MoodifierLive: Interactive and collaborative expressive music performance on mobile devices

Marco Fabiani, Gaël Dubus and Roberto Bresin
KTH Royal Institute of Technology
School of Computer Science and Communication
Dept. of Speech, Music and Hearing
Lindstedtsv. 24
100 44 Stockholm, Sweden
{himork,dubus,roberto}@kth.se

ABSTRACT

This paper presents *MoodifierLive*, a mobile phone application for interactive control of rule-based automatic music performance. Five different interaction modes are available, of which one allows for collaborative performances with up to four participants, and two let the user control the expressive performance using expressive hand gestures. Evaluations indicate that the application is interesting, fun to use, and that the gesture modes, especially the one based on data from free expressive gestures, allow for performances whose emotional content matches that of the gesture that produced them.

Keywords

Expressive performance, gesture, collaborative performance, mobile phone

1. INTRODUCTION

In the last five years, the evolution of mobile devices (in particular, mobile phones) and services has been fast and disruptive. Today, a very large part of the population owns a mobile phone, many of which are smartphones, devices that allow, among other things, to connect to the internet, listen to music, and run small applications and games. Following a trend in PC- and console-based video games, several interactive music mobile applications and games appeared that became instant best-sellers (e.g. Smule's *Ocarina*[1] and more recently the *Reactable mobile*[2]).

MoodifierLive is an application that aims at combining on a mobile-platform two different aspects of music-related research: automatic performance, and expressive gesture analysis. Mobile phones were used for their immediate availability, the ability to create an all-in-one solution (i.e. sound production and control device), and their connectivity options. *MoodifierLive* was developed in the contest of the FP7 EU ICT SAME Project[3] (Sound And Music For Everyone Everyday Everywhere Every-way, 2008-2010).

[1] http://www.smule.com/
[2] http://www.reactable.com/products/mobile/
[3] http://www.sameproject.eu/

NIME'11, 30 May–1 June 2011, Oslo, Norway.

The aim of automatic performance is to allow computers to play a musical score (e.g. MIDI file) in an expressive way, by imitating the techniques used by musicians. This is achieved by introducing deviations from the nominal values of acoustical parameters such as a tone's length and dynamic level, and tempo. The KTH rule system for music performance [11] is a large set of rules which define the value of these deviations analytically. The effects of the rules are cumulative, their relative contribution defined by a weighting factor. By changing the value of the weighting factors, one can modify the performance in real-time.

Music performance have a strong connection with movement and gesture [6]. Gestures produced by musicians do not only have a functional purpose (i.e. to produce a specific sound), but also help to convey her expressive intention. Expressive gestures, extracted from video analysis, have been previously used to control the KTH rule system [10]. In *MoodifierLive*, data from the phone's built-in accelerometer are analyzed and mapped to performance macro-rules to obtain corresponding expressive music performances.

2. MOODIFIERLIVE

MoodifierLive is a mobile phone application designed to work on Nokia S60 series mobile phones. It is written in Python, and requires the PyS60 interpreter[4] to be installed on the phone. It plays MIDI files expressively using the KTH rule system for music performance [10, 11] to control the main acoustical parameters of the musical performance (i.e. tempo, dynamics, and articulation).

The use of the performance rules allows even the non-musicians to obtain musically acceptable interpretations of a score, by offering high-level, more intuitive controllable parameters, which are automatically mapped to the low-level acoustical ones. For instance, a classic performance technique, the phrase arch, in which the first part of a musical phrase is played with *crescendo-accelerando* and the second with *decrescendo-ritardando*, is automatically applied by the *phrase-arch* rule (provided that the phrasing has been previously defined in the score file).

The application offers five different interaction modes (described in detail in Sect. 2.1-2.5), two of which give direct control over some of the numerous performance rules, while the remaining three simplify the control even further by introducing an additional mapping, from emotions or expressive gestures to rules.

2.1 Sliders mode

In the sliders mode the user has the ability to control the values of the four main performance rules: *Overall Tempo*

[4] http://www.pys60.org/

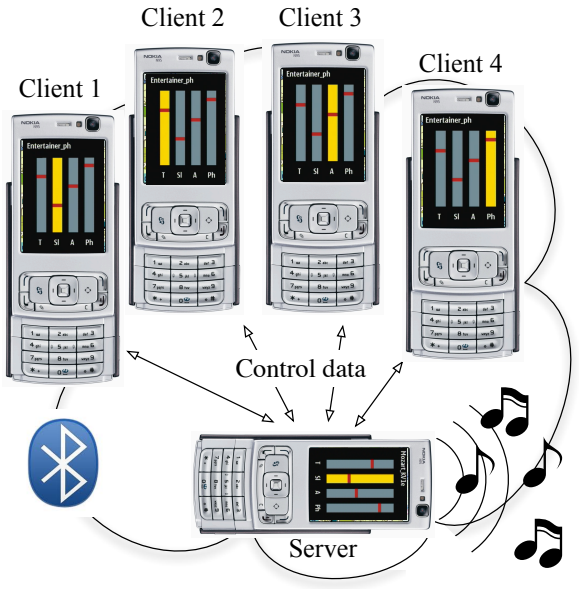

Figure 1: Schematic view of the collaborative interaction mode. Note the displays showing the four sliders controlling Overall Tempo, Overall Sound level, Overall Articulation, and Phrasing.

(T), *Overall Sound level* (Sl), *Overall Articulation* (A), and *Phrasing* (Ph). The sliders mode was designed to allow the naïve user to explore the effect of the single rules on the performance, as well as to allow more advanced users to fine-tune their performances. More rules are available, but to strike a balance between simplicity and usability, we used only the ones that we consider most important to obtain a good performance. Previous research as demonstrated that only tempo, sound level, and articulation account for about 90% of the communication of emotion in expressive music performance [12].

2.2 Collaborative mode

Playing together with other people is an important aspect of the music performance experience, which is both challenging and entertaining. An example of collaborative music performance with mobile phones was proposed in the CaMus^2 project [14]: the camera was used to navigate a marker sheet, and the extracted parameters (e.g. rotation, height) were mapped to MIDI control messages and sent to a PC running MIDI sound synthesis software. Our implementation of a collaborative performance experience allows up to four users to contribute to the creative process by taking control of one or more of the four parameters available in the sliders mode (section 2.1). This simple implementation of a collaborative mode does not aim at reproducing the complex interactions established in an ensemble, but it nevertheless introduces a social component to automatic music performance.

The mode is based on a server-client architecture (see Figure 2.1). In the current implementation, the main application (i.e. *MoodifierLive*) runs on a mobile phone acting as server, which is responsible for the music playback. The client phones (up to four), connected to the server via Bluetooth, are mere control devices. Once a client-server connection is established, the server begins sending regular status-update messages to the clients.

The status-update message is a string of characters, of variable length, beginning with a colon and ending with a semicolon, of the form `:tnn smm aii pjj xxx...x;`. The first twelve characters, divided in four groups of three, correspond to the four parameters (i.e. sliders). The first character in each group (i.e. `t`, `s`, `a`, and `p`) can assume one of three values: `F` if the parameter is free for booking; `B` if the parameter is booked by another client; `S` if the parameter is selected by the client. The second and third character in each group (i.e. `nn`, `mm`, `ii`, `jj`) contain the value of that parameter.The remaining characters (i.e. `xxx...xxx`) contain the name of the current score. For example, the message `:S20B11F59B44The_Entertainer;` means that the tempo parameter is selected by the current client, and has a value of 20; the overall dynamics and phrasing parameters are booked by other clients, and have values 11 and 44, respectively; the articulation parameter is free for booking and has a value of 59; the title of the score is "The Entertainer". The sliders on the client phones are updated with the values in the status messages. Their color indicates if they are booked or free.

The user of a client phone can book and change the value of more than one slider. Any user action generates a command message, which is sent to the server phone. The command message is a string of three characters, opened by a colon, and closed by a semicolon, of the form `:cni;`, where `c` indicates the command: `B` for book; `R` for release; `M` for modify. The second character, `n`, indicates the number of the parameter for which the command was issued (i.e. `0`, `1`, `2`, or `3`). Finally, `i` indicates if the value of the parameter has to be increased (`+`) or decreased (`-`) by a constant step, although only if `c` = `M`. For example, the command `M2+` orders the server to increment the articulation value.

2.3 *Navigate the performance* mode

To simplify the performance control, the use of expressive performances based on macro-rules was introduced [3]. These are sets of low-level rules with a specific set-up that corresponds to a specific emotional expression, for example happiness or sadness. The activity-valence space, previously used in [10], is employed as a simple bi-dimensional model to define different emotions. The activity-valence space boundaries are defined by the sets of values from an experiment [4] in which several expert musicians were asked to create expressive performances of a few musical pieces by setting the values of seven musical variables (tempo, sound level, articulation, phrasing, register, timbre, attack speed). The intermediate values in the bi-dimensional space are obtained by interpolation.

A virtual ball , confined within the screen boundaries, is used to "navigate" in the bi-dimensional space and thus in the space of possible expressive performances. The ball is controlled by tilting the phone, as if it was in a box (the built-in accelerometer is used to determine where the ball is rolling to). The position, size, and color of the ball are used to visually reinforce the perception of the emotion expressed by the music. The size of the ball is directly coupled with the activity, while its color changes following the findings of a previous study [2] that investigated the colors that listeners associate to expressive performances: yellow for happiness, red for anger, blue for sadness, and pink for tenderness.

2.4 *Marbles in a box* mode

Two gesture-based control modes based were designed to allow for a more intuitive interaction with the performance: *Marbles in a box*, and *Free movement* (see section 2.5). Hand gesture information is derived from the acceleration data provided by the phone's built-in accelerometer, and mapped to the activity-valence space (section 2.3).

The *Marbles in a box* is a metaphor to represent the phone as a box containing some marbles, which can be shaken and moved around. The same concept has been used in several other applications (see for example [16, 7]). The energy of the movement, computed as the squared modulus of the acceleration, is directly mapped to the activity, and the tilt of the phone to the valence. Holding the phone high over one's head (positive tilt) should display a positive emotion (feeling "high"), while holding it down parallel to one's leg (negative tilt), should represent a negative emotion (feeling "down"), as previous studies showed [6]. The energy of the movement is also visualized by changing the color of a marble on the phone screen, as in the *Navigate the performance* mode (see section 2.3). Energy and tilt (and thus activity and valence) are sampled at the same frequency as the accelerometer data ($f_s = 30\,\text{Hz}$), and smoothed with a running average over 40 samples (1.3 s).

2.5 *Free movement* mode

The second interaction mode to make use of the accelerometer to detect the user's gestures is called *Free movement*.This mode, unlike the *Marbles in a box*, was developed using data from actual gestures, collected and analyzed in a previous experiment [9]. Eight people were asked to freely perform gestures expressing one of four basic emotions, continuously for 10 seconds. After that, a classification tree was trained using features extracted from these data. The simple regression tree was chosen for its simplicity and because it requires very little computational power, important when it has to be implemented on a mobile device. The choice of a classifier instead of a regressor (i.e. returning continuous values of activity and valence) was dictated by the fact that the training data was categorical.

Two features were chosen to train the classification tree: the velocity and the jerkiness of the gesture, and in particular their Root Mean Square (RMS) values. Although several features from previous studies were considered [1, 11, 13], many were discarded for reasons such as the high correlation between them, and the low sensitivity of the accelerometer ($\pm$2g), which makes the estimation of the relative position of the phone unreliable. The velocity was computed by separately integrating the acceleration in the three directions after subtracting the average over a time window of about 1.3 seconds (to remove the gravity's bias), and then taking the absolute value of the resulting vector. The jerkiness was computed as the derivative of the acceleration [15]. While the velocity gives a rough indication of the energy of the gesture, the jerkiness is an index of how smooth or spiky the movements are.

Before the features extraction, each participant's data was standardized with its mean and standard deviation over all the performances. This followed the observation that the gestures were very similar between participants, but were performed with different intensities.

Cross-validation was used to determine the minimum-cost tree. The resulting tree slightly differs from the one described in [9], because slightly different features and different frame lengths for the averaging were used. The resulting minimum-cost tree is:

```
if Jerkiness (RMS) > 0.86
   ANGRY
else
   if Jerkiness (RMS) > -0.45
      HAPPY
   else
      if Velocity (RMS) > 0.0
         SAD
      else
         TENDER
```

Each one of the four basic emotions is then assigned a fixed value of activity and valence, based on the values obtained in [4], corresponding to the four corners of the activity-valence space in the *Navigate the performance* mode (section 2.3). The features are sampled and smoothed as in the *Marbles in a box* mode. The classification is thus performed 30 times per second.

3. EVALUATION

3.1 Public evaluation

MoodifierLive, developed in the framework of the SAME project [5], was demoed at two public events: the Agora Festival 2009 in Paris (France), and the Festival of Science 2010 in Genova (Italy). Questionnaires were handed out to the visitors in order to collect feedback.

For the version presented in Paris, only three interaction modes (sliders, *Navigate the performance*, and *Marbles in a box*) had been implemented. The response to the questionnaires was positive: the application was judge interesting and fun to use. Critics were directed towards the control: according to the respondents, it could have been made more interesting. This feedback lead to the development of the two other interaction modes, the collaborative and the *Free movement* modes.

The large size of the groups of visitors at the Festival of Science 2010 prevented us from letting people test the application themselves. The groups were only shown a short demonstration. Very few of them completed the questionnaires. For this reason, we decided to evaluate the two modes based on gestures (i.e. *Marbles in a box* and *Free movement*) in a more controlled experiment, carried out at our lab.

3.2 Experimental evaluation

A simple six-tasks experiment, which is described in more detail in [8], was used to evaluate the two gesture-based interaction modes. For each of the two modes, the participants (6M, 7F) were asked to produce, by shaking and moving the phone, three performances that expressed anger, happiness, and sadness. Before each task, they had some time to freely test the application. When ready, they pressed a key to record the performance. After each task, two questions were asked, to be answered on a seven-steps Likert scale (1 = "Not at all", 7 = "Very much"):

1. How successful were you in the task? How much did the performance correspond to the emotion you were supposed to express?
2. How well did your gesture correspond to the emotion you were supposed to express?

At the end of the experiment, the participants were asked to choose their favorite interaction mode. Accelerometer and performance data (i.e. activity and valence) were also logged in a file for later analysis.

Three goals were set for the experiment. First, to verify if the participants, without any explanation about the mapping from gestures to performance, would understand how to obtain the requested expressive performances. Second, to find out which mode worked better, and which was preferred by the participants. Third, to verify that the gestures corresponded to the emotion expressed by the performance.

The analysis of the log data showed that in the case of the *Marbles in a box* mode, the participants did not understand the connection between tilt and valence (see section 2.5): most of the time, they held the phone horizontal,

Table 1: Mean and standard deviation (Std) of the answers to questions 1 and 2 for the different modes. The questions were answered on a seven-steps Likert scale (1 = "Not at all", 7 = "Very much")

	Marbles		*Free*		Overall	
	Mean	Std	Mean	Std	Mean	Std
Q1	4.56	1.60	5.10	1.47	4.83	1.55
Q2	4.82	1.48	5.51	1.30	5.17	1.43

which resulted in a valence value around zero. On the other hand, the log data for the *Free movement* mode showed that the participants obtained a much wider range of values for activity and valence. This can be in part explained by the fact that, in this mode, the output from the classifier is discrete. Nevertheless, much better separation between emotions was obtained in the *Free movement* mode. This might reflect the fact that this mode was based on the analysis of free expressive gestures, and thus was more natural to understand.

The results from the log data were also reflected in the answers to the questions, which showed a strong preference (92%) for the *Free movement* mode. Statistical analysis of the answers to the second question also revealed that, according to the participants' perception, the agreement between the gestures and the emotions was significantly higher for the *Free movement* mode ($F(1, 12) = 8.748$, $p = 0.012$). All in all, the answers to the two questions revealed that the participants judged both modes to work relatively well (see Table 1).

4. CONCLUSIONS

We presented here *MoodifierLive*, an application in which findings from several previous studies have been implemented on a handheld device. The application allows for interactive control of rule-based automatic music performance, through five interaction modes, of which two based on gestures, and one collaborative. Results from evaluation showed that the application is interesting, fun and relatively intuitive to use.

The limits of the mobile phones used to test the application (i.e. Nokia N95), and specifically of the built-in accelerometer, were the reasons behind some design choices, especially the use of a simple classification tree for gesture recognition, based on a very limited number of features.

The use of more powerful devices would allow us to implement more advanced solutions to several problems. In the *Free movement* mode, a better classifier, which takes into consideration also the time evolution of gestures (e.g. [1]), would improve the expressive possibilities of the system. Furthermore, an automatic calibration of the gesture range would also improve the response of the system to a specific user.

A demo video showing the functionalities of *MoodifierLive* can be found at: `http://www.youtube.com/watch?v=m_9TMmTpjAw`.

5. ACKNOWLEDGMENTS

This study was partially funded by the Swedish Research Council (Grant Nr. 2010-4654), and by the EU SAME project (FP7-ICT-STREP-215749):
`http://www.sameproject.eu/`

6. REFERENCES

[1] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana. Continuous realtime gesture following and recognition. In S. Kopp and I. Wachsmuth, editors, *Gesture in Embodied Communication and Human-Computer Interaction*, volume 5934 of *LNAI*, pages 73–84. Springer, Heidelberg, 2010.

[2] R. Bresin. What is the color of that music performance? In *Proc. Int. Computer Music Conf. (ICMC2005)*, pages 367–370, Barcelona, Spain, 2005.

[3] R. Bresin and A. Friberg. Emotional coloring of computer-controlled music performances. *Computer Music J.*, 24(4):44–63, 2000.

[4] R. Bresin and A. Friberg. Emotion rendering in music: Range and characteristic values of seven musical variables. *CORTEX*, 2011, accepted for publication.

[5] A. Camurri, G. Volpe, H. Vinet, R. Bresin, E. Maestre, L. Javier, J. Kleimola, V. Välimäki, and J. Seppanen. User-centric context-aware mobile applications for embodied music listening. In *Proc. of the 1st International ICST conference on User Centric Media*, 2009.

[6] S. Dahl and A. Friberg. Visual perception of expressiveness in musicians' body movements. *Music Perception*, 24(5):433–454, 2007.

[7] A. DeWitt and R. Bresin. Sound design for affective interaction. In A. C. Paiva, R. Prada, and R. W. Picard, editors, *Proc. Affective computing and intelligent interaction (ACII2007)*, volume 4738 of *LNCS*, pages 523–533. Springer, Berlin / Heidelberg, 2007.

[8] M. Fabiani, R. Bresin, and G. Dubus. Sonification of emotional expressive gestures with automatic music performance on mobile devices. *J. Multimodal User Interfaces*, 2011, Submitted.

[9] M. Fabiani, G. Dubus, and R. Bresin. Interactive sonification of emotionally expressive gestures by means of music performance. In R. Bresin, T. Hermann, and A. Hunt, editors, *Proc. ISon 2010 - Interactive Sonification Workshop*, 2010.

[10] A. Friberg. pDM: an expressive sequencer with real-time control of the KTH music-performance rules. *Computer Music J.*, 30(1):37–48, 2006.

[11] A. Friberg, R. Bresin, and J. Sundberg. Overview of the KTH rule system for musical performance. *Advances in Cognitive Psychology, Special Issue on Music Performance*, 2(2-3):145–161, 2006.

[12] P. N. Juslin and P. Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5):770–814, 2003.

[13] M. Mancini, G. Varni, J. Kleimola, G. Volpe, and A. Camurri. Human movement expressivity for mobile active music listening. *Journal on Multimodal User Interfaces*, 4:27–35, 2010.

[14] M. Rohs and G. Essl. Camus^2 - collaborative music performance with mobile camera phones. In *Proc. Int. Conf. Advances in Computer Entertainment Technology (ACE2007)*, Salzburg, Austria, 2007.

[15] K. Schneider and R. F. Zernicke. Jerk-cost modulation during the practice of rapid arm movements. *Biological Cybernetics*, 60(3):221–230, January 1989.

[16] J. Williamson, R. Murray-Smith, and S. Hughes. Shoogle: Multimodal excitatory interfaces on mobile devices. In *Proc. Computer Human Interaction Conf. (CHI2007)*, San Jose, CA, USA, 2007.