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

(Card et al., 1978) that their task-based approach may allow for quantitative comparisons of musical interfaces.

Wanderley and Orio's framework is interesting but may have some drawbacks. The reduction of musical interaction to maximally simple tasks risks compromising the authenticity of the interaction, creating situations in which the affective and creative aspects of music-making are abstracted away. In other words, the reduction conflates *controllability* of a musical interface with *expressiveness* of that interface (Dobrian and Koppelman, 2006). The use of Likert-scale metrics also may have some difficulties. They are susceptible to cultural differences (Lee et al., 2002) and psychological biases (Nicholls et al., 2006), and may require large sample sizes to achieve sufficient statistical power (Göb et al., 2007).

Acknowledging the relative scarcity of research on the topic of live human–computer music-making, we may look to other areas which may provide useful analogies. The field of computer games is notable here, since it carries some of the features of live music-making: it can involve complex multimodal interactions, with elements of goal-oriented and affective involvement, and a degree of learning. For example, Barendregt et al. (2006) investigates the usability and affective aspects of a computer game for children, during first use and after some practice. Mandryk and Atkins (2007) use a combination of physiological measures to produce a continuous estimate of the emotional state (arousal and valence) of subjects playing a computer game.

In summary, although there have been some useful forays into the field of expressive musical interface evaluation, and some work in related disciplines such as that of computer games evaluation, the field could certainly benefit from further development. Whilst task-based methods are suited to examining usability, the *experience* of interaction is essentially subjective and requires alternative approaches for evaluation. In this paper we hope to contribute to this area by investigating two different evaluation approaches which we have examined: Discourse Analysis and a Turing Test method.

2.2. A qualitative approach: Discourse Analysis

When a sonic interactive system is created, it is not "born" until it comes into use. Its users construct it socially using analogies and contrasts with other interactions in their experience, a process which creates the affordances and contexts of the system. This primacy of social construction has been recognised for decades in much of the social sciences and psychology, but is often overlooked by technologists.

Discourse Analysis is an analytic tradition that provides a structured way to analyse the construction and reification of social structures in discourse (Banister et al., 1994, chapter 6; Silverman, 2006, chapter 6). The source data for DA are written text, which may be appropriately transcribed interviews or conversations.

Interviews and free-text comments are sometimes reported in studies on musical interfaces. However, often they are conducted in a relatively informal context, and only quotes or summaries are reported rather than any structured analysis, therefore providing little analytic reliability. DA's strength comes from using a *structured method* which can take apart the language used in discourses (e.g. interviews, written works) and elucidate the connections and implications contained within, while remaining faithful to the content of the original text (Antaki et al., 2004). DA is designed to go beyond the specific sequence of phrases used in a conversation, and produce a structured analysis of the conversational resources used, the relations between entities, and the "work" that the discourse is doing.

DA is not a single method but an analytic tradition developed with a social constructionist basis. Discourse-analytic approaches have been developed which aim to elucidate social power relations, or the details of language use. Our interest lies in understanding the conceptual resources brought to bear in constructing socially a new interactive artefact. Therefore we derive our approach from a Foucauldian tradition of DA found in psychology (Banister et al., 1994, chapter 6), which probes the reification of existing social structures through discourse, and the congruences and tensions within.

We wish to use the power of DA as part of a qualitative and formal method which can explore issues such as expressivity and affordances for users of interactive musical systems. Longitudinal studies (e.g. those in which participants are monitored over a period of weeks or months) may also be useful, but imply a high cost in time and resources. Therefore we aim to provide users with a brief but useful period of exploration of a new musical interface, including interviews and discussion which we can then analyse.

We are interested in issues such as the user's conceptualisation of musical interfaces. It is interesting to look at how these are situated in the described world, and particularly important to avoid preconceptions about how users may describe an interface: for example, a given interface could be: an instrument; an extension of a computer; two or more separate items (e.g. a box and a screen); an extension of the individual self; or it could be absent from the discourse.

In any evaluation of a musical interface one must decide the context of the evaluation. Is the interface being evaluated as a successor or alternative to some other interface (e.g. an electric cello vs. an acoustic cello)? Who is expected to use the interface (e.g. virtuosi, amateurs, children)? Such factors will affect not only the recruitment of participants but also some aspects of the experimental setup.

Our method is designed either to trial a single interface with no explicit comparison system or to compare two similar systems (as is done in our case study). The method consists of two types of user session, solo sessions followed

by group session(s), plus the Discourse Analysis of data collected.

We emphasise that DA is a broad tradition, and there are many designs which could bring DA to bear on evaluating sonic interactions. The method described in the following is just one approach.

2.2.1. Solo sessions

In order to explore individuals' personal responses to the interface(s), we first conduct solo sessions in which a participant is invited to try out the interface(s) for the first time. If there is more than one interface to be used, the order of presentation is randomised in each session.

The solo session consists of three phases for each interface:

Free exploration: The participant is encouraged to try out the interface for a while and explore it in their own way.

Guided exploration: The participant is presented with audio examples of recordings created using the interface, in order to indicate the range of possibilities, and encouraged to create recordings inspired by those examples. This is not a precision-of-reproduction task; precision-of-reproduction is explicitly not evaluated, and participants are told that they need not replicate the examples.

Semi-structured interview (Preece et al., 2004, chapter 13): The interview's main aim is to encourage the participant to discuss their experiences of using the interface in the free and guided exploration phases, both in relation to prior experience and to the other interfaces presented if applicable. Both the free and guided phases are video recorded, and the interviewer may play back segments of the recording and ask the participant about them, in order to stimulate discussion.

The raw data to be analysed are the interview transcript. Our aim is to allow the participant to construct their own descriptions and categories, which means the interviewer must be critically aware of their own use of language and interview style, and must (as far as possible) respond to the labels and concepts introduced by the participant rather than dominating the discourse.

2.2.2. Group session

To complement the solo sessions we also conduct a group session. Peer group discussion can produce more and different discussion around a topic, and can demonstrate the group negotiation of categories, labels, comparisons, and so on. The focus-group tradition provides a well-studied approach to such group discussion (Stewart, 2007). Our group session has a lot in common with a typical focus group in terms of the facilitation and semi-structured group discussion format. In addition we make available the interface(s) under consideration and encourage the participants to experiment with them during the session.

As in the solo sessions, the transcribed conversation is the data to be analysed. An awareness of facilitation technique is also important here, to encourage all participants to speak, to allow opposing points of view to emerge in a non-threatening environment, and to allow the group to negotiate the use of language with minimal interference.

2.2.3. Data analysis

Our DA approach to analysing the data is based on that of Banister et al. (1994, chapter 6), adapted to the experimental context. The DA of text is a relatively intensive and time-consuming method. It can be automated to some extent, but not completely, because of the close linguistic attention required. Our approach is summarised in Fig. 1 and consists of the following five steps:

(a) *Transcription*: The speech data are transcribed, using a standard style of notation which includes all speech events (including repetitions, speech fragments, pauses). This is to ensure that the analysis can remain close to what is actually said, and avoid adding a gloss which can add some distortion to the data. For purposes of analytical transparency, the transcripts (suitably anonymised) should be published alongside the analysis results.

(b) *Free association*: Having transcribed the speech data, the analyst reads it through and notes down surface impressions and free associations. These can later be compared against the output from the later stages.

(c) *Itemisation of transcribed data*: The transcript is then broken down by itemising every single object in the discourse (i.e. all the entities referred to). Pronouns such as "it" or "he" are resolved, using the participant's own terminology as far as possible. For every object an accompanying description of the object is extracted from that speech instance—again using the participant's own language, essentially by rewriting the sentence/phrase in which the instance is found.

The list of objects is scanned to determine if different ways of speaking can be identified at this point. For example, there may appear to be a technical music-production way of speaking, as well as a more intuitive music-performer way of speaking, both occurring in different parts of the discourse; they may have overlaps or tensions with each other. Also, those objects which are also "actors" are identified—i.e. those which act with agency/sentience in the speech instance; they need not be human.

It is helpful at this point to identify the most commonly occurring objects and actors in the discourse, as they will form the basis of the later reconstruction.

Fig. 2 shows an excerpt from a spreadsheet used during our DA process, showing the itemisation of objects and subjects, and the descriptions extracted.

(d) *Reconstruction of the described world*: Starting with the list of most commonly occurring objects and actors in the discourse, the analyst reconstructs the depictions of the world that they produce, in terms of the interrelations between the actors and the objects. This could for example be represented using concept maps. If different ways of speaking have been identified, there will typically be one reconstructed "world" per way of speaking. Overlaps and contrasts between these worlds can be identified. Fig. 3 shows an excerpt of a concept map representing a "world" distilled in this way.

The "worlds" we produce are very strongly tied to the participant's own discourse. The actors, objects, descriptions, relationships, and relative importances, are all derived from a close reading of the text. These worlds are essentially just a methodically reorganised version of the participant's own language.

(e) *Examining context*: One of the functions of discourse is to create the context(s) in which it operates, and as part of the DA process we try to identify such contexts, in part by moving beyond the specific discourse act. For example, the analyst may feel that one aspect of a participant's discourse ties in with a common cultural paradigm of an incompetent amateur, or with the notion of natural virtuosity.

In our design we have parallel discourses originating with each of the participants, which gives us an opportunity to draw comparisons. After running the previous steps of DA on each individual transcript, we compare and contrast the described worlds produced from each transcript, examining commonalities and differences. We also compare the DA of the focus-group session(s) against that of the solo sessions.

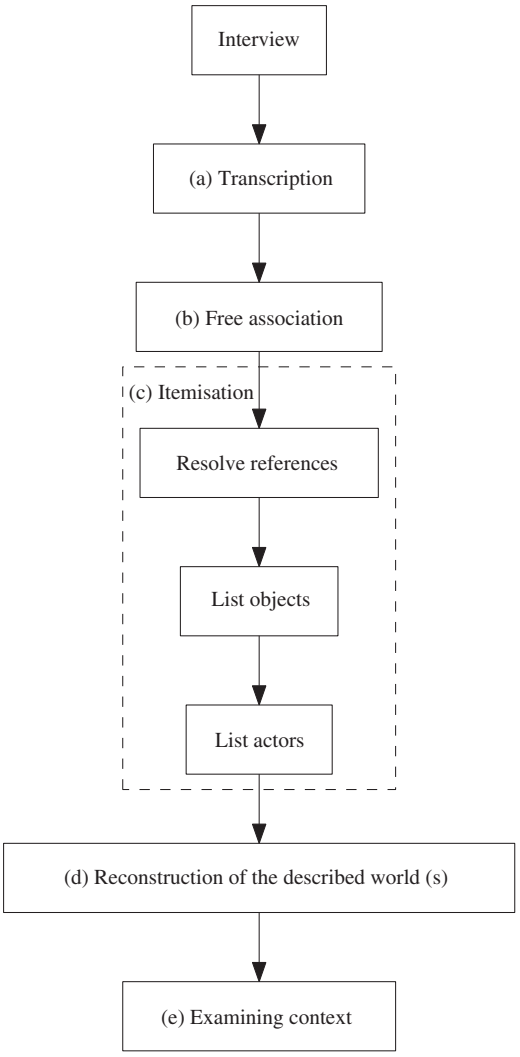

Fig. 1. Outline of our Discourse Analysis procedure.

The utility of this method will be explored through the case study in Section 3. We next consider a method designed to answer a more specific question.

2.3. A quantitative approach: a musical "Turing Test"

In interaction design, human-likeness is often a design goal (Preece et al., 2004, chapter 5). In sonic interactions and music, we may wish a system to emulate a particular

Transcription	Object (referent)	Description	Is actor?
I was trying to work out what the other person was,	Participant	was trying to work out what the other person was ((doing))	Y
yeah I'm curious to see how the other person did it,	((person recording the examples)) the other person	Ptcpt was trying to work out what this person was ((doing))	Y
Because it was more fun	((Y))	Ptcpt preferred this ((to X)) because it was more fun	
I just think the noises were a bit more,	the noises ((made by Y))	were a bit more, bit different	
you could come up with some slightly more funky noises.	((general person))	could come up with some slightly more funky noises ((in Y, than X))	Y
	noises	((general person)) could come up with some slightly more funky ones ((in Y, than X))	

Fig. 2. Excerpt from a spreadsheet used during the itemisation of interview data, for step (c) of the Discourse Analysis.

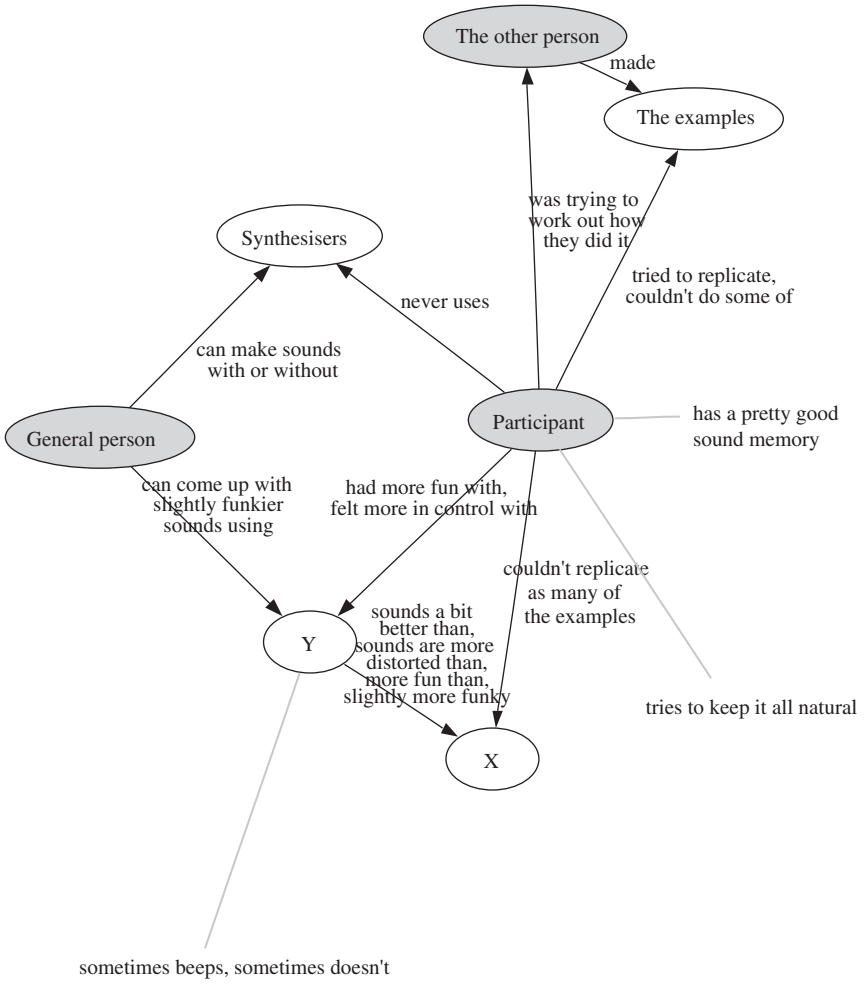

Fig. 3. An example of a reconstructed set of relations between objects in the described world. This is a simplified excerpt of the reconstruction for User 2 in our study. Objects are displayed in ovals, with the shaded ovals representing actors.

human musical ability. Therefore we employ an evaluation method that can investigate this specifically.

Turing's seminal paper (Turing, 1950) proposes replacing the question "can a computer think?", by an "Imitation Game", now commonly known as the *Turing Test*, in which the computer is required to imitate a human being in an interrogation. If the computer is able to fool a human interrogator a substantial amount of the time, then the computer can be credited with "intelligence".

There has been considerable debate around the legitimacy of this approach as a measure of artificial intelligence (e.g. Searle, 1980). However, without making any claims about the intelligence of musical systems, we can say that often they are designed with the aim of reacting or interacting in a human-like fashion. Therefore the degree of observer confusion between human and automated response is an appropriate route for evaluating systems which perform human-like tasks, such as score-based accompaniment or musical improvisation.

Algorithmic composition can involve imitation of style or the adherence to music theory rules, such as completing a four-part harmony. Pearce and Wiggins (2001) proposed a framework for the evaluation of algorithmic composition algorithms through a "discrimination test", analogous to the Turing Test, but without the element of interaction. This methodology was demonstrated by evaluating an automatic "drum and bass" composition tool. Pachet (2003) asked two judges to distinguish between his *Continuator* system and jazz pianist, Albert van Veenendaal, whose improvised playing it was emulating. David Cope also contrasted pieces in the style of Chopin, created through the use of his *Emmy* algorithm (Cope, 2001), with a lesser-known piece by the composer in front of an audience.

We are interested in the use of the Turing Test to evaluate interactive music systems. Where the computer could conceivably take the place of a skilled human, the formulation of the test can quantify the aesthetic impressions of listeners in an un-biased way. For example, a computer accompanist "learning" to play a piece with a soloist could be contrasted with an expert musician who would undertake the same rehearsal process behind a screen. Third-party observers can be used to carry out a discrimination test; however, when the soloist takes the role

of judge, the test further resembles Turing's original conception in which the judge is free to interact with the system. By analysing the degree of confusion (using a statistical test such as the Chi-Square Test), we can make numerical comparisons between systems and evaluate their relative success at this emulation.

The case study in Section 4 will look at applying a Turing Test to the evaluation of a real-time beat-tracking system. In fact we will illustrate a slight variation on the standard Turing Test approach, comparing three rather than two conditions. But our first case study is concerned with the Discourse Analysis approach.

3. Case study A: Discourse Analysis

Case study A was conducted in the context of a project to develop voice-based interfaces for controlling musical systems. Our interface uses a process we call *timbre remapping* to allow the timbral variation in a voice to control the timbral variation of an arbitrary synthesiser (Fig. 4). The procedure involves analysing vocal timbre in real-time to produce a multidimensional "timbre space", then retrieving the synthesis parameters that correspond best to that location in the timbre space. The method is described further by Stowell and Plumbley (2007).

In our study we wished to evaluate the timbre remapping system with beatboxers (vocal percussion musicians), for two reasons: they are one target audience for the technology in development; and they have a familiarity and level of comfort with manipulation of vocal timbre that should facilitate the study sessions. They are thus not representative of the general population but of a kind of "expert" user.

We recruited by advertising online (a beatboxing website) and around London for amateur or professional beatboxers. Participants were paid £10 per session plus travel expenses to attend sessions in our (acoustically isolated) university studio. We recruited five participants from the small community, all male and aged 18–21. One took part in a solo session; one in the group session; and three took part in both. Their beatboxing experience ranged from a few months to four years. Their use of technology for music ranged from minimal to a keen use of recording and effects technology (e.g. Cubase). The facilitator was a PhD student, known to the participants by his membership of the beatboxing website.

In our study we wished to investigate any effect of providing the timbre remapping feature. To this end we presented two similar interfaces: both tracked the pitch and volume of the microphone input, and used these to control a synthesiser, but one also used the timbre remapping procedure to control the synthesiser's timbral settings. The synthesiser used was an emulated General Instrument AY-3-8910 (General Instrument, early 1980s), which was selected because of its wide timbral range (from pure tone to pure noise) with a well-defined control space of a few integer-valued variables. Participants spent a total of around 30–60 min using the interfaces, and 15–20 min in interview. Analysis of the interview transcripts using the procedure of Section 2.2 took approximately 9 h per participant (around 2000 words each).

Fig. 4. Timbre remapping maps the timbral space of a voice source onto that of a target synthesiser.

We do not report a detailed analysis of the group session transcript here: the group session generated information which is useful in the development of our system, but little which bears directly upon the presence or absence of timbral control. We discuss this outcome further in Section 5.

In the following, we describe the main findings from analysis of the solo sessions, taking each user one by one before drawing comparisons and contrasts. We emphasise that although the discussion here is a narrative supported by quotes, it reflects the structures elucidated by the DA process—the full transcripts and Discourse Analysis tables are available online.[1] In the study, condition "X" was used to refer to the system with timbre remapping inactive, "Y" for the system with timbre remapping active.

3.1. Reconstruction of the described world

User 1 expressed positive sentiments about both X (without timbre remapping) and Y (with timbre remapping), but preferred Y in terms of sound quality, ease of use and being "more controllable". In both cases the system was construed as a reactive system, making noises in response to noises made into the microphone; there was no conceptual difference between X and Y—for example in terms of affordances or relation to other objects.

The "guided exploration" tasks were treated as reproduction tasks, despite our intention to avoid this. User 1 described the task as difficult for X, and easier for Y, and situated this as being due to a difference in "randomness" (of X) vs. "controllable" (of Y).

User 2 found the system (in both modes) "did not sound very pleasing to the ear". His discussion conveyed a pervasive structured approach to the guided exploration tasks, in trying to infer what "the original person" had done to create the examples and to reproduce that. In both Y and X the approach and experience was the same.

Again, User 2 expressed preference for Y over X, both in terms of sound quality and in terms of control. Y was described as more fun and "slightly more funky". Interestingly, the issues that might bear upon such preferences are arranged differently: issues of unpredictability

[1] http://www.elec.qmul.ac.uk/digitalmusic/papers/2008/Stowell08ijhcs--data/

were raised for Y (but not X), and the guided exploration task for Y was felt to be more difficult, in part because it was harder to infer what "the original person" had done to create the examples.

User 3's discourse placed the system in a different context compared to others. It was construed as an "effect plugin" rather than a reactive system, which implies different affordances: for example, as with audio effects it could be applied to a recorded sound, not just used in real-time; and the description of what produced the audio examples is cast in terms of an original sound recording rather than some other person. This user had the most computer music experience of the group, using recording software and effects plugins more than the others, which may explain this difference in contextualisation.

User 3 found no difference in sound or sound quality between X and Y, but found the guided exploration of X more difficult, which he attributed to the input sounds being more varied.

User 4 situated the interface as a reactive system, similar to Users 1 and 2. However, the sounds produced seemed to be segregated into two streams rather than a single sound—a "synth machine" which follows the user's humming, plus "voice-activated sound effects". No other users used such separation in their discourse.

"Randomness" was an issue for User 4 as it was for some others. Both X and Y exhibited randomness, although X was much more random. This randomness meant that User 4 found Y easier to control. The pitch-following sound was felt to be accurate in both cases; the other (sound effects/percussive) stream was the source of the randomness.

In terms of the output sound, User 4 suggested some small differences but found it difficult to pin down any particular difference, but felt that Y sounded better.

3.2. Examining context

Users 1 and 2 were presented with the conditions in the order XY; Users 3 and 4 in the order YX. Order-of-presentation may have some small influence on the outcomes: Users 3 and 4 identified little or no difference in the output sound between the conditions (User 4 preferred Y but found the difference relatively subtle), while Users 1 and 2 felt more strongly that they were different and preferred the sound of Y. It would require a larger study to be confident that this difference really was being affected by order-of-presentation.

In our study we are not directly concerned with which condition sounds better (both use the same synthesiser in the same basic configuration), but this is an interesting aspect to come from the study. We might speculate that differences in perceived sound quality are caused by the different way the timbral changes of the synthesiser are used. However, participants made no conscious connection between sound quality and issues such as controllability or randomness.

Taking the four participant interviews together, no strong systematic differences between X and Y are seen. All participants situate Y and X similarly, albeit with some nuanced differences between the two. Activating/deactivating the timbre remapping facet of the system does not make a strong enough difference to force a reinterpretation of the system.

A notable aspect of the four participants' analyses is the differing ways the system is situated (both X and Y). As designers of the system we may have one view of what the system "is", perhaps strongly connected with technical aspects of its implementation, but the analyses presented here illustrate the interesting way that users situate a new technology alongside existing technologies and processes. The four participants situated the interface in differing ways: either as an audio effects plugin, or a reactive system; as a single output stream or as two. We emphasise that none of these is the "correct" way to conceptualise the interface. These different approaches highlight different facets of the interface and its affordances.

The discourses of the "effects plugin" and the "reactive system" exhibit some tension. The "reactive system" discourse allows the system some agency in creating sounds, whereas an effects plugin only alters sound. Our own preconceptions (based on our development of the system) lie more in the "reactive system" approach; but the "effects plugin" discourse seemed to allow User 3 to place the system in a context along with effects plugins that can be bought, downloaded, and used in music production software.

During the analyses we noted that all participants maintained a conceptual distance between themselves and the system, and analogously between their voice and the output sound. There was very little use of the "cyborg" discourse in which the user and system are treated as a single unit, a discourse which hints at mastery or "unconscious competence". This fact is certainly understandable given that the participants each had less than an hour's experience with the interface. It demonstrates that even for beatboxers with strong experience in manipulation of vocal timbre, controlling the vocal interface requires learning—an observation confirmed by the participant interviews.

The issue of "randomness" arose quite commonly among the participants. However, randomness emerges as a nuanced phenomenon: although two of the participants described X as being more random than Y, and placed randomness in opposition to controllability (as well as preference), User 2 was happy to describe Y as being more random and also more controllable (and preferable).

A uniform outcome from all participants was the conscious interpretation of the guided exploration tasks as precision-of-reproduction tasks. This was evident during the study sessions as well as from the discourse around the tasks. As one participant put it, "If you're not going to replicate the examples, what are you gonna do?" This issue did not appear in our piloting.

A notable absence from the discourses, given our research context, was discussion which might bear on expressivity, for example the expressive range of the interfaces. Towards the end of each interview we asked explicitly whether either of the interfaces was more expressive, and responses were generally non-committal. We propose that this was because our tasks had failed to engage the participants in creative or expressive activities: the (understandable) reduction of the guided exploration task to a precision-of-reproduction task must have contributed to this. We also noticed that our study design failed to encourage much iterative use of record-and-playback to develop ideas. In Section 5 we suggest some possible implications of these findings on future study design.

We have seen the Discourse Analysis method in action and the information it can yield, about how users situate a system in relation to themselves and other objects. In the next section we will turn to consider an alternative evaluation approach based on the Turing Test, before comparing and contrasting the methods.

4. Case study B: a musical "Turing Test"

Our second case study concerns the task of real-time beat tracking with a live drummer. We have developed a beat tracker specifically for such live use, named "B-Keeper" (Robertson and Plumbley, 2007), which adjusts the tempo of an accompaniment so that it remains synchronised to a drummer.

We wished to develop a test suitable for assessing this real-time interaction. Established beat tracking evaluations exist, typically comparing annotated beat positions against ground-truths provided by human annotators (McKinney et al., 2007). However, these are designed for offline evaluation: they neglect the component of interaction, and do not attempt to judge the degree of "naturalness" or "musicality" of any variation in beat annotations.

Qualitative approaches such as Discourse Analysis described above could be appropriate. However, in this case we are interested specifically in evaluating the beat-tracker's designed ability to interact in a human-like manner, which the musical Turing Test allows us to quantify.

In our application of the musical Turing Test to evaluate the B-Keeper system, we decided to perform a three-way comparison, incorporating human, machine, and a third "control" condition using a steady accompaniment which remains at a fixed tempo dictated by the drummer. Our experiment is depicted in Fig. 5. For each test, the drummer gives four steady beats of the kick drum to set the tempo and start, then plays along to an accompaniment track. This is performed three times. Each time, a Human Tapper (one of the authors, AR) taps the tempo on the keyboard, keeping time with the drummer, but only for one of the three times will this be altering the tempo of the accompaniment. For these trials, controlled by the Human

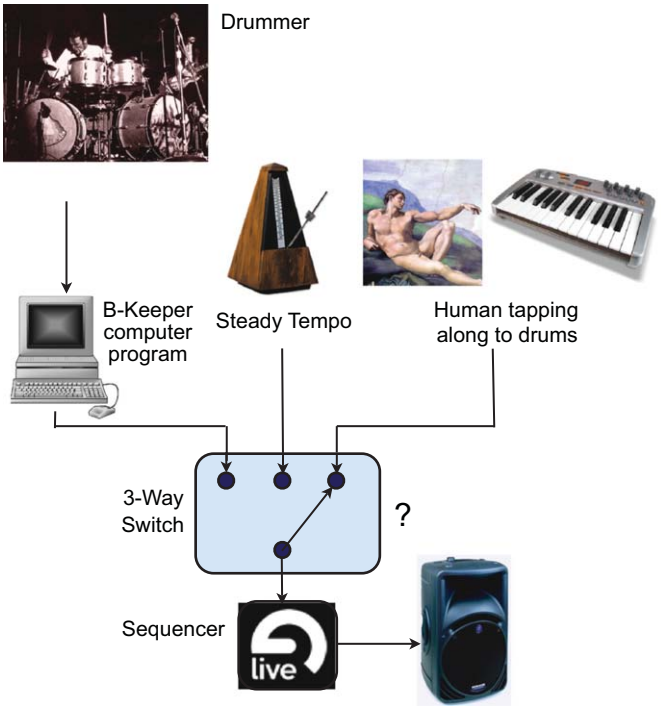

Fig. 5. Design set-up for the experiment. Three possibilities: (a) computer controls tempo from drum input; (b) Steady Tempo; (c) human controls tempo by tapping beat on keyboard.

Tapper, we applied a Gaussian window to the intervals between taps in order to smooth the tempo fluctuation, so that it would still be musical in character. Of the other two performances, one uses accompaniment controlled by the B-Keeper system and the other the same accompaniment but at a fixed tempo. The sequence in which these three trials happen is randomly chosen by the computer and only revealed to the participants after the test so that the experiment is *double-blind*, i.e. neither the researchers nor the drummer know which accompaniment is which. Hence, the quantitative results gained by asking for opinion measures and performance ratings should be free from any bias.

We are interested in the interaction between the drummer and the accompaniment which takes place through the machine. In particular, we wish to know how this differs from the interaction that takes place with the human beat tracker. We might expect that, if our beat tracker is functioning well, the B-Keeper trials would be "better" or "reasonably like" those controlled by the Human Tapper. We would also expect them to be "not like a metronome" and hence, distinguishable from the Steady Tempo trials.

We carried out the experiment with 11 professional and semi-professional drummers. All tests took place in an acoustically isolated studio space. Each drummer took the test (consisting of the three randomly selected trials) twice, playing to two different accompaniments. The first was based on a dance-rock piece first performed at Live Algorithms for Music Conference, 2006, which can be

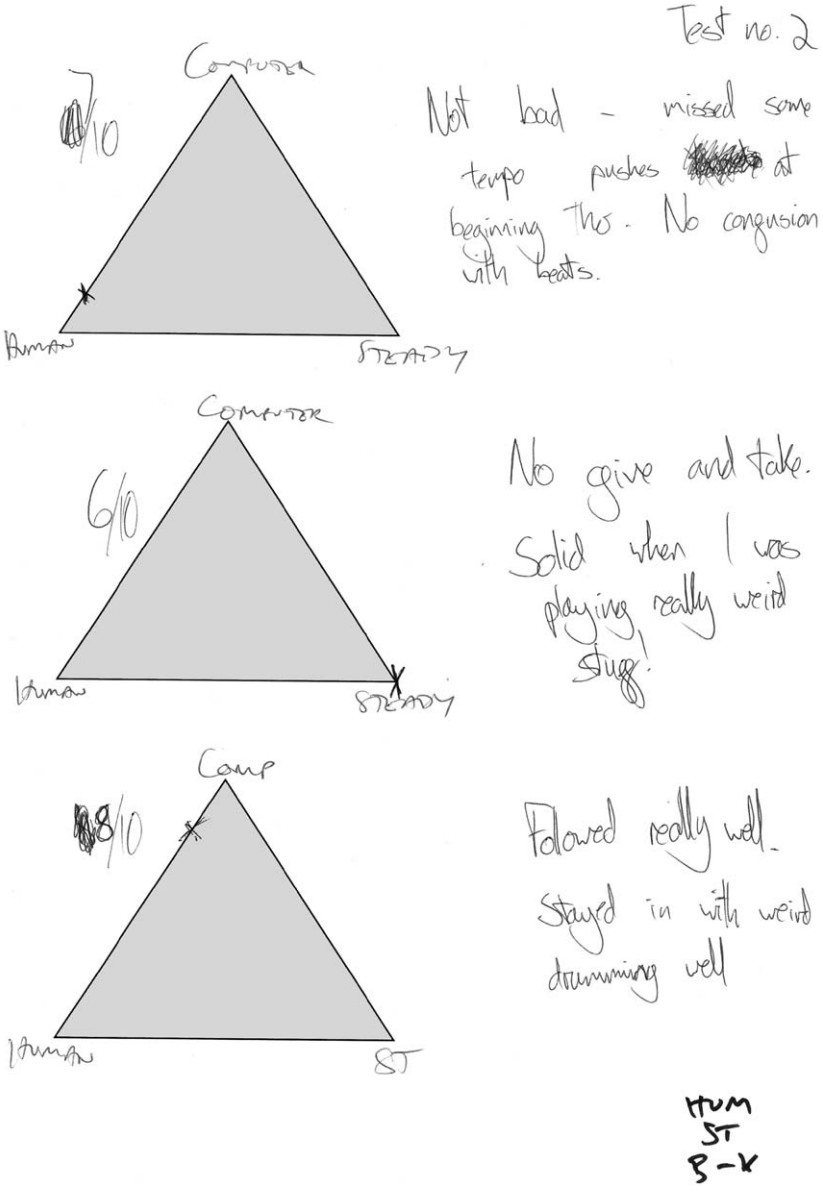

Fig. 6. Sample sheet filled in by a drummer.

viewed on the internet.[2] The second piece was a simple chord progression on a software version of a Fender Rhodes keyboard with some additional percussive sounds. The sequencer used was Ableton Live,[3] chosen for its time-stretching capabilities.

In the classic Turing Test, there would only be two possibilities: the human or the machine. However, since we wish to also contrast the beat tracker against a metronome as a control, we required a three-way choice. After each trial, we asked each drummer to mark an "X" on an equilateral triangle which would indicate the strength of their belief as to which of the three systems was responsible. The three corners corresponded to the three choices and the nearer to a particular corner they placed the "X", the stronger their belief that was the tempo-controller for that particular trial. Hence, if an "X" was placed on a corner, it would indicate certainty that was the scenario responsible. An "X" on an edge would indicate confusion between the two nearest corners, whilst an "X" in the middle indicates confusion between all three. This allowed us to quantify an opinion measure for identification over all the trials. The Human Tapper (AR) and an independent observer also marked their interpretation of the trial in the same manner.

In addition, each participant marked the trial on a scale of one to 10 as an indication of how well they believed that test worked as "an interactive system". They were also asked to make comments and give reasons for their choice. A sample sheet from one of the drummers is shown in Fig. 6.

[2] http://www.elec.qmul.ac.uk/digitalmusic/b-keeper
[3] http://www.ableton.com

4.1. Results

The participants' difficulty in distinguishing between controllers was a common feature of many tests and, whilst the test had been designed expecting that this might be the case, the results often surprised the participants when revealed, with both drummers and the researchers being mistaken in their identification of the controller. We shall contrast the results between all three tests, particularly with regard to establishing the difference between the B-Keeper trials and the Human Tapper trials and comparing this to the difference between the Steady Tempo and Human Tapper trials. In Fig. 7, we can see the opinion measures for all drummers placed together on a single triangle. The corners represent the three possible scenarios: B-Keeper, Human Tapper and Steady Tempo with their respective symbols. Each "X" has been replaced with a symbol corresponding to the actual scenario in that trial. In the diagram we can clearly observe two things:

- There is more visual separation between the Steady Tempo trials than the other two. With the exception of a relatively small number of outliers, many of the Steady Tempo trials were correctly placed near the appropriate corner. Hence, if the trial is actually steady then it will probably be identified as such.
- The B-Keeper and Human Tapper trials tend to be spread over an area centred around the edge between their respective corners. At best, approximately half of these trials have been correctly identified. The distribution does not seem to have the kind of separation seen for the Steady Tempo trials, suggesting that they have difficulty telling the two controllers apart, but could tell that the tempo had varied.

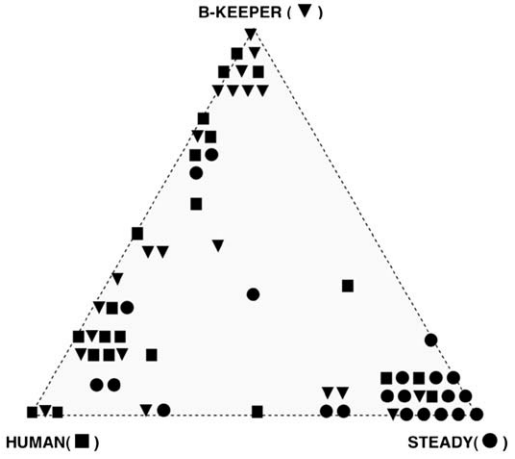

Fig. 7. Results illustrating where the 11 different drummers judged the three different accompaniments (B-Keeper, Human Tapper and Steady Tempo) in the test. The symbol used indicates which accompaniment it actually was (see corners). Where the participants have marked many trials in the same spot, as happens in the corners corresponding to Steady Tempo and B-Keeper, we have moved the symbols slightly for clarity. Hence, a small number of symbols are not exactly where they were placed. The raw data are available in co-ordinate form online (see footnote 1).

4.1.1. Analysis and interpretation

The mean scores recorded by all drummers are given in the first rows of Table 2. They show similar measures for correctly identifying the B-Keeper and Human Tapper trials: both have mean scores of 44%, with the confusion being predominantly between which of the two variable tempo controllers is operating. The Steady Tempo trials have a higher tendency to be correctly identified, with a score of 64% on the triangle.

Each participant in the experiment had a higher score for identifying the Steady Tempo trials than the other two. It appears that the Human Tapper trials are the least identifiable of the three and the confusion tends to be between the B-Keeper and the Human Tapper.

For analysis purposes, we can express the opinion measures from Fig. 7 as polarised decisions, by taking the nearest corner to be the participant's decision for that trial. In the case of points equidistant from corners, we split the decision equally. Table 3 shows the polarised decisions

Table 2
Mean identification measure results for all judges involved in the experiment.

Judge	Accompaniment track	Judged as		
		B-Keeper (%)	Human (%)	Steady (%)
Drummer	B-Keeper	**44**	37	18
	Human	38	**44**	17
	Steady	12	23	**64**
Human Tapper	B-Keeper	**59**	31	13
	Human	36	**45**	23
	Steady	15	17	**68**
Observer	B-Keeper	**55**	39	6
	Human	33	**42**	24
	Steady	17	11	**73**

Bold percentages correspond to the correct identification.

Table 3
Polarised decisions made by the drummer for the different trials.

Controller	Judged as		
	B-Keeper	Human	Steady
B-Keeper	**9.5**	8.5	4
Human Tapper	8	**10**	4
Steady Tempo	2	4	**16**

Table 4
Polarised decisions made by the drummer over the Steady Tempo and Human Tapper trials.

Controller	Judged as	
	Human Tapper	Steady Tempo
Human Tapper	12	4
Steady Tempo	5	14

made by drummers over the trials. There is confusion between the B-Keeper and Human Tapper trials, whereas the Steady Tempo trials were identified over 70% of the time. The B-Keeper and Human Tapper trials were identified 43% and 45% of the time respectively—little better than the 33% we would expect by random choice.

4.1.2. Comparative tests

In order to test the distinguishablility of one controller from the other, we performed a Chi-Square Test, calculated over all trials with either of the two controllers. If there is a difference in scores so that one controller is preferred to the other (above a suitable low threshold), then that controller is considered to be chosen for that trial. Where no clear preference was evident, such as in the case of a tie or neither controller having a high score, we discard the trial for the purposes of the test.

Thus, for any two controllers, we can construct a table of which decisions were correct. The table for comparisons between the Steady Tempo and the Human Tapper trials is shown in Table 4. We test against the null hypothesis that the distribution is the same for either controller, corresponding to the premise that the controllers are indistinguishable.

The separation between Steady Tempo and Human Tapper trials is significant ($\chi^2(3,22) = 8.24$, $p < 0.05$), meaning participants could reliably distinguish them. Partly this might be explained from the fact that drummers could vary the tempo with the Human Tapper controller but the Steady Tempo trials had the characteristic of being metronomic.

Comparing the B-Keeper trials and the Human Tapper trials, we get the results shown in Table 5. No significant difference is found in the drummers' identification of the controller for either trial ($\chi^2(3,22) = 0.03$, $p > 0.5$). Whilst B-Keeper shares the characteristic of having variable tempo and thus is not identifiable simply by trying to detect a tempo change, we would expect that if there was a *machine-like* characteristic to the B-Keeper's response, such as an unnatural response or unreliability in following tempo fluctuation, syncopation and drum fills, then the drummer would be able to identify the machine. It appeared that, generally, there was no such characteristic and drummers had difficulty deciding between the two controllers.

Table 5
Table contrasting decisions made by the drummer over the B-Keeper and Human Tapper trials.

Controller	Judged as	
	Human Tapper	B-Keeper
Human Tapper	9	8
B-Keeper	8	8

From the above, we feel able to conclude that the B-Keeper performs in a satisfactorily human-like manner in this situation.

4.1.3. Ratings

In addition to the identification of the controller for each trial, we also asked each participant to rate each trial with respect to how well it had worked as an interactive accompaniment to the drums. The frequencies of ratings aggregated over all participants (drummers, Human Tapper and independent observer) are shown in Fig. 8. The Steady Tempo accompaniment was consistently rated worse than the other two. The median values for each accompaniment are shown in Table 6. The B-Keeper system has generally been rated higher than both the Steady Tempo and the Human Tapper accompaniment.

The differences between the B-Keeper ratings and the others were analysed using the Wilcoxon signed-rank test (Mendenhall et al., 1989, section 15.4). These were found to be significant ($W = 198$ (Human Tapper) and $W = 218$ (Steady Tempo), $N = 22$, $p < 0.05$).

This analysis of user ratings is a relatively traditional evaluation, in line with Likert-scale approaches. Our results demonstrate that the framework of the musical Turing Test allows for such evaluation, but also adds the extra dimension of direct comparison with a human. It is encouraging that not only did the beat tracker generally receive a high rating whether judged by the drummer or by an independent observer, but that its performance was sufficiently human-like to confuse participants as to which was the beat tracker and which the Human Tapper (Section 4.1.2). This suggests that musically the beat tracker is performing its task well.

5. Discussion

The two evaluation methods described in Section 2 are designed to evaluate live interactive musical systems, without reducing the musical interaction to unrealistically

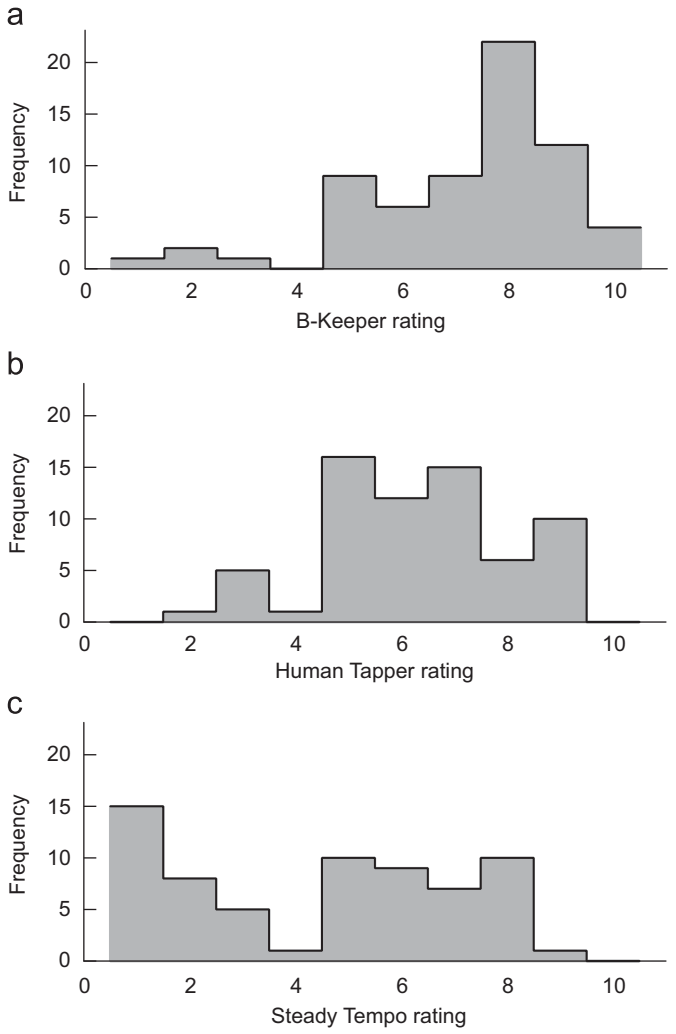

Fig. 8. Frequencies of ratings for the three scenarios: B-Keeper (upper), Human Tapper (middle) and Steady Tempo (lower).

Table 6
Median ratings given by all participants for the different scenarios.

Judge	Median rating		
	B-Keeper	Human Tapper	Steady Tempo
Drummer	7.5	5.5	5
Human	8	6.5	4
Observer	8	7	5
Combined	**8**	**6**	**5**

simple tasks. In the case studies, we have seen some of the possibilities afforded by the methods.

Firstly, the Discourse Analysis method can extract a detailed reconstruction of users' conceptualisation of a system. Our investigation of a voice-controlled interface provides us with interesting detail on the interaction between such concepts as controllability and randomness in the use of the interface, and the different ways of construing the interface itself. These findings would be difficult to obtain by other methods such as observation or questionnaire.

However, we see evidence that the discourses obtained are influenced by the experimental context: the solo sessions, structured with tasks in using both variants of our interface, produced discourse directly related to the interface; while the group session, less structured, produced wider-ranging discourse with less content bearing directly on the interface. The order of presentation also may have made a difference to the participants. It is clear that the design of such studies requires a careful balance: experimental contexts should be designed to encourage exploration of the interface itself, while taking care not to "lead" participants in unduly influencing the categories and concepts they might use to conceptualise a system.

Secondly, the musical Turing Test method can produce a quantitative result on whether a system provides an interactive experience similar to that provided by a human—despite the fact that we cannot evaluate such similarity directly. Our case study found that both participants and observers exhibited significant confusion between the B-Keeper and the Human Tapper, but not between the B-Keeper and the Steady Tempo. Preference ratings alone tell us that the B-Keeper provides a satisfactory experience, but the confusion data tell us more: that B-Keeper achieves its aim of synchronising a piece of music with the tempo variations of a live drummer, in a manner similar to that obtained if a human performs the synchronisation.

The musical Turing Test approach is of course limited to situations in which a system is intended to emulate a human musician, or perhaps to emulate some other system. It cannot be applied to the vocal timbre-mapping system of Section 3, since for that there is no reference against which to compare. However, emulation of human abilities is not uncommon in the literature: for example the Continuator (Pachet, 2003), designed to provide a naturalistic "call and response" interaction with a keyboard player; or BBCut (Collins, 2006), designed to produce real-time "beat-slicing" effects like a Drill'n'bass producer. A more general method such as our DA method could be used on these systems, and could produce useful information about users' cognitive approach to the systems, perhaps even illuminating the extent of human-like affordances. However, the musical Turing Test gives us a more precise analysis of this specific facet of musical human–computer interaction, and for example enables numerical comparison between two systems.

Our two methods reflect two approaches, qualitative and quantitative; but further, they illustrate two philosophical attitudes. The musical Turing Test derives useful numerical results, but at the cost of imposing a predetermined conceptual framework onto the interactive situation, including for example the concept of humanlikeness which can be attained to a greater or lesser degree. Our qualitative DA-based approach represents a fairly strong social constructionist attitude in which the key categories and

concepts are not predetermined but are considered an important outcome of the analysis. This can reveal aspects of users' conceptualisation of an interface, but has the drawback that results are difficult to reduce down to simple comparable elements such as statistics.

Having explored our two methods, we are in a position to compare and contrast them with approaches used by other investigators, and then to work towards recommendations on the applicability of different methods to different contexts.

5.1. Comparison with other approaches

A useful point of comparison is the approach due to Wanderley and Orio (2002), involving user trials on "maximally simple" tasks followed by Likert-scale feedback. As previously discussed, this approach raises issues of task authenticity, and of the suitability of the Likert-style questionnaire. Indeed, Kiefer et al. (2008) investigate the Wanderley and Orio approach, and find qualitative analysis of interview data to be more useful than quantitative data about task accuracy. The Wanderley and Orio method may therefore only be appropriate to cases in which the test population is large enough to draw conclusions from Likert-scale data, and in which the musical interaction can reasonably be reduced or separated into atomic tasks. We suggest the crossfading of records by a DJ as one specific example: it is a relatively simple musical task that may be operationalised in this way, and has a large user-base. (We do not wish to diminish the DJ's art: there are creative and sophisticated aspects to the use of turntables, which may not be reducible to atomic tasks.)

One advantage of the Wanderley and Orio method is that Likert-scale questionnaires are very quick to administer and analyse. In our case study of the Discourse Analysis approach, the ratio of interview time to analysis time was approximately 1:30 or 1:33, a ratio slightly higher than the ratio of 1:25–1:29 reported for observation analysis of video data (Barendregt et al., 2006). This long analysis time implies practical limitations for large groups.

Our approaches (as well as that of Wanderley and Orio) are "retrospective" methods, based on users' self-reporting after the musical act. We have argued that concurrent verbal protocols and observation protocols are problematic for experiments involving live musicianship. A third alternative, which is worthy of further exploration, is to gather data via physiological measurements. Mandryk and Atkins (2007) present an approach which aims to evaluate computer-game-playing contexts, by continuously monitoring four physiological measures on computer-game players, and using fuzzy logic to infer the players' emotional state. Analogies between the computer-gaming context and the music-making context suggest that this method could be adopted for evaluating interactive music systems. However, there are some issues which would need to be addressed:

- Most importantly, the inference from continuous physiological variables to continuous emotional state requires more validation work before it can be relied on for evaluation.
- The evaluative role of the inferred emotional state also needs clarification: the mean of the *valence* (the emotional dimension running from happiness to sadness) suggests one simple figure for evaluation, but this is unlikely to be the whole story.
- Musical contexts may preclude certain measurements: the facial movements involved in singing or beatboxing would affect facial electromyography (Mandryk and Atkins, 2007), and the exertion involved in drumming will have a large effect on heart-rate. In such situations, the inference from measurement to emotional state will be completely obscured by the other factors affecting the measured values.

We note that the literature, the present work included, is predominantly concerned with evaluating musical interactive systems from a performer-centred perspective. Other perspectives are possible: a composer-centred perspective (for composed works), or an audience-centred perspective. We have argued in Section 1 that the performer should typically be the primary focus of evaluation; but in some situations it may be appropriate to perform e.g. audience-centred evaluation. Our methods can be adapted for use with audiences—indeed, the independent observer in our musical Turing Test case study takes the role of audience. However, for audience-centred evaluations it may be the case that other methods are appropriate, such as voting or questionnaire approaches for larger audiences. Labour-intensive methods such as DA will tend to become impractical with large audience groups.

A further aspect of evaluation focus is the difference between solo and group music-making. Wanderley and Orio's set of simple musical tasks is only applicable for solo experiments. Our evaluation methods can apply in both solo and group situations, with the appropriate experimental tasks for participants. The physiological approach may also apply equally well in group situations.

6. Recommendations

From our studies, we suggest that an investigator wishing to formally evaluate an interactive music system, or live music interface, should consider the following:

(1) *Is the system primarily designed to emulate the interaction provided by a human, or by some other known system?*If so, the musical Turing Test method can be recommended.
(2) *Is the performer's perspective sufficient for evaluation?*In many cases the answer to this is "yes", although there

may be cases in which it is considered important to design an experiment involving audience evaluation. Many of the same methods (interviews, questionnaires, Turing-Test choices) are applicable to audience members—and because audiences can often result in large sample sizes compared against performer populations, survey methods such as Likert scales are more likely to be appropriate.

(3) *Is the system designed for complex musical interactions, or for simple/separable musical tasks?*If the latter, then Wanderley and Orio's approach using simplified tasks may hold some attraction. If the former, then we recommend a more situated evaluation such as our Discourse Analysis approach, which avoids the need to reduce the musical interaction down to atomic tasks.

(4) *Is the system intended for solo interaction, or is a group interaction a better representation of its expected use pattern?*The experimental design should reflect this, using either solo or group sessions.

(5) *How large is the population of participants on which we can draw for evaluation?*Often the population will be fairly small, which raises issues for the statistical power of quantitative approaches. Qualitative approaches should then be considered.

One of the key themes in our recommendations is that the design of an evaluation experiment should aim as far as possible to reflect an authentic context for the use of the system. Experimental design should include a phase which encourages use and exploration of the system. Approaches such as our Discourse Analysis of interview data can then be applied in a wide variety of cases to probe the participants' cognitive constructs produced during the experiment. Discourse Analysis is not the only way to analyse interview data (Silverman, 2006), and others may be worth pursuing; we have argued for Discourse Analysis as a principled approach which extracts a structured picture of the described world from a relatively small amount of interview data.

In any design using interview data, it is important that the facilitator has a reflexive awareness of their own use of language, able to avoid "leading" participants in their choice of concepts and language. It is also important that the reporting of the experiment demonstrates the difference between formal and informal qualitative analysis: a formal qualitative analysis makes clear the route from data to conclusions, by describing the methodological basis and the steps taken to process the data, and ideally by publishing transcripts, etc.

Approaches based on continuous physiological measures (Mandryk and Atkins, 2007) may become viable for evaluating interactive systems, although there are at present some issues to be resolved, discussed above. We consider this a topic for future research, rather than an approach to be generally recommended at present, although we look forward to developments in this area.

Finally, from our experience we repeat the advice given by others (Kiefer et al., 2008) that the importance of piloting should not be underestimated, as it can reveal issues with an experimental design that do not otherwise become apparent beforehand.

7. Conclusions

Traditional HCI evaluation methods, such as task-oriented methods or "talk-aloud" protocols, have problems when applied to live human–computer music-making. In this article we have considered approaches that may usefully be applied to evaluating such interactions, and have presented two specific methods—based on Discourse Analysis and on the Turing Test—along with case studies.

Our Discourse Analysis method aims to characterise the conceptual structures participants bring to bear in rendering an interface in their social context. Our musical interactive Turing Test is intended for a more specific situation, where a system aims to emulate some aspect of human musical performance. Both approaches succeed in evaluating sonic interactive systems from a performer-oriented perspective, without reducing the interaction to atomic tasks that might compromise the authenticity of the situation.

We hope that our recommendations (Section 6) may be a useful starting-point for others to conduct evaluations in authentic musical contexts. More generally, we believe that this area is underexplored and needs much more research, such as the further development of structured approaches to analysing user talk (both within and outside the traditions of Discourse Analysis), or the application of physiological measures to music-making situations.

Acknowledgements

Dan Stowell and Andrew Robertson are supported by Doctoral Training Account research studentships from the EPSRC.

References

Antaki, C., Billig, M., Edwards, D., Potter, J., 2004. Discourse analysis means doing analysis: a critique of six analytic shortcomings. Discourse Analysis Online 1 (1). URL: ⟨http://extra.shu.ac.uk/daol/articles/v1/n1/a1/antaki2002002-paper.html⟩.

Banister, P., Burman, E., Parker, I., Taylor, M., Tindall, C., 1994. Qualitative Methods in Psychology: A Research Guide. Open University Press, Buckingham.

Barendregt, W., Bekker, M.M., Bouwhuis, D., Baauw, E., 2006. Identifying usability and fun problems in a computer game during first use and after some practice. International Journal of Human–Computer Studies 64 (9), 830–846.

Borchers, J.O., 2001. A pattern approach to interaction design. AI & Society 15 (4), 359–376.

Buxton, W., Sniderman, R., 1980. Iteration in the design of the human–computer interface. In: Proceedings of the 13th Annual Meeting, Human Factors Association of Canada, pp. 72–81.

Card, S., Moran, T., Newell, A., 1983. The Psychology of Human–Computer Interaction. Lawrence Erlbaum Associates, Hillsdale, NJ.

Card, S.K., English, W.K., Burr, B.J., 1978. Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT. Ergonomics 21 (8), 601–613.

Collins, N., 2006. BBCut2: integrating beat tracking and on-the-fly event analysis. Journal of New Music Research 35 (1), 63–70.

Collins, N., d'Escrivan, J. (Eds.), 2007. The Cambridge Companion to Electronic Music. Cambridge University Press, Cambridge.

Cope, D., 2001. Virtual Music: Computer Synthesis of Musical Style. MIT, Cambridge, MA.

de Poli, G., 2004. Methodologies for expressiveness modelling of and for music performance. Journal of New Music Research 33 (3), 189–202.

Dobrian, C., Koppelman, D., 2006. The 'E' in NIME: musical expression with new computer interfaces. In: Proceedings of New Interfaces for Musical Expression (NIME). IRCAM, Centre Pompidou Paris, France, pp. 277–282. URL: ⟨http://www.nime.org/2006/proc/nime2006_277.pdf⟩.

Ericsson, K.A., Simon, H.A., 1996. Protocol Analysis: Verbal Reports as Data, revised ed. Massachusetts Institute of Technology, Cambridge, MA.

Fels, S., 2004. Designing for intimacy: creating new interfaces for musical expression. Proceedings of the IEEE 92 (4), 672–685.

General Instrument, early 1980s. GI AY-3-8910 Programmable Sound Generator datasheet.

Göb, R., McCollin, C., Ramalhoto, M.F., 2007. Ordinal methodology in the analysis of Likert scales. Quality and Quantity 41 (5), 601–626.

Goebl, W., 2004. Computational models of expressive music performance: the state of the art. Journal of New Music Research 33 (14), 203–216.

Grant, S., Aitchison, T., Henderson, E., Christie, J., Zare, S., McMurray, J., Dargie, H., 1999. A comparison of the reproducibility and the sensitivity to change of visual analogue scales, Borg scales, and Likert scales in normal subjects during submaximal exercise. Chest 116 (5), 1208–1217.

Hunt, A., Wanderley, M.M., 2002. Mapping performer parameters to synthesis engines. Organised Sound 7 (2), 97–108.

Kiefer, C., Collins, N., Fitzpatrick, G., 2008. HCI methodology for evaluating musical controllers: a case study. In: Proceedings of the International Conference on New Interfaces for Musical Expression (NIME), pp. 87–90.

Lee, J.W., Jones, P.S., Mineyama, Y., Zhang, X.E., 2002. Cultural differences in responses to a Likert scale. Research in Nursing and Health 25 (4), 295–306.

Levitin, D.J., McAdams, S., Adams, R.L., 2003. Control parameters for musical instruments: a foundation for new mappings of gesture to sound. Organised Sound 7 (2), 171–189.

Mandryk, R.L., Atkins, M.S., 2007. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. International Journal of Human–Computer Studies 65 (4), 329–347.

McKinney, M.F., Moelants, D., Davies, M.E.P., Klapuri, A., 2007. Evaluation of audio beat tracking and music tempo extraction algorithms. Journal of New Music Research 36 (1), 1–16.

Mendenhall, W., Wackerly, D.D., Scheaffer, R.L., 1989. Mathematical Statistics with Applications, fourth ed. PWS-Kent.

Nicholls, M.E.R., Orr, C.A., Okubo, M., Loftus, A., 2006. Satisfaction guaranteed: the effect of spatial biases on responses to Likert scales. Psychological Science 17 (12), 1027–1028.

Pachet, F., 2003. The Continuator: musical interaction with style. Journal of New Music Research 32 (3), 333–341.

Pearce, M., Wiggins, G., 2001. Towards a framework for the evaluation of machine compositions. In: Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science, pp. 22–32.

Peretz, I., Zatorre, R.J., 2005. Brain organization for music processing. Annual Review of Psychology 56 (1), 89–114.

Polfreman, R., 2001. A task analysis of music composition and its application to the development of Modalyser. Organised Sound 4 (1), 31–43.

Preece, J., Rogers, Y., Sharp, H., 2004. Interaction Design. Wiley, New York.

Robertson, A., Plumbley, M.D., 2007. B-Keeper: a beat-tracker for live performance. In: Proceedings of the International Conference on New Interfaces for Musical Expression (NIME), New York, USA, pp. 234–237.

Salamé, P., Baddeley, A., 1989. Effects of background music on phonological short-term memory. The Quarterly Journal of Experimental Psychology Section A 41 (1), 107–122.

Searle, J., 1980. Minds, brains and programs. Behavioural and Brain Sciences 3, 417–457.

Silverman, D., 2006. Interpreting Qualitative Data: Methods for Analysing Talk, Text and Interaction, second ed. Sage Publications Inc, Beverley Hills, CA.

Stewart, D.W., 2007. Focus Groups: Theory and Practice. Sage Publications, Beverley Hills, CA.

Stowell, D., Plumbley, M.D., 2007. Pitch-aware real-time timbral remapping. In: Proceedings of the Digital Music Research Network (DMRN) Summer Conference, July. URL: ⟨http://www.elec.qmul.ac.uk/digitalmusic/papers/2007/StowellPlumbley07-dmrn.pdf⟩.

Turing, A., 1950. Computing machinery and intelligence. Mind 59, 433–460.

Wanderley, M.M., Orio, N., 2002. Evaluation of input devices for musical expression: borrowing tools from HCI. Computer Music Journal 26 (3), 62–76.

Wharton, C., Rieman, J., Lewis, C., Polson, P., 1994. The cognitive walkthrough method: a practitioner's guide. In: Nielsen, J., Mack, R. (Eds.), Usability Inspection Methods. Wiley, New York, pp. 105–140.