Evaluation of an Interactive Music Performance System in the Context of Irish Traditional Dance Music

Marco Amerotti
KTH Royal Institute of Technology
Stockholm, Sweden
amerotti@kth.se

Bob L. T. Sturm
KTH Royal Institute of Technology
Stockholm, Sweden
bobs@kth.se

Steve Benford
University of Nottingham
Nottingham, United Kingdom
steve.benford@nottingham.ac.uk

Hugo Maruri-Aguilar
Queen Mary University of London
London, United Kingdom
h.maruri-aguilar@qmul.ac.uk

Craig Vear
University of Nottingham
Nottingham, United Kingdom
craig.vear@nottingham.ac.uk

ABSTRACT

We present a preliminary evaluation of an interactive, real-time, and co-creative performance system for Irish Traditional Dance music. We focus on how this musical partnership is experienced by a human musician performing with it in four aspects: enjoyability, musicality, humanness and responsiveness. Our preliminary study with seven traditional musicians reveals that they find playing with the system to be enjoyable, and appreciated its musicality; but they scored its humanness and responsiveness less highly. These findings suggest that such real-time performance systems might bring an enjoyable "otherness" to musical performance, even for traditional forms of music. Finally, we discuss experimental considerations for a future study involving more participants.

Author Keywords

Interactive Performance Modeling, Performance Evaluation, Traditional Music

CCS Concepts

•**Applied computing** → **Sound and music computing;** *Performing arts;* •**Human-centered computing** → *User studies;*

1. INTRODUCTION

Music performance modeling targets the rendering of realistic performances whether improvised or from a given score, solo or together with musicians. The literature mostly addresses classical music between the 18th and 20th centuries, particularly solo piano performance [7]. There is a relative abundance of datasets of classical music, e.g., MAESTRO [12], whereas other styles lack well-organized corpora of aligned scores and performances. There is not much research in real-time interactive performance modeling; when systems are considered "interactive" it is in the sense that they allow for some parameter choice or performance planning, without any reaction to a simultaneous musical performance (exceptions include [2, 6, 10, 9, 8]). Finally, the evaluation of music performance modeling has yet to converge on a specific methodology.

We consider an interactive music performance modeling system engineered for Irish traditional dance music [1]. How well does this system do in the context of co-creative performance? How can we measure its success, quantitatively and qualitatively? This kind of music performance is interesting because it features extensive elaboration of traditional melodies, from simple ornamentation to complex extemporization; it can be played both solo and in a heterophonic fashion, enabling the study of musical interaction; finally, it allows us to investigate what domain-specific issues arise when modeling traditional music performance.

2. RELATED WORK

The RENCON experience [15] is interesting to consider: it consists of multiple performance modeling contests between 2002 and 2011, where participating systems compete to render a peformance. Various evaluation methodologies have been employed, mostly using a qualitative number-based scale on high-level attributes, such as "naturalness" or "artificial or human". One contest included an open-ended questionnaire. Alongside a piece of choice by the contestant, almost all contests require a mandatory classical piece to be performed. The RENCON experience considers also "interactive" systems [15], but the focus is on interactive performance planning, rather than real-time co-performance.

In [4], emotional descriptors are mentioned as viable high-level control descriptors for interactive real-time systems. To artistic quality and expressivity, performance recordings and listening tests are mentioned (e.g., as in [13, 14, 11, 16]), while usability and level of engagement tests can help assess the quality of interaction. Questionnaires are useful to obtain a general idea of the evaluated system from the users' perspective.

To the best of our knowledge, very few systems target the issue of real-time steerable performance. One example is pDM [10], a real-time version of Director Musices [5] and the CaRo [6] system. However, both systems lack any in-depth evaluation, apart from taking part in RENCON.

NIME'24, 4–6 September, Utrecht, The Netherlands.

3. MODELING IRISH TRADITIONAL DANCE MUSIC PERFORMANCE

"LOERIC" [1] is a real-time, rule-based performance system for Irish traditional dance music. The version of LOERIC we evaluate is the same as in the original paper, with the exception that the dynamics control function was averaged with a high-loud correlation function, derived from empirical testing: higher pitches will tend to have higher dynamics than lower pitches. LOERIC allows for real-time interaction and steering of the performance: a MIDI control change signal representing a desired level of "intensity" can be fed into the system. For this experiment, system-participant interaction is achieved as follows: while LOERIC is running, a concurrent program monitors microphone input, dynamically mapping amplitude between 0 and 127 according to a temporal neighborhood. This information is sent as a control signal to LOERIC ten times per second using MIDI. LOERIC then computes the weighted sum of several control functions with the user-generated control value (details in [1]). The weight of the performer's control signal is controlled by the "human impact" parameter, which we vary in our experiment. This simple setup allows LOERIC to adapt to the participant's playing and, although naive, this approach has proved to be a valid and enjoyable interaction modality in our previous testing.

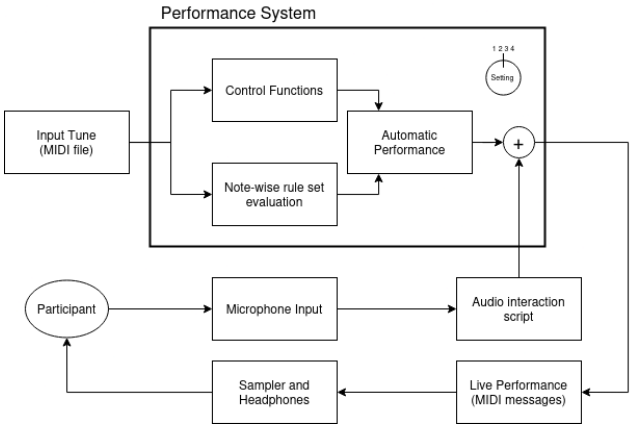

Figure 1: Our experimental setup. The performance system evaluates control functions and rule sets to generate an automatic performance. An external control signal is computed from the microphone input of the participant and then fed into the system to steer the performance. The generated performance is synthesized using a sampler and output to the participant via headphones.

4. METHODS

We conducted preliminary experiments by playing with the system to assess what factors should be considered. We were interested in seeing how the experience of performing with and without LOERIC changes, that is playing against an unexpressive MIDI playback of a tune, versus playing the same tune but interpreted by LOERIC with varying human impact. We compare four performance systems: unexpressive MIDI playback (S_1), LOERIC with no interaction (human impact set to 0, S_2), LOERIC with partial human control (human impact set to 0.5, S_3), and LOERIC with complete human control (human impact set to 1, S_4).

Our experimental procedure is as follows for a participant:

- the participant books the experiment and specifies what instrument they will use and what tune they will play, providing a hyperlink to a score;
- participant is introduced to the experiment;
- participant reads an information sheet and signs a consent form;
- participant is asked to tune their instrument;
- participant is asked to confirm their choice of tune and to play a fragment of it to determine their preferred tempo;
- participant to fill in the first page of the questionnaire, featuring their name and self-assessed level in Irish traditional music;
- when ready, a metronome plays middle C for 16 beats at 120 BPM, and participants are asked to match it, to later align the recordings;
- the trial begins and the participant plays their tune twice with the first system;
- the trial ends with the participant answering the four questions;
- the above two steps are repeated for the three other systems;
- finally, the order of the systems is revealed and the participant is asked to comment on the experience.

Given what we found in the literature, we choose to ask high-level questions concerning four attributes, these being enjoyability, humanness, musicality and responsiveness on a 10-point Likert scale. The four questions we ask after each trial are:

- **Q1**: "How enjoyable was playing with the system?" (1: not at all, 10: very enjoyable);
- **Q2**: "How human did the performance feel?" (1: not at all, 10: very human);
- **Q3**: "How musical did the performance feel?" (1: not at all, 10: very musical);
- **Q4**: "How responsive did the performance feel?" (1: not at all, 10: very responsive).

Q1 attempts to capture how enjoyable the experience with the system is; Q2, how much the system feels like a human performer; Q3, how musically coherent and stylistically appropriate the performance is, considering the context and practice of Irish Traditional Dance music; Q4, how responsive and adaptive the system feels.

The order of the systems for each participant is randomized and concealed. Each participant wears headphones to hear the system, which uses a sampled mandolin-like instrument.[1] We use this sound over other options because it provides a strong attack and is easier to hear. The entirety of each experiment is recorded.

[1]https://www.decentsamples.com/product/stella-mandolin

5. RESULTS

In total, $N = 7$ participants took part in the experiment, of which four self-assessed their Irish music level/experience as "Beginner", two as "Intermediate" and one as "Professional". The instruments of choice were flute (1), tin whistle (1), Irish flute (1), violin/fiddle (3), and acoustic guitar (1). The tunes selected by the participants were four jigs ("The Lilting Banshee", "The Rolling Waves", "The Wheels of the World", "Paddy Fahy's Jig") and three reels ("Drowsy Maggie", "The Green Fields of Rossbeigh" and "Bag of Spuds"). Two participants used sheet music as an aid. Below we first present the quantitative results, and then the qualitative results.

Table 1: Mean rating for each question and system with unbiased standard deviation. Each combination is computed with 7 measurements, one from each participant.

	Setting 1	Setting 2	Setting 3	Setting 4
Q1	6.43 ± 1.62	7.0 ± 1.63	7.0 ± 1.53	5.14 ± 2.41
Q2	5.29 ± 1.98	6.29 ± 1.25	6.0 ± 1.83	4.0 ± 2.45
Q3	5.43 ± 1.4	7.0 ± 1.83	6.29 ± 1.25	4.29 ± 1.6
Q4	4.43 ± 2.64	5.71 ± 2.14	5.71 ± 1.98	4.43 ± 2.44

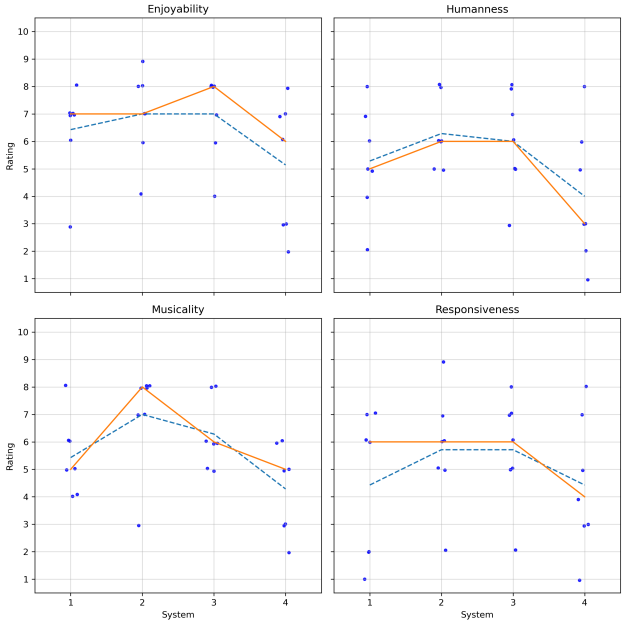

Figure 2: Ratings for each question across systems. The median is shown as a solid line, and the mean as a dashed line. Top left: enjoyability. Top right: humanness. Bottom left: musicality. Bottom right: responsiveness.

	System 1	System 2	System 3	System 4
Trial 1	8	4	0	16
Trial 2	4	4	12	8
Trial 3	4	12	12	0
Trial 4	12	8	4	4

Table 2: Number of observations for each system and trial pair. The factors are not orthogonal as the conditions in [3] are not met.

Figure 2 shows the ratings for each question across systems and Table 1 shows their mean and unbiased standard deviation. We analyze this data in the following section.

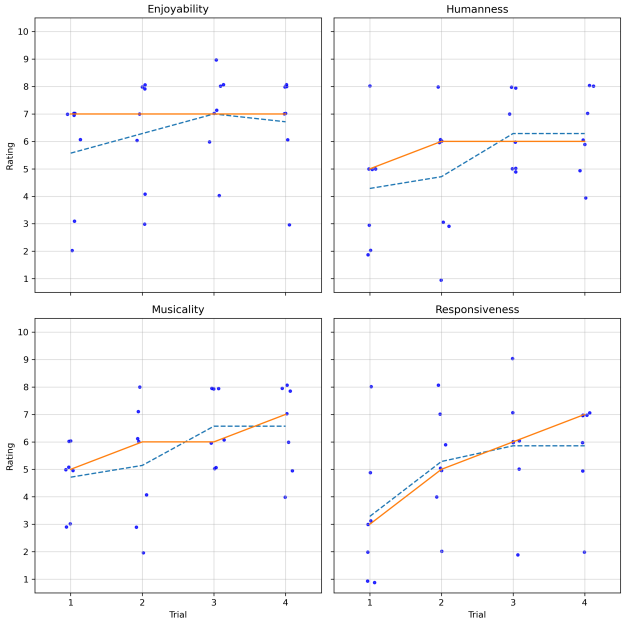

Figure 3: Ratings for each question across trials. The median is shown as a solid line, and the mean as a dashed line. Top left: enjoyability. Top right: humanness. Bottom left: musicality. Bottom right: responsiveness.

We now report some of the feedback participants left. Interviews were conducted in English, which was not the first language for any of the participants.

- **Participant 1** [Beginner]: *My rating of enjoyability increased from system to system, by having played with it for a while you get a sense of what to expect. With the second [S_3] and fourth [S_2] one I noticed unexpected individual notes, it felt like interacting with something that has its own ideas, but awkward in the sense that I didn't feel it fit. I tested the system to follow a more "round" rhythm (i.e. "swung"), and it wouldn't follow me on that. I feel it difficult to comment on whether the ornamentation is human-like, my knowledge of Irish music is thin, and the sound kind of distracted me. I didn't feel a huge difference between the systems.*

- **Participant 2**[Beginner]: *When you call someone on the phone and you get a playback of your own voice, I felt that a bit. It did some interesting things and that is why I laughed. I said it was very enjoyable, but that's because it sounded funny.*[2] *It felt like it amplified my wrongs sometimes.*

- **Participant 3**[Professional]: *The first two [S_2 and S_1] were difficult to discern from one another. The third one [S_3] was the most pleasant, I could match it like a banjo player, there were some notes that would be slightly more marked, and others that it would take out, it felt quite close to what I would hear from an Irish banjo player. The fourth one [S_4] overdid it. I chose the speed in the beginning, but normally I use quite a bit of swing, while the system didn't. That was one of the main things that were saying "this is a machine".*

[2]The participant is referring to the sound of slides through the sampling synthesizer.

- **Participant 4**[Beginner]: *The second one [S_4] sounded very weird.*[3] *It was the easiest the fourth time. I had the feeling that some of the notes the system produced were based on my playing. Adaptiveness was not so easy to understand. If I had slowed down, would the system have done the same? When I played wrong notes, did the system pick that up?*

- **Participant 5**[Beginner]: *I put very low rankings on responsiveness because I didn't manage to get what I expected from the system. I felt it was too slow, I tried to speed up and it didn't, maybe a bit in the second [S_2] or third one [S_3]. In the first one [S_4], if it hadn't been the first one, my rating might have changed because I started trying to get it to do what I wanted later in the experiment. The second one sometimes almost faded away and stopped playing and that made me focus too much on it. For the third one [S_3], it was very boring. In the other ones, I was constantly trying to guess what it was trying to do and what I was supposed to do to get it to do what I wanted to do.*

- **Participant 6**[Intermediate]: *The second [S_3] and third [S_2] ones felt better, but I think that I got better at playing with the system. It felt responsive. The last one [S_1] was all over the place. With that sound, some of the expression felt out of place. It feels weird that I rated very high the third one [S_2], I was adapting to the system and not the other way around. It's a nice experience but it needs to be practiced, the same way that you practice with a human.*

- **Participant 7**[Intermediate]: *The first time you get nervous about coming in at the right time. In the B part, I played differently.*[4] *The tone of the mandolin is a little hard: it would be more comfortable to have a softer sound. It could be developed into a very useful pedagogic instrument. Interaction is very interesting.*

6. DISCUSSION

Considering the theory of experimental design as in [3], this experiment can be studied as a row-column design, involving two sets of blocking factors – "trial" and "participant" – and to each combination of these factors, the treatment "system" was applied. These factors need to be checked for orthogonality, to validate that conclusions from the data can be stated independently for each factor in the study. A pairwise check between the factors trial, system and participants shows that the pair trial-system (Table 2) is not orthogonal, as per Theorem 10.5 in [3]. We cannot benefit from the removal of variability when we include the blocking factors in the study: thus, we cannot say that the order in which participants experienced each system does not have an impact on their ratings. Some participants explicitly felt this was the case. Careful consideration of the experimental design is needed to account for such an effect, as discussed in the next section.

We now examine the ratings for each question. Considering enjoyability, S_2 and S_3 obtained higher mean scores. These correspond to LOERIC with a value of human impact of 0 and 0.5, respectively. S_1 (unexpressive MIDI playback) has the third highest mean score, and LOERIC having human impact of 1 is rated lowest. This could be related to the very high variability in tempo, dynamics, and ornaments introduced with the last parameter setting. The ratings of humanness follow a similar trend, with a more demarcated separation between S_1, S_2, S_3, and S_4.

Considering musicality, S_2 (LOERIC without any interaction) receives the highest marks. This could be related to the consistency of the performance in its expressive parameters, without unexpected changes deriving from participant-system interaction. The ratings for responsiveness do not match the nature of each system: there is no difference in score between LOERIC with and without interaction. Furthermore, S_4 is rated in the same way as S_1. One speculation is that the high sensitivity of the control signal in S_4 was perceived as a lack of control on the musician's side; in fact variability and lack of control are recurrent themes in the participant's feedback.

When analyzing the latter, we find recurring considerations (participants 1, 4, 5, 6) on how the experience changes as the experiment progresses. Another common observation concerns the score-following abilities of the system. Participants 1, 4, and 7 played some variations of the tune they selected and wished for the performance system to adapt to their playing.

Participants 1 and 3 in particular commented on the lack of swing; for participant 3, who is also a professional traditional musician, this immediately gave away that they were playing with a machine. We also find comments on the weirdness of the system and wrong notes (participants 2, 4, and 6). These refer to the "slide" ornament, implemented through MIDI pitch bend messages, the rendition of which can vary greatly depending on the synthesis software. This might have impacted the ratings of humanness and musicality, implying that choosing an appropriate sound plays an important role in the experiment's design. Participants 1 and 7 found the sound to be distracting or unpleasant. Regarding wrong notes, participants found them sometimes out of place and confusing during playing. The implementation of "errors" naively relies on a random variation in pitch within a certain interval, but this is not an effective strategy.

Participants 4 and 5 reported that it was unintuitive to understand what "responsiveness" meant, not having any previous expectations or references. A more comprehensive experiment could include longer interaction times with the system. No system obtained a mean score or median over 6 for "responsiveness"; nonetheless, the performance system offered a playground for participants to try out their musical ideas, which makes it a valid sandbox to conduct research in musical interaction.

Finally, participant 5 noted how playing with the system requires practice in the same way as playing with a human, and participant 7 expressed LOERIC's validity as a possible educational tool. This tells us that the system can be perceived as more than an accompanist and thus power complex ways of musical co-creativity and interaction.

7. CONCLUSION

Designing an adequate experimental procedure was one of the main challenges, given the limited resources on interactive performance evaluation. What should be tested and what questions should be asked? For example, participants found it difficult to assess the quality of attributes like "responsiveness" and their feedback shows great differences even within the same setting. Formulating evaluation criteria that are useful for the researcher and clear to the participant has proved to be a complex task. The participants' overall experience varies with their skills and backgrounds: this might justify the selection of a more homoge-

[3]As before, referring to how sliding is implemented.

[4]The tune that was used was compared to a notated one the participant brought and there are minor differences in some motivic elements.

nous group in the future if we are not able to enlarge our sample size; however, it is difficult to find participants who have the required experience in Irish traditional dance music and who are willing to engage in the project and will not be biased toward or against the use of AI in music performance. Finally, designing the experiment so that we can perform certain statistical tests is a difficult task when the number of total participants is unknown *a priori*.

Considering the lack of orthogonality, for a future study we consider arranging trial-system pairs for each participant based on Latin squares. A limitation of this approach is that both the number of participants and the number of levels of trial have to be multiples of the number of treatments of factor system. Subsequent experiments will be informed by this strategy and will allow a more advanced statistical analysis.

8. ETHICS STATEMENT

The details of this experiment, including its consent form and information sheet, were reviewed by a dedicated ethics committee at the Royal Institute of Technology (KTH) and were deemed to not need any special considerations: research-ethical risks in the study are judged minimal, and the personal data collected does not raise ethical concerns. Participants were found through the networks of the investigators. Study participation was volunteered and was not remunerated.

9. ACKNOWLEDGMENTS

Portions of this work are outcomes of projects that have received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (MUSAiC, Grant agreement No. 864189). We gratefully acknowledge the support of the UKRI through the "UKRI Trustworthy Autonomous Systems Hub" programme (grant EP/V00784X/1).

10. REFERENCES

[1] M. Amerotti, S. Benford, B. L. T. Sturm, and C. Vear. A Live Performance Rule System informed by Irish Traditional Dance Music. In *Proc. of the 16th International Symposium on Computer Music Multidisciplinary Research*, page 277–288, Nov. 2023.

[2] T. Baba, M. Hashida, and H. Katayose. "VirtualPhilharmony"': A Conducting System with Heuristics of Conducting an Orchestra. In *Proc. International Conference on New Interfaces for Musical Expression*, pages 263–270, Feb. 2018.

[3] R. A. Bailey. *Design of comparative experiments*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2008.

[4] R. Bresin and A. Friberg. Evaluation of computer systems for expressive music performance. In *Guide to Computing for Expressive Music Performance*, pages 181–203. 2013.

[5] R. Bresin, A. Friberg, and J. Sundberg. Director musices: The kth performance rules system. In *Proceedings of SIGMUS-46*, pages 43–48, 2002.

[6] S. Canazza, G. De Poli, and A. Rodà. CaRo 2.0: An Interactive System for Expressive Music Rendering. *Advances in Human-Computer Interaction*, 2015:1–13, 2015.

[7] C. E. Cancino-Chacón, M. Grachten, W. Goebl, and G. Widmer. Computational Models of Expressive Music Performance: A Comprehensive and Critical Review. *Frontiers in Digital Humanities*, 5, 2018.

[8] M. Fabiani, A. Friberg, and R. Bresin. Systems for Interactive Control of Computer Generated Music Performance. In A. Kirke and E. R. Miranda, editors, *Guide to Computing for Expressive Music Performance*, pages 49–73. Springer London, London, 2013.

[9] A. Friberg. Home conducting: Control the overall musical expression with gestures. In *Proc. of the International Computer Music Conference*, pages 479–482, 2005.

[10] A. Friberg. pDM: An Expressive Sequencer with Real-Time Control of the KTH Music-Performance Rules. *Comput. Music J.*, 30:37–48, Mar. 2006.

[11] J. Gillick, A. Roberts, J. Engel, D. Eck, and D. Bamman. Learning to Groove with Inverse Sequence Transformations. In *Proc. International Conference on Machine Learning*, July 2019.

[12] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *Proc. International Conference on Learning Representations*, 2019.

[13] D. Jeong, T. Kwon, Y. Kim, K. Lee, and J. Nam. Virtuosonet: A hierarchical rnn-based system for modeling expressive piano performance. In *Proc. International Society for Music Information Retrieval Conference*, 2019.

[14] D. Jeong, T. Kwon, Y. Kim, and J. Nam. Graph Neural Network for Music Score Data and Modeling Expressive Piano Performance. In *Proc. International Conference on Machine Learning*, pages 3060–3070, May 2019.

[15] H. Katayose, M. Hashida, G. De Poli, and K. Hirata. On Evaluating Systems for Generating Expressive Music Performance: the Rencon Experience. *Journal of New Music Research*, 41(4):299–310, Dec. 2012.

[16] A. Maezawa, K. Yamamoto, and T. Fujishima. Rendering music performance with interpretation variations using conditional variational rnn. In *Proc. International Society for Music Information Retrieval Conference*, 2019.