# MUSIC OF 18 PERFORMANCES: EVALUATING APPS AND AGENTS WITH FREE IMPROVISATION

*Charles Martin*
Research School of Computer Science
Australian National University
charles.martin@anu.edu.au

*Henry Gardner*
Research School of Computer Science
Australian National University
henry.gardner@anu.edu.au

*Ben Swift*
Research School of Computer Science
Australian National University
ben.swift@anu.edu.au

*Michael Martin*
Research School of Finance, Actuarial Studies & Applied Statistics
Australian National University
michael.martin@anu.edu.au

## ABSTRACT

We present a study where a small group of experienced iPad musicians evaluated a system of three musical touch-screen apps and two server-based agents over 18 controlled improvisations. The performers' perspectives were recorded through surveys, interviews, and interaction data. Our agent classifies the touch gestures of the performers and identifies new sections in the improvisations while a control agent returns similar messages sourced from a statistical model. The three touch-screen apps respond according to design paradigms of reward, support, and disruption. In this study of an ongoing musical practice, significant effects were observed due to the apps' interfaces and how they respond to agent interactions. The "reward" app received the highest ratings. The results were used to iterate the app designs for later performances.

## 1. INTRODUCTION

This paper describes the evaluation of a system of touch-screen musical instrument apps and server-based computational agents in a controlled study of 18 free-improvised performances. Free-improvised ensemble music is performed without any plan for the performance and our system is designed to react to the performance structure that emerges while a group is playing. Improvisations can be considered to be segmented by new musical ideas (Stenström 2009, pp. 58–59) and our ensemble-tracking agent searches for these new ideas by classifying and analysing performers' touch gestures. Three musical apps have been developed for the Apple iPad platform that receive messages from this agent and react by updating their interfaces in real-time. Each of the three apps encodes a different behavioural model of interaction in ensemble improvisation. A "reward" model gives performers access to new notes at each new section of the performance, a "disruption" model interrupts performers who stay on one gesture for too long, and a "support" model plays complementary sounds when performers focus on individual gestures.

A group of three touch-screen musicians with more than a year of performance experience with the apps were participants in the study. While concert experience had suggested that the ensemble-tracking agent interacted with the group accurately and could enhance improvisation, a formal experiment was conducted to evaluate the agent system under controlled conditions and compare the three apps. To assess the accuracy of the ensemble-tracking agent, a control agent was developed that generates similar messages randomly from a statistical model. In a methodology that combined a balanced experimental design with rehearsal processes, the group performed a series of 18 improvisations on all combinations of the three iPad interfaces and the two agents. We performed quantitative analyses of survey ratings from the musicians, and on the number of agent messages sent during performances, as well as qualitative analysis compiled from interviews.

The results support the effectiveness of our ensemble-tracking agent, although the source of agent interventions was seen as less important than how the apps responded. The app condition was found to have a significant main effect on the performer's responses to several questions, including the quality and level of creativity in performances. The app featuring the "reward" model showed the most positive response with the performers actively seeking out interaction with the agent when using this app. The performers articulated problems with the other two apps while still finding ways to use them in interesting improvisations and their responses were used to redesign the apps for later performances. Following a review of prior work in this field, in Section 2 we will describe the construction of our

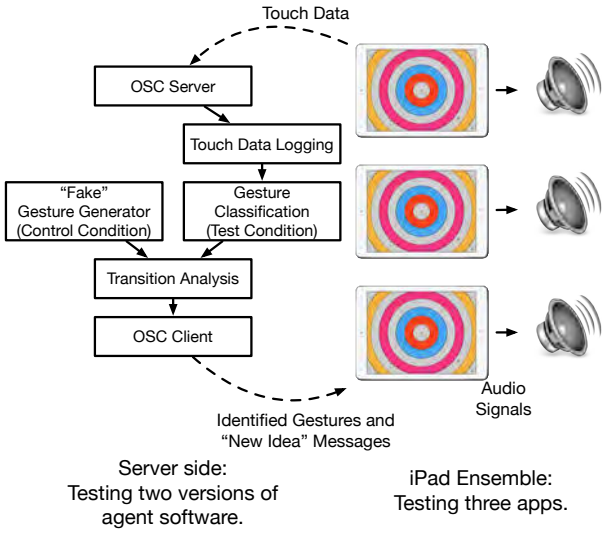


**Figure 1.** A system diagram of our agent software interacting with our iPad instruments. In the test condition, touch messages are classified as gestures by a Random Forest classifier, while in the control, gestures are generated from a statistical model disconnected from the performers' current actions.

system of apps and agents, Section 3 will describe our experimental design, and our results will be analysed and discussed in Section 4.

### 1.1 Related Work

The "Laptop Orchestra" (Bukvic et al. 2010; Trueman 2007) (LO), where multiple performers use identical hardware and software in musical performance, is now established in computer music practice and has an expanding compositional repertoire (Smallwood et al. 2008) and pedagogy (Wang et al. 2008b). These ensembles often use artificial intelligence agents as ensemble members (Martin et al. 2011) or as a "conductor" (Trueman 2007) to provide cohesive direction of broad musical intentions. In our study, two fundamental designs of these mediating agents are evaluated: one using a statistical Markov model (Ames 1989), and one using machine learning algorithms to follow the performers (Fiebrink et al. 2009).

With the emergence of powerful mobile devices such as smartphones and tablets, ensembles of "mobile music" (Gaye et al. 2006; Jenkins 2012) performers have appeared, taking advantage of the many sensors, touch screen interfaces, and convenient form-factors of these devices (Tanaka 2010). These ensembles have used phones to perform gamelan-like sounds (Schiemer and Havryliv 2007), sensor-based music (Wang et al. 2008a) or explore touch interfaces (Oh et al. 2010). Both smartphones (Swift 2013) and tablets (Martin et al. 2014) have been used in improvising ensembles and Williams (2014) has noted their utility in exploratory, collaborative music making. While mobile instruments have often been aimed towards beginners (Wang 2014; Wang et al. 2011), we have been developing a long term musical practice by experienced performers.

There are a wide range of approaches for evaluating

| # | Code | Description | Group |
|---|---|---|---|
| 0 | N | Nothing | 0 |
| 1 | FT | Fast Tapping | 1 |
| 2 | ST | Slow Tapping | 1 |
| 3 | FS | Fast Swiping | 2 |
| 4 | FSA | Accelerating Fast Swiping | 2 |
| 5 | VSS | Very Slow Swirling | 3 |
| 6 | BS | Big Swirling | 3 |
| 7 | SS | Small Swirling | 3 |
| 8 | C | Combination of Swirls and Taps | 4 |

**Table 1.** Touch-screen gestures that our classifier is trained to identify during performances. When gestures are summarised in transition matrices, the gesture groups are used instead, producing $5 \times 5$ matrices.

new digital musical instruments, but it is generally accepted that the performer is the most important stakeholder (O'Modhrain 2011), particularly when performing improvised music. Gurevich et al. (2012) have used a grounded-theory approach to identify styles and skills that emerge when multiple participants engage with very simple electronic instruments. Fiebrink et al. (2011) asked users to repeatedly evaluate interactive musical systems that use machine learning across a number of novel criteria, and this "direct evaluation" was found to have more utility than a typical cross-validation approach for machine learning systems. A long-term ethnographic study of the Reactable table-top surface observed collaborative and constructive processes (Xambó et al. 2013) in video footage of improvised performances. Ethnographic techniques have also been used over natural rehearsal and development processes such as for Unander-Scharin et al.'s (2014) "Vocal Chorder", where an autobiographical design process transitioned into an interface developed for other performers. Our study uses a reheasal-as-research methodology where multiple performances are evaluated in a single session through short surveys and interviews. As our iPad ensemble had already established a performance practice (Martin 2014) they were able to test the six experimental conditions with 18 improvisations in one session, an unprecedented number in musical interface evaluation.

## 2. SYSTEM DESIGN

The following sections detail the construction of our iPad apps, ensemble-tracking agent, and control agents. An overview of the system architecture is given in Figure 1 which shows the two important parts of our agent software: a gesture classification system which uses machine learning algorithms to identify each performer's actions, and a performance tracking system which analyses the ensemble's transitions between touch screen gestures to identify important "new-idea" moments in these performances. We also describe a "fake" gesture generator, used as an experimental control, where gestures were generated randomly from a statistical model derived from a live performance. The gesture generator takes the place of the classifier while other parts of the system remain the same.

Our iOS iPad apps are developed in Objective-C, with the `libpd` library used for audio synthesis. Our server

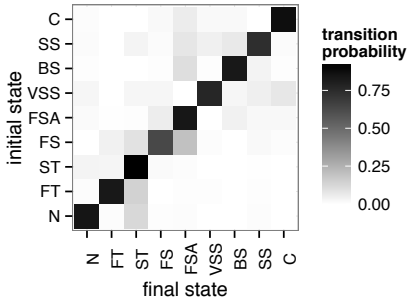


**Figure 2.** Plot of the transition matrix used for the Markov model in the generative agent.

software is developed in Python and communicates with the iPad apps over a WiFi network connection using the OSC message format (Freed and Schmeder 2009). The iPad apps send messages to the server for each touch-screen event and the server logs all of this touch information for gesture classification and for later analysis.

### 2.1 Gesture Classifier

Previous work has identified a vocabulary of gestures used by expert percussionists on touch-screen interfaces (Martin et al. 2014). Our agent is able to identify nine of these touch-screen gestures (see Table 1) using each performer's touch-data at a rate of once per second. The server records these gestures and also sends them back to the performers' iPads.

Classification is accomplished using a Random Forest classifier algorithm (Breiman 2001) provided by Python's `scikit-learn` package. This was trained using examples of the touch-screen gestures recorded in a studio session by our app designer. The input to the classifier is a feature vector of descriptive statistics from the last five seconds of each performer's touch data. The timing parameters for our classifier were tuned by trial and error and previous research (Martin et al. 2015) has shown that our classifier has a mean accuracy of 0.942 with standard deviation 0.032 under cross-validation, comparable to other systems that recognise command gestures (Wobbrock et al. 2007).

### 2.2 Generating Fake Gestures

In order to evaluate the effect of our gesture classifying agent (CLA) on performances in our experiment, we developed a contrasting system that generates fake gestures (GEN) to be used as a control. As the rest of our agent software remains the same (see Figure 1), the fake gestures and fake "new-idea" messages would be recorded and reported back to the iPads in the same way as with the ensemble-tracking agent.

To build this control agent, a live touch-screen performance of the iPad ensemble was analysed with our classification system and the resulting sequence of states was used to construct a first-order Markov model. The concept of using a Markov model to generate data is a common design pattern in computer music and Ames (1989) has described how it can be used to algorithmically compose melodies or vary other musical parameters. In our case, the model was used to generate fake gesture classifications similar to the gestural output of our touch-screen ensemble. As it is statistically similar to the changes induced by the classifying agent, but decoupled from the performers' actual gestures, our generative agent was used as a control in our experiment to expose the effect of an intelligent mediation of live performance.

### 2.3 Transitions, Flux, and New Ideas

Our classifying agent is designed to identify the musical sections present in improvised musical performances and pass this information to the iPad interfaces operated by the performers. A "new-idea" in our system is defined as a moment in the performance where, for the whole ensemble, transitions between different gestures increases sharply over 30 seconds. The implementation of the system is more fully explained in previous research (Martin et al. 2015), but is presented here in brief.

An improvised musical performance can be modelled as a sequence of abstract musical gestures for each performer in the ensemble. In the present study, these gestures are either identified by the gesture classifier (CLA) or generated by our statistical model (GEN). Transitions between different gestures over a certain window of time can be summarised by a transition matrix $P$ constructed in a similar way to the transition matrix of a Markov chain (Swift et al. 2014).

The matrices for each performer can be averaged to summarise the whole ensemble's transition activity. Our agent software compares transition matrices by applying a matrix measure, flux, which is defined as follows:

$$\mathrm{flux}(P) = \frac{\|P\|_1 - \|\mathrm{diag}(P)\|}{\|P\|_1} \qquad (1)$$

where $\|P\|_1 = \sum_{i,j} |p_{ij}|$ is the element-wise 1-norm of the matrix $P$ and $\mathrm{diag}(P)$ is the vector of the main diagonal entries of $P$.

The flux measure is equal to 0 when all the non-zero elements of the transition matrix $P$ are on the main diagonal, that is, when performers never change gesture. The measure will be equal to 1 when no performer stays on the same gesture for two subsequent classifications.

In our agent software, the flux of the ensemble is calculated each second for the two preceding 15 second windows of gestures reduced to their "groups" (see Table 1). If the flux of the ensemble has increased over these windows by a certain threshold, the system sends a new-idea message to the performers's iPads. The iPad apps include a rate-limiting function that prevents them reacting to several measurements of the same new-idea event by ignoring messages for at least 10 seconds after responding to a new-idea. The timing parameters and threshold for detecting new-ideas were tuned by trial and error within the research group. As well as reporting new-idea events, our agent sends the entire gesture classification sequence to the apps, which are able to respond to long sequences of identical gestures (old ideas) as well as the new-idea messages.

### 2.4 iPad Apps

Three different iPad apps were chosen from a repertoire

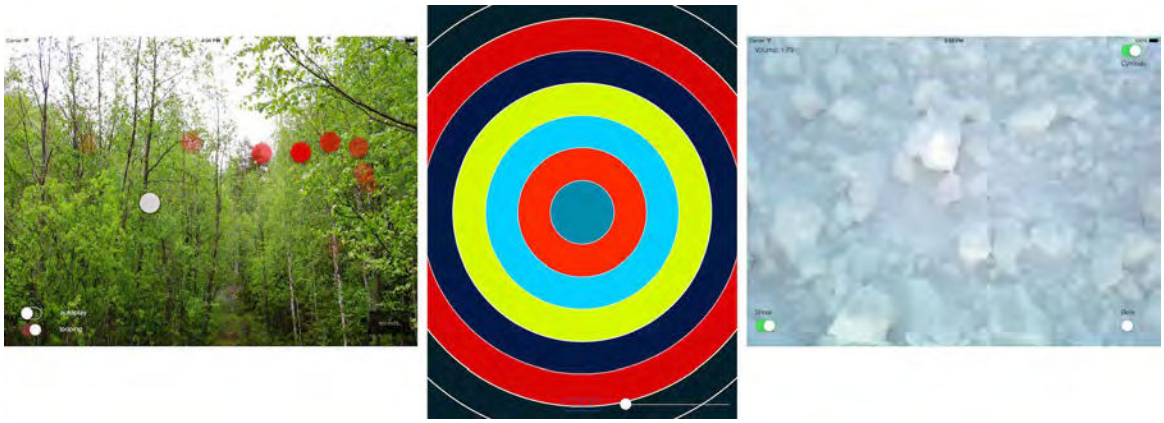

**Figure 3.** The three apps used in this study, from left to right: Bird's Nest (BN), Singing Bowls (SB), and Snow Music (SM).

of six apps created by our designer and routinely used by our ensemble. The three apps, Bird's Nest (BN), Singing Bowls (SB), and Snow Music (SM), are shown in Figure 3. Each app features a free-form touch area where tapping produces short sounds and swiping or swirling produces continuous sounds with volume controlled by the velocity of the moving touch point. While these apps share a paradigm for mapping touch to sound, their different sound material and contrasting designs for interaction with our agents make them three distinct instruments.

Bird's Nest and Snow Music present nature-inspired interfaces to the performers. In Bird's Nest, performers create a soundscape from a northern Swedish forest with bird samples, field recordings, percussive sounds, and a backdrop of images collected at that location. Snow Music emulates a bowl of amplified snow, where performers manipulate field recordings of snow being squished, smeared, stomped and smashed. Singing Bowls presents users with an annular interface for performing with percussive samples. Rings on the screen indicate where touches will activate different pitches of a single-pitched percussion sample. Tapping a ring will activate a single note while swirling on a ring will create a sustained sound reminiscent of that of Tibetan singing bowls.

The apps' response to messages from the agent followed three distinct paradigms. Bird's Nest was designed to **disrupt** the musicians' performance. Based on gesture feedback from the agent, the app would watch for runs of identical gestures and then switch on looping and autoplay features in the user interface in order to prompt new actions by the performers. New-idea messages were used to randomise the sounds available to the user from a palette of sample and pitched material.

Snow Music used a **supportive** paradigm. The app would watch for sequences of similar gestures and activate extra layers of complementary sounds. For instance, the app would support a run of tapped snow sounds by layering the taps with glockenspiel notes while a backdrop of generative bell melodies would be layered on top of the sound of continuous swirling gestures. When the performer moves on to other gestures, the supportive sounds were switched off. New-idea messages in Snow Music changed the pitches of the supportive sounds and the snow samples available to the performer. While the actions of the supportive sounds were shown on the screen, the performers were not able to control them directly.

| Set | Perf. 1 | Perf. 2 | Perf. 3 |
|---|---|---|---|
| 0 | orientation | | |
| 1 | SM, CLA | BN, GEN | SB, CLA |
| 2 | BN, CLA | SB, GEN | SM, GEN |
| 3 | SB, CLA | SM, CLA | BN, GEN |
| 4 | SB, GEN | BN, CLA | SM, GEN |
| 5 | BN, GEN | SM, CLA | SB, CLA |
| 6 | SM, GEN | SB, GEN | BN, CLA |
| 7 | interview | | |

**Table 2.** The experiment schedule showing the balanced ordering of apps and agents. The experiment was performed in one session divided by breaks into six groups of three five minute performances.

Finally, the Singing Bowls app **rewarded** the player's exploration of gestures with new pitches and harmonic material. This app only allows the performer to play a limited number of pitches at a time. When the ensemble's performance generates a new-idea message, the app rewards the players by changing the number and pitches of rings on the screen. The pitches are taken from a sequence of scales so that as the performers explore different gestures together, they experience a sense of harmonic progression.

## 3. EXPERIMENT

Our experiment took the form of a lab-based study under controlled conditions. Although analogous to a rehearsal call for professional musicians in its length and artistic intent—a performance of this ensemble actually took place some four weeks later at an art exhibition—the research intent of this experiment meant that it was quite an unusual rehearsal from the musicians' perspective.

In the experiment, two agents (a classifying agent: CLA, and a generative agent: GEN) were crossed with three iPad apps (Bird's Nest: BN, Singing Bowls: SB, and Snow Music: SM) to obtain the six independent conditions. The

**Figure 4.** The ensemble setup for the lab study shown from one of two camera angles. Each performer's sound was dispersed through a large loudspeaker directly behind them and simultaneously recorded.

ensemble were asked to perform improvisations limited to five minutes each and to immediately fill out questionnaires after each improvisation. It was determined that 18 of these sessions would fit into a standard three-hour rehearsal session which allowed for three trials of each of the six independent conditions.

The entire rehearsal was divided into six sets of three performances (see Table 2) preceded by an orientation and followed by an open-ended interview. In each set, the musicians used each app once and the order of apps was permuted between sets in a balanced design following Williams (1949) to offset local learning effects. Successive performances with each app alternated between the two agents. The experiment was blinded insofar as the performers were aware that two agents were under investigation but were not made aware of the difference between them or of which agent was used in each performance.

The experiment took place in an acoustically treated recording studio (see Figure 4). The performers were seated in the recording room while the two experimenters were present in a separate control room. The experiment was video recorded with two angles[1] which allowed the performers' faces and screens to be seen. The sound of each iPad was recorded from the headphone output in multitrack recording software[2] and simultaneously diffused through large monitor speakers behind the performers. Audio from a microphone directly in front of the ensemble as well as from a microphone in front of the experimenter was also recorded to capture discussion during the experiment and during the post-session interview. In each performance session all touch-interaction messages from the three performers' iPads were recorded (even though only the CLA agent made use of this information), as were the messages returned to the performers by the agents.

### 3.1 Participants

The participants in this study (Performer A, Performer B, and Performer C) are members of an ensemble established to perform improvised music with the apps and agents under investigation as well as acoustic percussion instruments. All three participants are professional percussionists and had worked together previously in educational and professional contexts. The fourth member of this ensemble (Experimenter A) was also the designer of the apps and agents but did not participate in the performances in this study. A second researcher (Experimenter B) assisted with running the study. The two experimenters are also experienced musicians.

Over the 14 months prior to this study, the performers had engaged in a longitudinal process of rehearsals and performances parallel to the development of the apps and agent. The process and other results of this longitudinal study have previously been reported (Martin 2014; Martin and Gardner 2015; Martin et al. 2014).

The three performers were chosen to participate in the present study due to their high level of skill and experience in iPad performance and their capacity for self-evaluation. Cahn (2005, pp. 37–38) has written about the strong learning effect present in new improvisation ensembles, where members overcome initial inhibitions to test the limits of newfound musical freedom with "severe departures from normal music making". This phase is followed by a plateau of thoughtful free-improvisation where "listening and playing come into more of a balance". The significant experience by the performers in this study meant that all of the performances recorded had the potential to be of high quality.

### 3.2 Questionnaires

At the end of each performance, the performers filled out written surveys consisting of the following questions on a five point Likert-style scale (Very Bad, Bad, Neutral, Good, Excellent). The two experimenters present during the lab study were also surveyed on Question 1.

1. How would you rate that performance?
2. How would you rate the level of creativity in that performance?
3. How did the agent's impact compare to having it switched off?
4. How well were you able to respond to the app's actions?
5. How well were you able to respond to the other players' actions?
6. How was the app's influence on your own playing?
7. How was the app's influence on the group performance?

The format and content of our questionnaire follows other evaluations of improvised performance, including Eisenberg and Thompson (2003), in evaluating overall quality, creativity, and ensemble interaction, however we added specific questions to evaluate the overall impact of the agents and the changes that they caused in the apps.

## 4. RESULTS

In the following sections we analyse and discuss the data collected in the study session. This corpus consists of 57 minutes of interviews, 92 minutes of performances,

[1] The video recorders were a GoPro HERO 3+ Black and a Zoom Q2HD both set to record in 1920*1080 resolution.

[2] The audio recording was made through a Presonus Firepod interface in Apple Logic Studio.

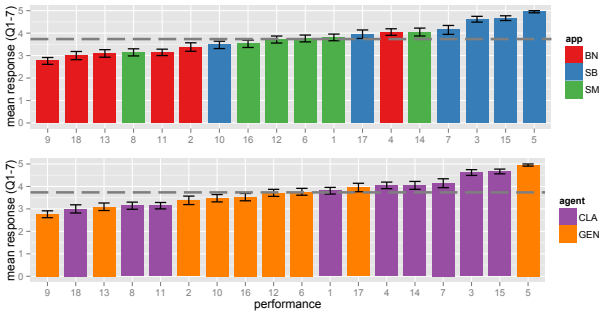


**Figure 5.** Performances ordered by the mean response to all questions. The grand mean is shown as a dashed horizontal line. The distribution of apps is striking with SB eliciting the highest ratings. Two thirds of the CLA agent performances appear above the mean.

32.2MB of touch and interaction data as well as the experimenters' notes. We will first discuss the data from surveys and agent-app interaction before considering the interview responses.

### 4.1 Survey Data

The survey responses from each question were analysed separately using univariate two-way repeated measures analysis of variance (ANOVA) procedures to determine the significance of main and interaction effects of the two indepedent variables (app and agent). Post-hoc Bonferroni-corrected paired $t$-tests were used to assess significant variations between each of the six experimental conditions. This is a standard procedure for significance testing used in human-computer interaction studies (Lazar et al. 2010).

Results from five of the seven questions (1,2,4,6,7) were found to be significant and will be considered below in detail. The other questions (3,5) were not significantly affected by the change of apps and agents. The normal variations of musical interactions in between five minute performances may have affected these questions more than the independent variables.

#### *4.1.1 Mean Response*

Figure 5 shows the mean response to all questions, yielding a holistic overview of the results. For the apps, this figure shows that, in general, Singing Bowls was rated higher than Snow Music which was higher than Bird's Nest. For the agents, performances with the classifying agent were, in general, more highly rated than those with the generative agent, with six of the nine classifier performances appearing above the grand mean.

#### *4.1.2 Performance Quality and Creativity*

Questions 1 and 2 of the survey concerned the overall level of quality and creativity in each performance, the distribution of responses to these questions are shown as a box plot (McGill et al. 1978) in Figure 6. Considering all responses to Question 1 in the survey (including the two experimenters), the app used had a significant effect on the perception of quality in the performances, $F(2, 8) = 5.006, p < 0.05$. The main effect of the agent and the interaction effect were not found to be significant.

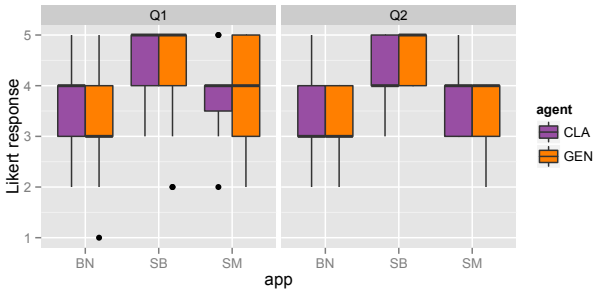


**Figure 6.** Distribution of performance quality (Question 1) and creativity (Question 2) ratings by app and agent. For both questions, the app had a significant effect on ratings.

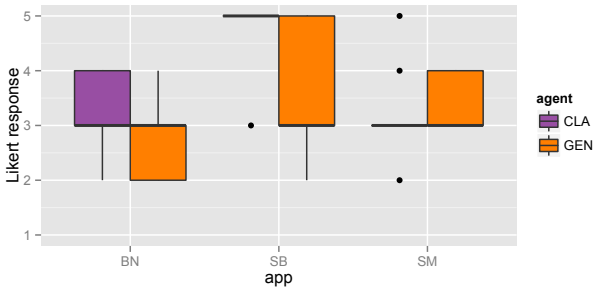


**Figure 7.** Distribution of ratings of individual performers' responses to the agents' actions (Question 4). The main effect of the app and an interaction effect between app and agent were found to be significant.

Bonferroni-corrected paired $t$-tests revealed that, without considering the agent, performances with the Singing Bowls app were of significantly higher quality than with Snow Music ($p = 0.04$) and Bird's Nest ($p = 0.002$).

A significant main effect of app was also observed on the performers' ratings of the level of creativity in their performances, $F(2, 4) = 8.699, p < 0.05$. Bonferroni-corrected paired $t$-tests only showed that performances of Singing Bowls with the generative agent were rated as significantly more creative ($p < 0.05$) than Snow Music with the generative agent and Bird's Nest with either agent.

#### *4.1.3 Responding to the App's Actions*

The performers were surveyed on how well they were able to respond to changes in the app interfaces caused by the agent (Question 4). Although the agent and app worked together as a system to change the interface, on the survey we called this "the app's actions" as from the performers' perspective, they were only aware of changes in the app.

A box plot of the results are shown in Figure 7. There was a significant effect of the app ($F(2, 4) = 13.32, p < 0.05$) and a significant interaction effect between agent and app ($F(2, 4) = 7.75, p < 0.05$). The effect of the agent was of borderline significance ($F(1, 2) = 16, p = 0.0572$). Post-hoc Bonferroni-corrected pairwise $t$-tests revealed that the performers were able to respond to the Singing Bowls app in combination with the classifying agent significantly better than for the other two apps with either agent ($p < 0.05$) but with only borderline significance against the Singing Bowls app with the generative

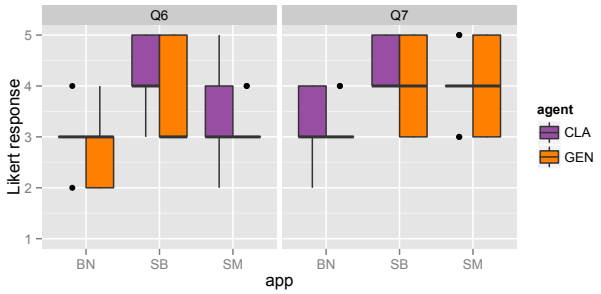


**Figure 8.** Distribution of responses about the influence of the app and agent on the performers' own playing (Q6), and the group's playing (Q7). In Q6, the app had a significant effect, while in Q7, the agent appears to have been more important.

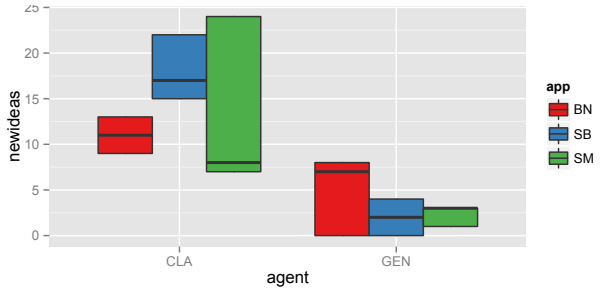


**Figure 9.** Distribution of the number of new-idea messages sent during performances. The GEN agent failed to match the CLA agent's behaviour, but the app also had an effect.

agent ($p = 0.11$). These tests revealed that when using the Bird's Nest and Snow Music apps and the classifying agent, performers reported that they were better able to respond to the app's actions than with the generative agent, although significance was borderline ($p < 0.1$).

#### *4.1.4 App/Agent Influence*

Questions 6 and 7 both relate to the influence of the app and agent on the performance with the former asking about impact on the individual and the latter on the group. By univariate ANOVA, the effect of the app on the performers' own playing was found to be significant ($F(2, 4) = 137.2$, $p < 0.01$)). The effect of the agent on the group performance was of borderline significance ($F(1, 2) = 16$, $p = 0.0572$).

A multivariate ANOVA on both outcomes showed significance only for the app's effect ($F(2, 4) = 4.238, p < 0.05$). These results suggest that although the app interface was the most important factor in the participants' perceptions of their own playing, the agent was a more important factor when considering the group.

#### *4.1.5 New Ideas*

As discussed in Section 2, the generative agent produced randomised gesture classifications based on a statistical distribution derived from a previous performance of the iPad ensemble. We had hoped that this agent would act as a control in our experiment by producing a similar number of new-idea messages but at times which did not correlate

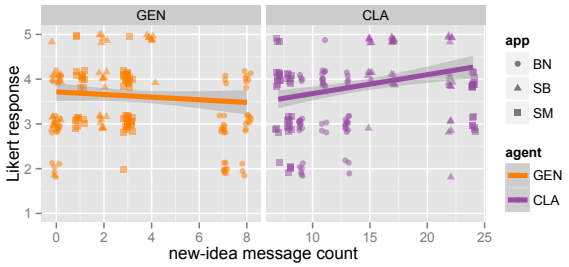


**Figure 10.** Number of new-idea messages in each performance plotted against performers' responses to all questions. The lines show a linear model with standard error for messages generated by the GEN and CLA agents. Responses are lower for more GEN new-ideas, but higher for more CLA new-ideas.

with activity in the live performance. However, from Figure 9, it is clear that the classifying agent produced more new-idea messages than the generative agent. We can investigate this difference by treating the number of new-idea messages as a dependent variable.

A two-way ANOVA showed that only the effect of the agent on the number of the new-idea messages was significant ($F(1, 12) = 24.19, p < 0.001$). Although the app's effect on new-ideas was not found to be significant, the number of new-ideas generated with Singing Bowls and the classifying agent appears higher than with Snow Music or Bird's Nest (Figure 9). This suggests that the musicians may have performed more creatively with Singing Bowls, cycling through numerous gestures as an ensemble.

Figure 10 shows the performers' responses to all questions against the number of new-ideas in each performance. A linear model of responses for each agent suggests that ratings decline as the generative agent produced more new-idea messages, while ratings increase as the classifying agent produced more messages. This may suggest that for the generative agent, more changes due to new-idea messages annoy the performers as they do not necessarily correspond to their actions. For the classifying agent, large numbers of new-idea messages may have been triggered by particularly creative and engaged performances which elicited higher ratings. While the generative agent did not produce the same numbers of new-idea messages as the classifying agent, if it had, the performers' responses may have been more negative.

### 4.2 Qualitative Analysis

Video recordings were made of the orientation briefing, the 18 performances, and the post-experiment interview. Thematic analysis (Braun and Clarke 2006) of a transcription of the interview revealed that the performers' experiences were shaped by the three apps and their interaction with the agents. This qualitative analysis was used to direct a redesign of two of the apps leading up to a concert performance of the ensemble four weeks after the experiment.

From the interview data, and confirming the analysis of the Likert data for Question 1 and 2, the performers

were most satisfied with the Singing Bowls app. It was noted that Singing Bowls was "familiar as an instrument ... responds as we're used to" (Perf. B) and that the "time went quickly in performances" (Perf. B). The ensemble noticed some performances (with the generative agent) where the Singing Bowls app "didn't change at all" (Perf. C). Rather than being discouraged, the performers tried actively to "get it to respond by copying and mimicking and getting everybody to change" (Perf. A). Because of this positive reception, Singing Bowls was deemed a success and its design was not revisited before the concert performance.

In marked contrast to Singing Bowls, performances with Snow Music felt like they "went on forever" (Perf. A), suffering from a lack of structure and motivation to keep playing with the smaller palette of snow sounds. The performers suggested that the "use of space" (i.e. silence) in Snow Music performances could improve quality and allow focus on particular sounds. The interaction with the supporting sounds was described as "lovely" (Perf. A) and "would play some really good stuff" (Perf. C). In response to these comments, design revisions were made to add to the palette of snow sounds and to refine the synthesis system to be more expressive. A sequence of harmonies was added to the pitched sounds in this app to encourage the group to continue exploring throughout the performance.

Bird's Nest performances suffered the lowest ratings from the performers who felt annoyed and "isolated" (Perf. A) by the disruptive interaction between the app and agent and found it "really hard" (Perf. C) to use creatively. While the app's sounds were "pretty" (Perf. A) it was "hard to have that flow of ideas between people" (Perf. C). It was noted that Bird's Nest was "less similar than an instrument" (Perf. B) than the other apps and that the sounds were "long... and the change in pitch is... less perceptible" (Perf. C). Following these comments, Bird's Nest was extensively revised for later concerts. The "autoplay" feature and the disruptive control of the "looping" function were removed. A sequence of images with corresponding scales and subsets of the sound palette was devised to form a compositional structure for performances. The "sounds" button was retained to refresh the palette of sounds for each scene, but, as with Singing Bowls, movement through the compositional structure depended on new-idea messages. The synthesis system for playing bird samples was refined to play sounds of varying, but usually much shorter, length.

The qualitative analysis suggested that, from the performers' point of view, the source of the agent's interventions (either responding to their gestures or generated from a model) was not as important as the way that the apps *responded* to these interventions. The "rewarding" paradigm used in Singing Bowls was the most successful in engaging the performers' interest. It was notable that with Singing Bowls the performers sought out agent interactions, particularly when the agent did not respond as was the case with the generative agent.

## 5. DISCUSSION

The primary limitation of the present study is the small number of participants surveyed. We surveyed only three participants with very specialised skills, so the generalisation of their responses is limited. The goal of this study was not to evaluate performances by inexperienced players but by practiced iPad musicians with an important stake in the quality of the instruments they use. We studied an expert iPad ensemble with extensive performance experience and as a result, we were able to examine more experimental conditions and more improvisations than would be feasible with beginners. As far as we are aware, no controlled studio-based study of 18 touch-screen performances by the same ensemble has been previously attempted. Given the strong preference for the Singing Bowls app, future studies with more participants may be warranted that focus only on this app to reduce the number of required trials.

The multitrack audio and video recordings of the 18 improvised performances and corresponding touch gesture data were important outcomes of this study. Other studies have used detailed logs of improvisations as a basis for analyses of keyboard (Gregorio et al. 2015; Pressing 1987) and live-coding (Swift et al. 2014) performances. We propose that performing similar analyses on our recorded performances could lead to further understanding of the structure of touch-screen improvisation and improvements in the ability of our gesture-classifying agent to track such performances.

## 6. CONCLUSION

Our system for ensemble touch-screen musical performance includes two server-based agents and three iPad apps. One agent classifies performers' gestures to track new ideas while the other generates similar messages from a statistical model. The three iPad apps use the responses from these agents to support, disrupt, and reward gestural exploration in collaborative improvised performances.

We have presented the results of an evaluation of this system's use in real musical performances in a formal, order-balanced study that considered surveys, interviews and interaction data by an expert iPad ensemble with 14 months of experience. The participants' high skill level allowed us to introduce a novel experimental design where a total of 18 performances over six conditions were recorded and evaluated in one rehearsal session.

Different apps were found to have a significant main effect on the performers' perception of performance quality and creativity, how well they were able to respond to interface changes, and the app's influence on individual playing. The main effect due to the agent was found only to have borderline significance on the app's influence on the group performance and the perfomers' ability to respond to interface changes. However, this question did reveal a significant interaction effect between the app and agent conditions.

While significant effects due to the agent were somewhat elusive, the study revealed that our generative agent produced significantly fewer new-idea messages than the classifying agent. Modelling the performer responses with respect to the number of new-idea messages suggests that performances with many classified new-ideas were rated highly, but frequent generated new-ideas may have had a negative impact on ratings.

The results of our study lead us to conclude that the design of an agent's interaction with a creative interface can make or break a performer's positive perception of this interaction; this design can also limit or enhance dynamic and adventurous playing. While experienced performers can create high quality, creative performances mediated by agents of many designs, connecting the agent to their actions seems to have a positive effect on how they respond to interface changes and their perception of the group performance. Rewarding users for their collaborative exploration was found to be an especially engaging paradigm for supporting creativity. The idea of disrupting performers' flow to encourage more creative interaction was roundly rejected in both quantitative and qualitative results.

This study has been a snapshot in the participants' ongoing artistic practice, and the recommendations from the performers have already been taken into account in updates to the apps for subsequent performances. Design guidelines for agent-app interaction will be further articulated in future work. While this study has concerned expert performers, given the broad interest in touch-screen computing, future investigations could consider a wider range of performers, and particularly users in musical education.

## 7. REFERENCES


C. Ames. The Markov process as a compositional model: A survey and tutorial. *Leonardo*, 22(2):175–187, 1989.

V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

L. Breiman. Random Forests. *Machine Learning*, 45(1): 5–32, 2001.

I. I. Bukvic, T. Martin, E. Standley, and M. Matthews. Introducing L2Ork: Linux Laptop Orchestra. In *Proc. NIME '10*, pages 170–173, 2010.

W. L. Cahn. *Creative Music Making*. Routledge, 2005.

J. Eisenberg and W. F. Thompson. A matter of taste: Evaluating improvised music. *Creativity Research Journal*, 15(2-3):287–296, 2003.

R. Fiebrink, D. Trueman, and P. R. Cook. A metainstrument for interactive, on-the-fly machine learning. In *Proc. NIME '09*, 2009.

R. Fiebrink, P. R. Cook, and D. Trueman. Human model evaluation in interactive supervised learning. In *Proc. CHI '11*, pages 147–156. ACM Press, 2011.

A. Freed and A. Schmeder. Features and future of Open Sound Control version 1.1 for NIME. In *Proc. NIME '09*, pages 116–120, 2009.

L. Gaye, L. E. Holmquist, F. Behrendt, and A. Tanaka. Mobile music technology: report on an emerging community. In *Proc. NIME '06*, pages 22–25, 2006.

J. Gregorio, D. S. Rosen, M. Caro, and Y. E. Kim. Descriptors for perception of quality in jazz piano improvisation. In *Proc. NIME '15*, 2015.

M. Gurevich, A. Marquez-Borbon, and P. Stapleton. Playing with constraints: Stylistic variation with a simple electronic instrument. *Computer Music Journal*, 36(1): 23–41, 2012.

M. Jenkins. *iPad Music: In the Studio and on Stage*. Taylor & Francis, 2012.

J. Lazar, J. Feng, and H. Hochheiser. *Research Methods in Human-Computer Interaction*. John Wiley & Sons, West Sussex, UK, 2010.

A. Martin, C. T. Jin, and O. Bown. A toolkit for designing interactive musical agents. In *Proc. OzCHI '11*, pages 194–197. ACM Press, 2011.

C. Martin. Making improvised music for iPad and percussion with Ensemble Metatone. In *Proceedings of the Australasian Computer Music Conference*, pages 115–118, Fitzroy, Vic, Australia, 2014. ACMA.

C. Martin and H. Gardner. That syncing feeling: Networked strategies for enabling ensemble creativity in iPad musicians. In *Proc. CreateWorld '15*, Brisbane, Australia, 2015. AUC.

C. Martin, H. Gardner, and B. Swift. Exploring percussive gesture on iPads with Ensemble Metatone. In *Proc. CHI '14*. ACM Press, 2014.

C. Martin, H. Gardner, and B. Swift. Tracking ensemble performance on touch-screens with gesture classification and transition matrices. In *Proc. NIME '15*, 2015.

R. McGill, J. W. Tukey, and W. A. Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.

J. Oh, J. Herrera, N. Bryan, L. Dahl, and G. Wang. Evolving the mobile phone orchestra. In *Proc. NIME '10*, 2010.

S. O'Modhrain. A framework for the evaluation of digital musical instruments. *Computer Music Journal*, 35(1), 2011.

J. Pressing. The micro- and macrostructural design of improvised music. *Music Perception: An Interdisciplinary Journal*, 5(2):pp. 133–172, 1987.

G. Schiemer and M. Havryliv. Pocket Gamelan: Interactive mobile music performance. In *Proc. IS-CHI '07*, pages 716–719. Research Publishing, 2007.

S. Smallwood, D. Trueman, P. R. Cook, and G. Wang. Composing for laptop orchestra. *Computer Music Journal*, 32(1):pp. 9–25, 2008.

H. Stenström. *Free Ensemble Improvisation*. Number 13 in ArtMonitor. Konstnärliga fakultetskansliet, University of Gothenburg, Sweden, 2009.

B. Swift. Chasing a feeling: Experience in computer supported jamming. In S. Holland, K. Wilkie, P. Mulholland, and A. Seago, editors, *Music and Human-Computer Interaction*, Springer Series on Cultural Computing, pages 85–99. Springer London, 2013.

B. Swift, A. Sorensen, M. Martin, and H. J. Gardner. Coding Livecoding. In *Proc. CHI '14*. ACM Press, 2014.

A. Tanaka. Mapping out instruments, affordances, and mobiles. In *Proc. NIME '10*, pages 88–93. University of Technology Sydney, June 2010.

D. Trueman. Why a laptop orchestra? *Organised Sound*, 12(2):171–179, August 2007.

C. Unander-Scharin, A. Unander-Scharin, and K. Höök. The vocal chorder: Empowering opera singers with a large interactive instrument. In *Proc. CHI '14*, pages 1001–1010. ACM Press, 2014.

G. Wang. Ocarina: Designing the iphone's magic flute. *Computer Music Journal*, 38(2):8–21, 2014.

G. Wang, G. Essl, and H. Penttinen. Do mobile phones dream of electric orchestras? In *Proc. ICMC '08*, August 2008a.

G. Wang, D. Trueman, S. Smallwood, and P. R. Cook. The laptop orchestra as classroom. *Comput. Music J.*, 32(1): 26–37, 2008b.

G. Wang, J. Oh, and T. Lieber. Designing for the iPad: Magic Fiddle. In *Proc. New Interfaces for Musical Expression 2011*, pages 197–202. University of Oslo, 2011.

D. A. Williams. Another perspective: The iPad is a real musical instrument. *Music Educators Journal*, 101(1): 93–98, 2014.

E. J. Williams. Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Chemistry*, 2(2):149–168, 1949.

J. O. Wobbrock, A. D. Wilson, and Y. Li. Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes. In *Proc. UIST '07*, pages 159–168. ACM Press, 2007.

A. Xambó, E. Hornecker, P. Marshall, S. Jordà, C. Dobbyn, and R. Laney. Let's jam the Reactable: Peer learning during musical improvisation with a tabletop tangible interface. *ACM Trans. Comput.-Hum. Interact.*, 20: 36:1–36:34, 2013.