Available online at www.sciencedirect.com
ScienceDirect
Int. J. Human-Computer Studies 67 (2009) 960–975
International Journal of Human-Computer Studies
www.elsevier.com/locate/ijhcs

Evaluation of live human–computer music-making: Quantitative and qualitative approaches

D. Stowell*, A. Robertson, N. Bryan-Kinns, M.D. Plumbley

Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK

Received 21 August 2008; received in revised form 16 March 2009; accepted 19 May 2009
Available online 23 June 2009

Abstract

Live music-making using interactive systems is not completely amenable to traditional HCI evaluation metrics such as task-completion rates. In this paper we discuss quantitative and qualitative approaches which provide opportunities to evaluate the music-making interaction, accounting for aspects which cannot be directly measured or expressed numerically, yet which may be important for participants. We present case studies in the application of a qualitative method based on Discourse Analysis, and a quantitative method based on the Turing Test. We compare and contrast these methods with each other, and with other evaluation approaches used in the literature, and discuss factors affecting which evaluation methods are appropriate in a given context.

Keywords: Music; Qualitative; Quantitative

1. Introduction

Live human–computer music-making, with reactive or interactive systems, is a topic of recent artistic and engineering research (Collins and d'Escrivan, 2007, esp. Chapters 3, 5, 8, 10). However, the formal evaluation of such systems is relatively little-studied (Fels, 2004). As one indicator, we carried out a survey of recent research papers presented at the conference on New Interfaces for Musical Expression (NIME—a conference about user interfaces for music-making). It shows a consistently low proportion of papers containing formal evaluations (Table 1).

A formal evaluation is one presented in rigourous fashion, which presents a structured route from data collection to results (e.g. by specifying analysis techniques). It therefore establishes the degree of generality and repeatability of its results. Formal evaluations, whether quantitative or qualitative, are important because they provide a basis for generalising the outcomes of user tests, and therefore allow researchers to build on one another's work.

Live human–computer music-making poses challenges for many common HCI evaluation techniques. Musical interactions have creative and affective aspects, which means they cannot be described as tasks for which e.g. completion rates can reliably be measured. They also have dependencies on timing (rhythm, tempo, etc.), and feedback interactions (e.g. between performers, between performer and audience), which further problematise the issue of developing valid and reliable experimental procedures.

Evaluation could be centred on a user (performer) perspective, or alternatively could be composer-centred or audience-centred (e.g. using expert judges). In live musical interaction the performer has privileged access to both the intention and the act, and their experience of the interaction is a key part of what determines its expressivity. Hence in the following we focus primarily on performer-centred evaluation, as have others (e.g. Wanderley and Orio, 2002).

"Talk-aloud" protocols (Ericsson and Simon, 1996, section 2.3) are used in many HCI evaluations. However, in some musical performances (such as singing or playing a wind instrument) the use of the speech apparatus for music-making precludes concurrent talking. More

*Corresponding author.
E-mail address: dan.stowell@elec.qmul.ac.uk (D. Stowell).

1071-5819/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijhcs.2009.05.007

Table 1
Survey of oral papers presented at the conference on New Interfaces for Musical Expression (NIME), indicating the type of evaluation described.

Evaluation type	NIME conference year		
	2006	2007	2008
Not applicable	*8*	*9*	*7*
None	18	14	15
Informal	12	8	6
Formal qualit.	1	2	3
Formal quant.	2	3	3
Total formal	3 *(9%)*	5 *(19%)*	6 *(22%)*

The last line indicates the total number of formal evaluations presented, also given as a percentage of the papers (excluding those for which evaluation was not applicable).

generally, speaking may interfere with the process of rhythmic/melodic performance: speech and music cognition can demonstrably interfere with each other (Salamé and Baddeley, 1989), and the brain resources used in speech and music processing partially overlap (Peretz and Zatorre, 2005), suggesting issues of cognitive "competition" if subjects are asked to produce music and speech simultaneously.

Other observational approaches may be applicable, although in many cases observing a participant's reactions may be difficult: because of the lack of objectively observable indications of "success" in musical expression, but also because of the participant's physical involvement in the music-making process (e.g. the whole-body interaction of a drummer with a drum-kit).

Some HCI evaluation methods use models of human cognition rather than actual users in tests—e.g. GOMS (Card et al., 1983)—while others such as cognitive walkthrough (Wharton et al., 1994) use structured evaluation techniques and guidelines. These are good for task-based situations, where cognitive processes are relatively well-characterised. However, we do not have adequate models of the cognition involved in live music-making in order to apply such methods. Further, such methods commonly segment the interaction into discrete ordered steps, a process which cannot easily be carried out on the musical interactive experience.

Another challenging aspect of musical interface evaluation is that the participant populations are often small (Wanderley and Orio, 2002). For example, it may be difficult to recruit many virtuoso violinists, human beatboxers, or jazz trumpeters, for a given experiment. Therefore evaluation methods should be applicable to relatively small study sizes.

In this paper we discuss current methods and present two methods developed specifically for evaluation of live musical systems, and which accommodate the issues described above.

1.1. Outline of paper

In Section 2 we first discuss existing methods in the literature, before presenting two particular methods for evaluation of live musical systems:

(1) A qualitative method using Discourse Analysis (DA) (Section 2.2), to evaluate a system by illuminating how users conceptually integrate the system into the context of use.
(2) A Turing-Test method, designed for the case when the system is intended to respond in a human-like manner (Section 2.3).

Sections 3 and 4 present case studies of these methods in action. Then in Section 5 we compare and contrast the methods with each other, and with other evaluation approaches described in the literature, and discuss factors affecting which approaches are appropriate in a given context. Section 6 aims to distil the discussion down to recommendations which may be used by a researcher wishing to evaluate an interactive musical system.

2. Approaches to evaluation

2.1. Previous work

There is a relative paucity of literature in evaluating live sonic interactions, perhaps in part due to the difficulties mentioned in Section 1. Some prior work has looked at HCI issues in "offline" musical systems, i.e. tools for composers (e.g. Buxton and Sniderman, 1980; Polfreman, 2001). Borchers (2001) applies a pattern-language approach to the design of interactive musical exhibits. Others have used theoretical considerations to produce recommendations and heuristics for designing musical performance interfaces (Hunt and Wanderley, 2002; Levitin et al., 2003; Fels, 2004; de Poli, 2004), although without explicit empirical validation. Note that in some such considerations, a "Composer → Performer → Audience" model is adopted, in which musical expression is defined to consist of timing and other variations applied to the composed musical score (Goebl, 2004; de Poli, 2004). In this work we wish to consider musical interaction more generally, encompassing improvised and interactive performance situations.

Wanderley and Orio (2002) provide a particularly useful contribution to our topic. They discuss pertinent HCI methods, before proposing a task-based approach to musical interface evaluation using "maximally simple" musical tasks such as the production of glissandi or triggered sequences. The authors propose a user-focused evaluation, using Likert-scale feedback (Grant et al., 1999) as opposed to an objective measure of gesture accuracy, since such objective measures may not be a good representation of the musical qualities of the gestures produced. The authors suggest by analogy with Fitts' law