Orio's article [11] from 2002 built a bridge between HCI usability evaluation methodology and computer music, reviewing current HCI research and suggesting ways in which it could be applied specifically to the evaluation of musical controllers. Since then, research in this area has been relatively sparse and the adoption of these evaluation techniques by the community seems relatively low. A review of the 2007 NIME proceedings, for example, showed that 37% of papers presenting new instruments described some sort of formal usability testing, though often not referenced to the wider HCI literature. One possible reason for the slow uptake of HCI methods is that the practicalities of carrying out a usability study are something of a black box as, understandably, papers tend to focus on results rather than methodology. It is clear that there is a lot more that could be done to draw on HCI; this paper is a modest response to this challenge by explicitly articulating the processes and lessons learnt in applying HCI methodology within the music field.

To date there is only limited HCI literature which focusses specifically on computer music. Höök [5] examines the use of HCI in interactive art, an area which shares common ground with computer music. She describes her methodology for evaluating interaction in an installation, and examines the issue of assessing usability when artists might want to build systems for unique rather than 'normal' users; music shares similar characteristics with art. Poepel [10] presents a method for evaluating instruments through the measurement of musical expressivity. This technique is based on psychology research on cues for musical expression; it evaluates players' estimations of a controller's capability for creating these cues. Wanderley and Orio [11] have conducted the most comprehensive review of HCI usability methodologies which can be applied to the evaluation of musical systems. They discuss the importance of testing within well defined contexts or metaphors, and suggest some that are commonly found in computer music. They propose the use of simplistic musical tasks for evaluation, and highlight features of controllers which are most relevant in usability testing: learnability, explorability, feature controllability and timing controllability. Their research fitted best with our objectives for evaluating the Wiimote, and had the largest influence of the methodology used.

We present a case study on the musical usability of the Nintendo Wiimote. This practical example will help to ground this research and provide a talking point for the employment of HCI evaluation for interactive systems. We will go on to review more recent developments in HCI, from the so called 'third paradigm', and discuss how they might be applied in our field in the future.

2. A CASE STUDY

The semi-ubiquitous Nintendo Wiimote is becoming popular with musicians, as can be seen from the multiplicity of demo videos on YouTube. This motivated us to carry out

Figure 1: The Nintendo Wiimote and the Roland HandSonic

a formal evaluation, asking the broad question *how useful is the wiimote as a musical controller?* Answering this question presents a number of challenges. What should be evaluated to give an overall picture of the device? How can the capabilities of the controller be judged with a minimum of influence on the results from software design and an acknowledgement of the potentially differing musical and gaming skill levels of the participants?

2.1 Experimental Design

The Wiimote is essentially a wireless 3-axis accelerometer with some buttons and an IR camera. Due to dependence on the force of gravity, only rotation around the roll and tilt axes are effective for accurate measurement. Following Wanderley and Orio's guidelines, we decided to test the core musical capabilities of the accelerometer using simplistic musical tasks; an evaluation of the basic functions could be extrapolated from this to help assess the Wiimote's use in more complex musical situations. The core functions that were tested were triggering (with drumming-like motions), continuous control using the roll and tilt axes, and gestural control using shape recognition. Continuous control was divided into two categories; precise and expressive.

Practical constraints had to be considered. The participants in the study were volunteers, so a balance had to be struck between the length of the experiment and the degree to which this might dissuade potential participants. A length of 30 minutes was selected.

Whenever possible, in order to give the participants a baseline for comparison of the Wiimote functions, a controller which represented a typical way of performing the musical tasks was provided. The Roland HPD-15 HandSonic was selected for this purpose, as it has a drum pad for comparison of triggering and knobs for comparison of continuous control tasks. The data from this controller would also provide a basis for statistical comparisons.

The triggering task involved participants drumming a set of simple patterns along with a metronome, to obtain timing data. The precise control task required participants to co-ordinate discrete changes of pitch of a sawtooth wave to the beats of a metronome, and was repeated once for each Wiimote axis as well as for turning a knob on the HandSonic. The expressive control task involved modifying the filter and grain density parameters of a synthesiser patch simultaneously. Finally, the gestural recognition task was to control five tracks of percussion, by muting and un-muting layers through casting shapes with the Wiimote. Before each task the participants were given a period of practice time; after each task they would be interviewed while the experience was still fresh, and asked about their preferred controller. We considered using the 'think aloud' method of gathering data during the tasks, but decided that this would be incompatible with a musical study as it would distract the participants' attention. All data was recorded for later analysis.

A script was written which described the events in the experiment and the wording of the interview questions in order to help the experimenter keep these constant for each participant. Participants were asked up front about factors which might affect the experiment such as musical experience and experience of using the Wiimote. After each task, questions probed their experience in using the controllers. To reduce learning effect, the order of use of the HandSonic and the Wiimote were alternated between participants.

A call for participants was sent out to university mailing lists and local musicians, 21 people volunteering in total. The study commenced as a rolling pilot, with experimental parameters being checked and adjusted until a stable setup had been reached. This was important in particular to adjust the difficulty of the tasks and to assess what could be fitted into the 30 minute runtime. The first four sessions ended up as pilot sessions, so the final results were taken from the remaining 17 participants.

During the study participants were videoed, to observe their gestures while using the Wiimote and also to record the interviews which occurred throughout the experiment. The SuperCollider audio software [9] was used to construct the experiment. This software allowed us to record a log file of the participants' actions which would be analysed later for quantitative results.

2.2 Post-Experiment Analysis

The initial data analysis fell into two main areas, the analysis of the qualitative interview data and of the quantitative log file data. Results of the analysis were stored in a MySQL database to facilitate flexible analysis later.

The analysis of the interview data happened over several stages. We used a process of reduction from the raw videos to a final document containing statements summarising the participants' answers to each question along with any interesting quotes. Key parts of answers from the interviews were transcribed from the video data and stored in the database. These quotes were then coded according to an emerging set of categories and then re-coded until the categorisation was stable. The categorised set of quotes was summarised to produce the final document of results. For the quantitative data, the log files were processed in SuperCollider to extract specific data such as timing information from the triggering task. This data was exported to MATLAB for statistical analysis using ANOVA and other tests.

Because we wish to concentrate on methodology, we only have space to give highlights of the results of this study. In interview, several people commented on the lack of physical feedback in the triggering task, saying that this made it difficult to determine the triggering point. The pitch task revealed some insights into the ergonomics of the device; some participants described how going past certain points of rotation felt unnatural. Some perceived it as less accurate than the HandSonic. Participants commented on the Wiimote's intuitive nature when used for expressive control. They described it as 'embodied', and some felt that it widened the scope of editing possibilities. In general, many participants commented on the fun aspect of using the Wiimote, even when they may have preferred the HandSonic. An overall

criticism was of the device's lack of absolute positioning capability. The statistics revealed little significant difference between the two controllers; participants displayed no overall preference and the timing errors showed no significant variance.

2.3 Reflections

What we've done by presenting this case study is to make explicit the practical issues of conducting a usability experiment. This is often mundane detail that gets omitted from experimental reports, but may be the type of essential detail that will make it easier to others to try out HCI methods.

As such, it is worth noting a few key points about how the study was implemented. Firstly, the importance of a pilot study is easy to under-estimate. The best way to expose flaws in a script is to put it into practice; for valid results the experimental parameters need to stay constant throughout the study, so flaws need to be removed at this early stage. In retrospect, the difficulty of some tasks in the Wiimote study could still have been better optimised at pilot stage to suit the range of participants' skill levels.

Secondly, an issue of particular importance in a musical usability study is allotted practice time. There's a lower limit on the time participants need to spend becoming accustomed to the features of an instrument; getting this amount wrong can result in unrepresentative attempts at a task, concealing the true results. Again, this is something which can be assessed during the pilot study.

Thirdly, the gathering of empirical data presents some challenges. In order for the data to be valid, the participants needed to perform the tasks in the same way, although getting people to perform a precise task can be difficult especially when you have creative people performing a creative task. There needs to be some built in flexibility in the tasks which allows for this.

Finally, the time and effort in transcribing interviews cannot be under estimated. Even supposing voice-recognition software of sufficient accuracy was available to help avoid the hard slog of manual annotation, it might be at the cost of the researcher not engaging so deeply with the data by parsing it themselves. An alternative approach is transcribing just the 'interesting' sections, which can save time, though this selection process entails some subjectivity. Tagging log file data for analysis was also a long process, as the correct data had to be found manually by comparison to the video; this could have been improved if the logging had been automatically synchronised to the video data.

3. DISCUSSION

The previous section discussed the details of applying the specific methodology we used in this study. It is also useful to reflect more generally on the structuring of the case study and the efficacy of the HCI evaluation. Was it useful to carry out the Wiimote usability study with the methods we chose? Where were the gaps in the results and how could the methodology be improved to narrow these gaps?

The most 'interesting' results came from analysis of the interview data. The interviews confirmed some expected results about the controller but more usefully brought up some unexpected issues that some people found with certain tasks, and some surprising suggestions about how the controller could be used. This is the kind of data that shows the benefits of conducting a usability study, the kind of data that is difficult to determine purely by intuition alone and that is best collected from the observations of a larger group of people. From the remaining results, the quantitative results provided objective backup to certain elements of the interview results, some useful data about the functional side of the controller, and insight into global trends of the participants. However, the conclusions reached from these results alone seemed to be a limited measure of the device compared to the subtlety of the participants' observations.

Did the study result in a complete answer in relation to the research question, *how useful is the Wiimote as a musical controller?* It's difficult to answer this objectively, but it can be observed that the results showed a detailed and intimate understanding of the controller in a musical context. One important thing the results do lack is any measure of the participants' experience while using the controller. The more interesting results came from post-task interviews, but there is no data about their experience in the moment while they were using the device, something that would seem important for a musical evaluation. This gap in the results is partly due to lack of technology and partly due to a lack of methodology. How can musicians self-report their experience while they are using a musical controller without disrupting the experience itself? Are there post-task evaluation techniques that can give a more accurate and objective analysis of a musical experience than an interview? More recent research in HCI is starting to address similar issues and can point to possibilities.

3.1 The 'Third Paradigm'

Kaye et. al. [7], in 2007, described a growing trend in HCI research towards *experience focused* rather than *task focused* HCI. With this trend comes the requirement for new evaluation techniques to respond to the new kinds of data being gathered. This trend is a response to the evolving ways in which technology is utilised as computing becomes increasingly embedded in daily life, a shift in focus away from productivity environments [8], and from evaluation of efficiency to evaluation of affective qualities [3]. As HCI is increasingly involved in other 'highly interactive' fields of computing such as gaming and virtual reality, the requirement for evaluating user experience becomes stronger. This new trend is known as the 'third paradigm', and researchers have started to tackle some of the challenges presented by this approach.

The Sensual Evaluation Instrument (SEI), designed by Isbister et.al. [6], is a means of self-reporting affect while interacting with a computer system. Users utilise a set of biomorphic sculptured shapes to provide feedback in real-time. Intuitive and emotional interaction occur cognitively on a sub-symbolic level so the system uses non-verbal communication in order to more directly represent this. With its sub-verbal reporting method, the SEI is a step in the right direction for evaluation of musical interfaces; however, as the reporting technique already involves some interaction itself, it could only be used effectively in less interactive contexts such as evaluating some desktop software. The most dynamic example of its use is from the designers' tests with a computer game, and they acknowledge in their results that it's not ideal for time-critical interfaces or tasks that require fine-grained data.

For more interactive tasks such as playing a musical controller, a non-interactive data gathering mechanism is essential, so the measurement of physiological data may yield realtime readings without interrupting the users' attention. Some studies concentrate on this area of evaluation. Chateau and Mersiols' AMUSE system [1] is designed to collect and synchronise multiple sources of physiological data to measure a user's instantaneous reaction while they interact with a computer system. This data might include eye gaze, speech, gestures and physiological readings such as EMG, ECG, EEG, skin conductance and pulse. Mandryk [8] examines the issues associated with the evaluation of affect

using these physiological measures; how to calibrate the sensor readings and how to correlate multi-point sensor data streams with single point subjective data. Both studies acknowledge that physiological readings are more valuable when combined with qualitative data. The challenge here is to interpret the data effectively and research needs to be done into how to calibrate this data for musical experiments.

Fallman and Waterworth [3] describe how the Repertory Grid Technique (RGT) can be used for affective evaluation of user experience. RGT is a post-task evaluation technique based on Kelly's Personal Construct Theory, and it involves eliciting qualitative constructs from a user which are then rated quantitatively. It sits on the border between qualitative and quantitative methods, allowing empirical analysis of qualitative data. RGT isn't ideal in a musical context as the data isn't collected in the moment of the experience it evaluates; however, it could be an improvement on interviews, and has the the practical advantage that the data analysis is less time-consuming.

A number of experience evaluation techniques attempt to gather data from multiple data sources in order to attempt to triangulate an overall result. This way of working brings the challenge of synchronising and re-integrating the data sources, and some researchers are creating tools to deal with this [2]. These kind of tools would have been of great value to the data analysis in the Wiimote study, especially because of the need for log file to video synchronisation.

Developments in *new HCI* research are encouraging, but how useful are they in a computer music context? All these methodologies need to be assessed specifically in terms of evaluation of musical experience as well as user experience.

4. CONCLUSION

We have examined current intersections between HCI evaluation methodology and computer music, presented a case study of an evaluation based on this methodology, and looked at some of the new research in HCI which is relevant to our field. The evaluation of the Wiimote produced some valuable insights into its use as a musical controller, but it lacked real-time data concerning the participants' experience of using the device. The *third wave* of HCI holds promising potential for computer music; the two fields share the common goal of evaluating experience and affect between technology and its users. The analysis of musical interfaces can be considered as a very specialised area of experience evaluation, though techniques for new HCI research are not necessarily immediately applicable to music technology. New research is needed to adapt and test these methodologies in musical contexts, and perhaps these techniques might inspire new research which is directly useful to musicians.

5. REFERENCES

[1] Noel Chateau and Marc Mersiol. Amuse: A tool for evaluating affective interfaces. In *CHI'05 Workshop on Evaluating Affective Interfaces - Innovative Approaces*, 2005.

[2] Andy Crabtree, Steve Benford, Chris Greenhalgh, Paul Tennent, Matthew Chalmers, and Barry Brown. Supporting ethnographic studies of ubiquitous computing in the wild. In *DIS '06: Proceedings of the 6th conference on Designing Interactive systems*, pages 60–69, New York, NY, USA, 2006. ACM.

[3] John Waterworth Daniel Fallman. Dealing with user experience and affective evaluation in hci design: A repertory grid approach. In *CHI'05 Workshop on Evaluating Affective Interfaces - Innovative Approaces*, 2005.

[4] Alan Dix, Janet Finlay, Gregory D. Abowd, and Russel Beale. *Human-Computer Interaction*. Prentice Hall, 3rd edition, 2004.

[5] Kristina Höök, Phoebe Sengers, and Gerd Andersson. Sense and sensibility: evaluation and interactive art. In *CHI '03*, pages 241–248, New York, NY, USA, 2003. ACM.

[6] Katherine Isbister, Kia Hook, Jarmo Laaksolahti, and Michael Sharp. The sensual evaluation instrument: Developing a trans-cultural self-report measure of affect. *International Journal of Human-Computer Studies*, 65:315–328, April 2007.

[7] Joseph 'Jofish' Kaye, Kirsten Boehner, Jarmo Laaksolahti, and Anna Staahl. Evaluating experience-focused hci. In *CHI '07: CHI '07 extended abstracts on Human factors in computing systems*, pages 2117–2120, New York, NY, USA, 2007. ACM.

[8] Regan Lee Mandryk. Evaluating affective computing environments using physiological measures. In *CHI'05 Workshop on Evaluating Affective Interfaces - Innovative Approaces*, 2005.

[9] James McCartney. Rethinking the computer music language: SuperCollider. *Computer Music Journal*, 26(4):61–8, 2002.

[10] Cornelius Poepel. On interface expressivity: a player-based study. In *NIME '05: Proceedings of the 2005 conference on New interfaces for musical expression*, pages 228–231, Singapore, Singapore, 2004. National University of Singapore.

[11] Marcelo Mortensen Wanderley and Nicola Orio. Evaluation of input devices for musical expression: Borrowing tools from hci. *Comput. Music J.*, 26(3):62–76, 2002.