Evaluating Interactive Music Systems: An HCI Approach

William Hsu
San Francisco State University
Department of Computer Science
San Francisco, CA USA
whsu@sfsu.edu

Marc Sosnick
San Francisco State University
Department of Computer Science
San Francisco, CA USA
msosnick@sfsu.edu

Abstract

In this paper, we discuss a number of issues related to the design of evaluation tests for comparing interactive music systems for improvisation. Our testing procedure covers rehearsal and performance environments, and captures the experiences of a musician/participant as well as an audience member/observer. We attempt to isolate salient components of system behavior, and test whether the musician or audience are able to discern between systems with significantly different behavioral components. We report on our experiences with our testing methodology, in comparative studies of our London and ARHS improvisation systems [1].

Keywords: Interactive music systems, human computer interaction, evaluation tests.

1. Introduction

We have been building interactive music systems that improvise with a saxophonist or other human instrumentalist ([1], [2]). From the human's real-time performance audio stream, our systems extract timbral and gestural features that are perceptually significant; this information is used to coordinate the performance of an ensemble of virtual improvising agents. In [1], we focused on our two latest systems, the *London* system, and the *Adaptive Real-time Hierarchical Self-monitoring (ARHS)* system. Based on observations of the systems in performances at Live Algorithms for Music 2006, NIME 2007, and at CNMAT in 2008, we tried to identify and address shortcomings by redesigning and fine-tuning system components.

When a system was relatively simple, enhancing its functionality usually led to more musical results. However, as the number of system components increased and their interactions increased in complexity, it became difficult to correlate design decisions to improvements in musicality; we felt the need for a more rigorous approach for comparative evaluation of design choices.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.
NIME09, June 3-6, 2009, Pittsburgh, PA
Copyright remains with the author(s).

It is relatively easy to verify the correct operation of a system component, or identify whether its effect is discernable by a human improviser or a listener. However, our goals are to achieve musically satisfying experiences for both the human improviser and the audience; a functioning component with observable effects may well be considered undesirable by a musician or listener.

HCI testing methodologies are rarely applied to the dynamic user/audience environment of an automatic improvisation system. An evaluation framework for such systems must address both aspects of our high-level goals: 1) the system should constitute a *usable* environment for an experienced human improviser to perform within, preferably for an extended period of time, and 2) the results of the performance should be *musically interesting* for an audience that is sympathetic to free improvisation.

In this paper, we attempt to identify some of the major issues associated with evaluation methodology of interactive music systems, propose a framework for comparative evaluations, and report preliminary results. In Section 2, we survey related work, both from the HCI area and the interactive music community. In Section 3, we describe our approach for developing testing methodologies for evaluating improvisation systems, and apply it to our London and ARHS systems. The design of our evaluative questionnaires is covered in Section 4. We report on our recent experiences with our evaluation methodology in Section 5, and discuss future work.

2. Related Work

Chapter 10 of [4] contains a survey of recent interactive music systems (IMSs). Such systems work mostly with pitch, with timbre playing a minor role; see for example Voyager [5]. In our London and ARHS systems, timbre is an integral and dynamic factor in sensing, analysis, and interaction management; see [1], [2] for details.

We are currently interested in evaluation frameworks for comparing IMSs, using approaches and techniques from Human Computer Interaction (HCI). IMSs can be thought of broadly as human-computer interfaces, with the musician providing input through a microphone to the system, and the musician and audience reacting to the audio output produced by the computer. HCI techniques lend themselves well to the development of such systems.

Collins [4] observed that the evaluation of IMSs "has often been inadequately covered in existing reports…" He

proposed three suggestions for evaluating IMSs: 1) Technical criteria related to tracking success or cognitive modeling; 2) The reaction of an audience; 3) The sense of interaction for the musicians who participate. We will address the latter two in our framework and procedures.

In [3], Ariza describes a variety of listening tests for evaluating generative music systems. Most of these tests ask of listeners a high-level question, such as whether a specific piece of music was composed by a human, or by a generative system. Ariza observed: "The lack of systematic evaluation of aesthetic artifacts in general is traditionally accepted: evaluation is more commonly found as aesthetic criticism, not experimental methodology."

Our focus is on testing procedures to distinguish between the musical behavior of two systems. We would capture data from the point of view of both a musician performing with the system and a listener observing the performance, and then evaluate the "aesthetic artifacts".

Ariza proposed that many listening tests provide "no more than a listener survey." While we agree with his findings, we believe that, despite the musical biases of users and listeners, there is value in this data. The important consideration then is whether we can produce more than "musical judgements" through a testing methodology that examines these very judgements. We will address this in our questionnaire design (Section 4).

Freeman [6] has used short surveys to collect feedback from audiences for his interactive pieces with audience participation, such as Flock. Survey questions tend to be high level, such as whether a respondent "had fun" or "enjoyed participating." "Test runs" were performed before the performance, but details on organization and data collection were not clearly documented.

In [7], listeners were asked to subjectively evaluate bars of existing works; genetic algorithms were then used to generate compositions based on listener preferences. The final analysis of the success of the resulting compositions comes down to a subjective "satisfaction" level. HCI testing has also been applied to musical input devices; see for example [8].

3. Evaluation Framework

We approached our design using Sharp, et al's DECIDE framework [9]. An overview of this approach follows. We will expand on some of these issues in the next few sections.

D: *Determine the Goals*. Our goals are to develop testing methodologies for evaluating interactive music systems for improvisation. The tests will capture experiences of musicians improvising with the IMSs, and audiences observing performances with the IMSs. The results will provide both qualitative and quantitative data for evaluating different IMSs, and guide us in the design of future systems.

E: *Explore the Questions*. What are the common environments in which a musician or listener might experience the IMSs? What are their important behavioral components? Are the differences discernable by the musician or audience? Do these differences result in more or less musically useful results?

C: *Choose the Approach and Methods*. For their simplicity, we decided on simple paper or equivalent electronic questionnaires; we will discuss their design in Section 4. Audio recordings of conversations with participating musicians were made after test sessions.

I: *Identify the practical Issues*. The differing musical experience of audience members and musicians, as well as the variability of venues and performances, need to be accounted for. Collection of data in vivo is another difficult challenge.

D: *Deal with ethical issues*. Privacy of participants is this project's major ethical issue.

E: *Evaluate, interpret, and present data*. This is the ongoing part of our research.

3.1 Working with the musician in rehearsal

We initially focused on capturing the musician's experience with our London and ARHS systems. We decided to work with experienced free improvisers with rich timbral and gestural palettes.

We postulated that a musician would probably need one or more *rehearsals* with each IMS. One important consideration is the amount of information about each IMS that should be made available to the musician before the rehearsal. In classical HCI-based comparative studies of two software applications or variants, detailed information about each variant is usually *not* given to the users beforehand. The concern is such information will bias a user toward one or the other variant. Hence, we felt that a *naïve* rehearsal—where the musician had no information about system behavior, material choice etc.—might capture interesting information about the ease of use of a system. In subsequent rehearsals or performance, more information about a system would be provided to the musician; it might also be interesting to compare the musician's experiences before and after receiving system specifications.

We also felt that the duration of each rehearsal should be chosen carefully. A rehearsal should be long enough for the musician to discover and exercise interesting modes of interaction, but not so long as to be exhausting.

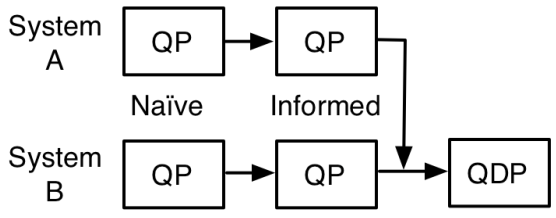

Figure 1. Questionnaire administration for rehearsal. (QP: non-differential questions about each system, QDP: differential questions encompassing both systems)

Our testing procedure attempts to be as unobtrusive as possible, working itself into the natural flow of rehearsal and performance being observed. During a rehearsal, the musician works with the researcher to discover the dynamics of the IMS. We break the rehearsal itself into two sections: 1) a short naïve introductory section (the musician receives no briefing of the internal details of the IMS being tested); 2) the musician is briefed on relevant details of the IMS, after which we have a longer informed rehearsal section. Both sections are recorded. After each section, the musician fills out a questionnaire. See Figure 1 for questionnaire administration workflow.

This rehearsal setup is repeated for each IMS being tested. Hence, for our project comparing the London and ARHS systems, there are four rehearsal sections (two per system), with four questionnaires being administered. At the end of the four rehearsal sections, a fifth *differential questionnaire* comparing the two systems is administered.

3.2 Working with the musician and audience in performance

In a performance setting, the musician will perform two preferably sequential sets, one with each IMS being tested. As discussed in Section 3.1, before each set, the audience will not be told which IMS is involved, to avoid possible bias. Systems are only identified by their order in the set. Following the performance, the musician and audience members who wish to participate fill in questionnaires.

As with rehearsals, we gave some thought to the duration of a performance. An IMS may demonstrate interesting behavior in a relatively short time window, but for various reasons fail to sustain interest in a long performance. This might be an important consideration when comparing two IMSs.

To capture qualitative data, we plan to collect feedback from audience members at a performance; recordings of the performance will also be made available after performances, and interested listeners will be encouraged to provide feedback. The issue of listener preferences for different musical genres is one we would like to avoid for now. We currently focus on audience members who are already experienced listeners of free improvisation or abstract electroacoustic music. In the listener's questionnaire, we ask audience members to rate their previous listening experience.

During the entire test, we document overall testing parameters such as the testing environment, the duration of each section, etc. We also make an audio recording for future reference and for further listening tests.

4. Questionnaire Design

In developing the questionnaires, we needed to ensure that the results were more than what Ariza calls "musical judgments" [3]. By making the tests differential—comparing two different systems' performance—we hoped to narrow the subjective domain.

Though a differential comparison is similar to a musical Turing test [3], we desired richer data than the binary answer that such a test provided. For example, qualitative information could help the developer understand how the modules developed are being perceived by the musician and audience. To move beyond "musical judgments", we need to identify and isolate relatively concrete *behavioral components* for each system. These components will of course vary from system to system. We designed questions that might help identify potential enhancements and be used to distinguish between the two systems. Musicians were asked to rate statements such as "the system was responsive to short-term changes in performance" and "the system facilitated discovery of new musical combinations". Other statements addressed more general, high-level impressions such as "I would [perform | attend a performance with] this system again."

For rating each statement, we used a modified Likert scale [10], with 1 being *strongly agree*, and 5 being *strongly disagree*; one can also respond "N/A" or no answer. Space for comments is provided.

The musician fills out questionnaires at various points in rehearsal and performance, as described in Section 3. After a performance, audience members are encouraged to fill out questionnaires, made available to them at the performance venue. To avoid ethical issues such as privacy or coercion, we will emphasize that response to a questionnaire is entirely voluntary, and that each response is anonymous and may be used for purposes of research.

The methodology and questionnaires have undergone many revisions. The current version of the questionnaire may be found at http://userwww.sfsu.edu/~whsu/IMSHCI.

5. Recent Experiences and Future Work

So far we have focused primarily on evaluation tests and questionnaires from the musician's point of view. We have worked closely with saxophonists John Butcher and James Fei in the development and testing of the London and ARHS systems. Rehearsals with Butcher took place in June 2008 at Guerilla Recording Studio (Oakland CA), followed by a performance at CNMAT (Berkeley CA). Rehearsals with Fei took place in December 2008 at Harvestworks (New York).

We initially expected the feedback from Butcher and Fei to be fairly clearcut, i.e., clearly preferring the more developed ARHS system. After all, the ARHS system is functionally more complex than the London system, and has enhancements specifically targeting the London system's shortcomings. At ICMC 2008, during the presentation of [1], we had played short audio clips (about 2 minutes each) of Butcher working with each system; informal audience feedback afterwards indicated that the ARHS system was preferred. (However, we clearly identified which system was involved in each clip; also, each clip was chosen to highlight the capabilities of each

system.) Hence, we were very surprised with the feedback from Butcher and Fei, after our more formalized tests.

As discussed in Section 3.1, with each IMS, we had Butcher and Fei start with a short *naïve rehearsal*, in which they were given almost no information about the system being tested. We felt that this captured a common situation in free improvisation, where improvisers would meet and perform for the first time, without prior discussions of the performance. We had hoped that after levels were set, the musician would simply start improvising with the IMS. Through performance, s/he would discover how each system worked, and possibly identify the differences between systems. The musician would fill out a questionnaire after the naïve rehearsal, to document her/his experience in the discovery process.

Butcher and Fei both found it difficult to identify differences between the London and ARHS systems in the naïve rehearsal. In fact, both felt that, in the short initial rehearsal, it was easier in some respects to work with the simpler London system, with its phrase-oriented playing. The more complex ARHS system is sensitive to short-term performance changes; it seems to encourage both musicians to play with rapid transitions and "choppy" material. This change in the musicians' performance in turn causes the ARHS system to make frequent adjustments, resulting in a dynamic feedback loop. It is not clear why the slowly developing playing of the London system was preferred in the short naïve rehearsal. Butcher did agree that the simpler London system felt predictable in an extended session, which was not surprising.

In interviews following the rehearsals, both musicians indicated that direction from the programmer would be useful in setting a context for performance. Fei pointed out that a performer would have at least a vague idea of the musical context of an improvisation; for example, a saxophone player would work differently with a loud free jazz rhythm section, versus with quieter acoustic instruments. Butcher also suggested that the musician be asked to play with each system with several different approaches, for example as a soloist, then as a duo partner, etc. In this light, the information obtained from "naïve discovery" seems of limited value.

We plan to drop the initial naïve rehearsals in the future. Instead, the researcher will start by giving the musician an overview of the system being tested. Then the researcher suggests a musical context or progression of gestural and material choices, such as "play long tones for about a minute, followed by short gestures with rapid timbral variations", to elicit specific behavioral responses from the IMS. The musician starts the initial rehearsal section according to the suggestions. A second free rehearsal section, with no restrictions or pre-arranged material choices, will follow.

The development of this evaluation methodology is an ongoing process. We look forward to future testing with larger audiences and a wider variety of musicians. As mentioned, the most recent version of the questionnaire is available online. We are also working on making recordings available online, and implementing an automated system for collecting feedback from listeners. We look forward to and encourage input from the community.

6. Acknowledgements

We especially wish to thank John Butcher and James Fei for their ongoing patience and help in developing this methodology, and CNMAT and Harvestworks for their support of our work.

References

[1] W. Hsu, "Two Approaches for Interaction Management in Timbre-aware Improvisation Systems" in *Proceedings of the ICMC,* Belfast, UK, 2008.

[2] W. Hsu, "Managing Gesture and Timbre for Analysis and Instrument Control in an Interactive Environment" *Proceedings of the Int. Conf. on NIME,* Paris, France, 2006.

[3] C. Ariza, "The interrogator as critic: The questionable relevance of Turing tests and aesthetic tests in the evaluation of generative music systems." in *Computer Music Journal,* 33(1), 2009, pp. 1-23.

[4] *The Cambridge Companion to Electronic Music,* N. Collins and J. d'Escrivan, Eds. Cambridge, UK: Cambridge University Press, 2007, pp. 171-184.

[5] G. Lewis, "Too Many Notes: Computers, Complexity and Culture in *Voyager*" in *Leonardo Music Journal,* vol. 10, 2000.

[6] J. Freeman and M. Godfrey, "Technology, Real-time Notation, and Audience Participation in Flock," in *Proceedings of the ICMC*, Belfast, UK, 2008.

[7] M. Unehara and T. Onisawa. "Music Composition System Based on Subjective Evaluation," in *IEEE Int. Conf. on Systems, Man and Cybernetics*, 2003, pp. 980-986.

[8] M. Wanderly and N. Orio, "Evaluation of Input Devices for Musical Expression: Borrowing Tools from HCI." Computer Music Journal. 2002, vol. 26, 62-76.

[9] H. Sharp, Y. Rogers, and J. Preece, *Interaction Design,* New York, Wiley, 2007.

[10] "Tech 3286: Assessment methods for the subjective evaluation of the quality of sound programme material – Music". P. Laven, Ed., Available: http://www.ebu.ch/CMSimages/fr/tec_doc_t3286_tcm7-10487.pdf.