INTERACTIVE SONIFICATION OF EMOTIONALLY EXPRESSIVE GESTURES BY MEANS OF MUSIC PERFORMANCE

Marco Fabiani, Gaël Dubus, Roberto Bresin

The Royal Institute of Technology
Stockholm, Sweden
{himork,dubus,roberto}@kth.se

ABSTRACT

This study presents a procedure for interactive sonification of emotionally expressive hand and arm gestures by affecting a musical performance in real-time. Three different mappings are described that translate accelerometer data to a set of parameters that control the expressiveness of the performance by affecting tempo, dynamics and articulation. The first two mappings, tested with a number of subjects during a public event, are relatively simple and were designed by the authors using a top-down approach. According to user feedback, they were not intuitive and limited the usability of the software. A bottom-up approach was taken for the third mapping: a Classification Tree was trained with features extracted from gesture data from a number of test subject who were asked to express different emotions with their hand movements. A second set of data, where subjects were asked to make a gesture that corresponded to a piece of expressive music they just listened to, were used to validate the model. The results were not particularly accurate, but reflected the small differences in the data and the ratings given by the subjects to the different performances they listened to.

1. INTRODUCTION

The strong coupling between motion and sound production, and in particular between body gestures and music performance has been investigated and documented in recent years (for an overview see [1]).

In the work presented in this paper the focus is on the relationship between body gestures and emotionally expressive music performance. The idea behind the application presented here is to use music and music performance rules to mediate the sonification of gesture data that contain emotional cues. This is a slightly different approach to sonification, if compared to the usual mapping of (reduced) data to, for example, sound synthesis parameters. We apply a higher level mapping in which the meaning of gestures is identified and mapped into the expressive meaning conveyed by a music performance. Although it is possible that part of the emotional content of the data is blurred by the intrinsic emotional content of the select piece of music, it is nevertheless accepted ([2] for an overview) that it is possible to express different basic emotion through changes in the performance of a piece of music.

A software called PyDM was developed that allows real-time control of an expressive music performance. It uses the KTH rules system for musical performance [3] to map different emotions (*e.g.* *happiness*, *anger*, *sadness*, *tenderness*) to time varying modifications of tempo, sound level and articulation. Rules can also be controlled independently to achieve more fine-tuned results. PyDM uses a special score file format where information from a MIDI score is augmented with pre-computed rule values. During playback, these values are weighted and summed to obtain the desired performance. The various parameters can be controlled remotely via OSC [1] messages. For this experiment, the messages were sent from a mobile phone, that was used as a remote controller to collect gesture data using the built-in accelerometer.

One way the user can control the emotional expression is by navigating in the so-called Activity-Valence space: different basic emotions can be placed in a 2D space where activity is on the the horizontal axis (*e.g.* low activity for *sadness* and *tenderness*, high activity for *happiness* and *anger*), and Valence on the vertical axis (*e.g.* positive Valence for *happiness* and *tenderness*, negative Valence for *sadness* and *anger*). In PyDM, a colored circle can be moved around in the Activity-Valence space using the mouse to "navigate" through the emotional space. The color of the circle changes according to the emotion, following a study by Bresin [4], whereas the size of the circle changes with the degree of activity (large for high activity, small for low activity).

2. BASIC EMOTIONAL EXPRESSION AND ITS SONIFICATION

In this paper, three different approaches to mapping gesture data to expressive performance parameters are presented. The first, and most basic, mapping is the direct control of the values of Activity-Valence ("Balance the performance"). In a simple virtual two dimensional space the user moves and tries to balance a virtual ball, positioning it in the area corresponding to the desired emotion. The position of the ball in the space is computed using the data from the phone's accelerometer.

A second approach tries to map different gestures directly to Activity-Valence values. The metaphor used for this approach is that of a small box filled with marbles (thus the name, "Marbles in a box") which is shaken in different ways to express different emotional states. The mapping, in this case, is less direct. The accelerometer data are analyzed in real-time on a frame-by-frame basis (the frame size can be set by the user). The Root Mean Square of the acceleration, which is related to the energy of the movement, or quantity of motion, is directly mapped to the Activity value. The sampling frequency of the phone's accelerometers is $f_s = 33$ Hz. In the application, a frame length $F = 40$ samples is normally used, which means the Activity and Valence values are updated every 1.2 seconds. The Valence value is coupled to the tilt of the phone: a vertical, upward position corresponds to maximum positive Valence; a horizontal position corresponds to a neutral Valence; a vertical, downward position corresponds to a maximum negative Valence. This mapping was designed by observing that positive Valence emotions can be expressed with "hands up" gestures (and thus, the phone is held in a vertical, upward position); on the other hand, negative Valence emotions can be expressed with "hands down" gestures. The "Marbles in a box" mapping, although slightly more related to the actual data and based on the

[1] Open Sound Control

normal behavior of the users, is rather arbitrary, and somehow *imposed* on the user. As a consequence the user must learn and follow the mapping to obtain the desired emotion. These considerations led to the design of a third mapping, which is extensively described in the following section.

From a sonification point of view, and according to the taxonomy proposed by Hermann [5], our first approach (direct navigation through he Activity-Valence space) constitutes a simple Parameter-Mapping Sonification, where the position is mapped to the rules of the KTH system for music performance via the program PyDM. The second and the third approaches can somehow be considered as hybrid methods, since they make use of a model of different complexity to associate the user's gestures with a position in this intermediary space, followed by the aforementioned parameter-mapping method.

3. DATA-DRIVEN EMOTIONAL MAPPING: PILOT EXPERIMENT

The "Balance the performance" and "Marbles in a box" mappings were tested during the Agora Festival 2009 in Paris [2] with a large number of users, during an event to display different mobile applications developed during the SAME project [3]. From the formal feedback provided by 36 users and from personal conversations it emerged that, although the PyDM application was fun and interesting to use, the control part based on gestures could be made more interesting and engaging. This led us to consider a different approach to data mapping, based on more advanced gesture recognition. For this reason, a pilot experiment has been designed to collect emotional gesture data. Different features can be extracted from the data and analyzed to expose possible commonalities between different users in expressing the same emotion. The common features can then be used to train a model that recognizes the different basic emotions and maps them to a musical performance.

3.1. Data collection

Since the first experiments with the accelerometers built-in in the mobile phone, it emerged that their small range (about $\pm 2\,g$) limits the effectiveness of the gesture control: data quickly saturate when fast gestures are performed. For this reason, we decided to use, alongside the phone's built-in accelerometer, an accelerometer with a wider range ($\pm 6\,g$), the WiTilt V3 [4]. It comes in a small enclosure, and the data are sent via Bluetooth. The sampling rate of the WiTilt was set at $80\ Hz$. For the data collection in the pilot experiment, we attached the WiTilt to the phone (iPhone 3G) using strong rubber bands. Data from the iPhone were sent through a WiFi network using the OSC protocol. Both WiTilt and iPhone accelerometer data were saved along with timestamps to allow for a comparisons of the two. The experiment was controlled through a Python script that managed the different connections and saved the data to text files for later analysis.

3.2. Experiment design

The experiment comprised three parts: calibration, emotional gesture without music, and emotional gesture with music. In the first part of the experiment, the calibration, subjects were asked to perform a fast movement, and then a slow movement, and were given 10 seconds for each one of the two gestures.

For the second task (emotional gesture without music), subjects were asked to perform four gestures that expressed the four basic emotions *happiness*, *anger*, *sadness* and *tenderness/love*. The order of the emotions was randomly chosen, and subjects were given 10 seconds to perform each one of the four gestures.

The final task, the more complex, comprised three parts. First, subjects were asked to listen to one of 16 musical clips[5], between 10 and 20 seconds long, and rate it on four different scales according to how much *happy*, *sad*, *angry* and *tender* they perceived each musical excerpt. The scales had values from 1 to 7, where 1 corresponded to "not at all" and 7 corresponded to "very much". Each clip could be listened only once. The rating was introduced to compare the emotion perceived by subjects with the intended emotion of the musical clips, and with the gestures. The 16 clips were created from a combination of four melodies and four sets of expressive performance parameters, and produced using MIDI files and a high quality synthesizer. The four melodies were specifically composed at McGill University for this type of experiment, and to be inherently expressing one of the four basic emotions [6]. For the expressive performance, seven musical parameters (tempo, sound level, articulation, phrasing, register, instrument, and attack speed) were varied according to a set of values used in a previous experiment conducted by Bresin and Friberg [7]. The effectiveness of the values for the four basic emotions was verified in [8] and will not be discussed here. To give an example, the *happy* performance had a fast tempo, *staccato* articulation, high sound level and bright timbre (trumpet), whereas a *sad* performance had a slow tempo, *legato* articulation, low sound level and dull timbre (flute).

After listening and rating one musical excerpt, the order of which was randomly chosen, the subject was given 10 seconds to perform a gesture that represented the music she had just listened to. We decided not to let the subject perform the gesture *while* listening to the music because we wanted to remove the influence of "directing" the music as much as possible, which would have meant reducing the task to just keeping the tempo.

Eight subjects (six male, two female) were recruited among students and researchers at the Dept. of Speech, Music and Hearing at KTH. They were aged between 24 and 44. All except one had some musical experience playing an instrument. They all actively listened to music on a regular daily basis. The subjects participated to the experiment without receiving any compensation.

3.3. Data analysis

In the following analysis, the calibration data mentioned in Sec. 3.2 was not used. The iPhone data were compared to the WiTilt data, and it was shown that the correlation between the two signals was very high for the *happy*, *sad* and *tender* gestures (~ 0.95 on average), while it was lower for the *angry* gestures (~ 0.8). This reflects the fact that the *angry* gestures were faster and more impulsive, thus saturating the output from the iPhone accelerometers. We decided to use only the WiTilt data for the gesture analysis. An example of the accelerometer signals for one of the subjects is shown in Fig. 1.

3.3.1. Features extraction

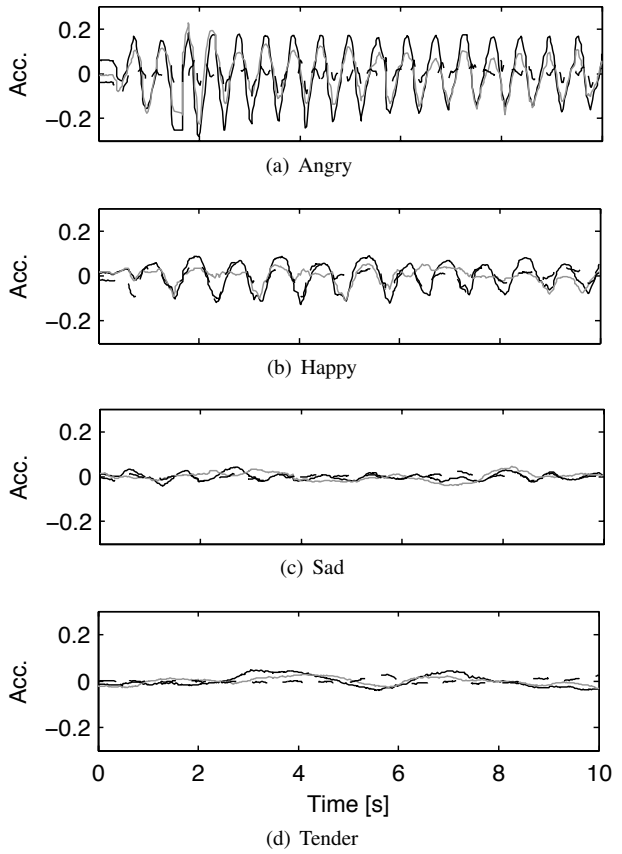
A set of features was chosen that could well describe the different characteristics of the emotional gestures. Some of these features were also used in other applications, such as the "Fishing game" [9], presented at the Agora Festival 2009. Different features were extracted from the signals, such as frequency, periodicity and energy. An estimate of the velocity in the three directions was computed by integrating the acceleration over time and subtracting the

[2] Agora Festival 2009: `http://agora2009.ircam.fr/`

[3] SAME, FP7-ICT-STREP-215749, `http://sameproject.eu/`

[4] WiTilt V3: `http://www.sparkfun.com/`

[5] Musical clips used in the experiment: `http://www.speech.kth.se/music/papers/2010_MF_ISon/`



(a) Angry

(b) Happy

(c) Sad

(d) Tender

Figure 1: Scaled acceleration data for Subject 5 (black line: x-axis; grey line: y-axis; dashed line: z-axis). Fig. (a) shows an *angry* gesture; (b) shows a *happy* gesture; (c) shows a *sad* gesture; (d) shows a *tender* gesture.

mean to remove the bias from Earth's gravity. The *jerkiness* of the signal, which is defined in [10] as the Root Mean Square of the derivative of the acceleration, was also extracted. Means and standard deviations of the different features were finally computed.

3.3.2. Gesture modeling

Different models from machine learning were considered to automatically classify gestures, such as Classification/Regression Trees, Neural Networks, Support Vector Classifiers, and Fuzzy Classifiers (a Fuzzy Classifier was previously used for a similar task by Friberg in [11]). We decided to start by testing the simplest option, a Classification Tree, which can also be easily implement on a low power device such as a mobile phone. By visual inspection it was clear that differences within subject for each emotion were quite significant, but the absolute values of the features between subjects were rather different. For this reason, the data from each subject were first standardized by subtracting the mean and dividing by the standard deviation of the data from all the four emotional gestures for that particular subject. As a consequence of the standardization, the data used for the classification were the relative differences between different emotions, instead of the features' absolute values. The drawback of doing so is that before a new user starts using the system, a calibration is required to collect data for the standardization. This can be done explicitly by asking the user to perform the four basic emotional gestures, or by adaptively correcting the standardization parameters during the normal use of the application.

Table 1: Confusion matrix for the classification of the gestures performed after listening to the expressive clips. The rows contain the expected emotion, the columns the predicted emotion.

	Angry	Happy	Sad	Tender
Angry	0	**28**	9	4
Happy	0	**26**	3	3
Sad	0	4	7	**21**
Tender	0	4	11	**17**

From a scatter plot it was possible to see that most of the features were strongly correlated to the energy of the signal. In the end, it was clear that the best candidates for a simple Classification Tree were the mean jerkiness and the mean velocity. The Classification Tree was trained using vectors of feature values extracted from the gestures performed without the music. Cross-validation was used to determine the minimum-cost tree. The resulting tree was:

```
if Mean Jerkiness > 0.78
   ANGRY
else
   if Mean Jerkiness > -0.48
      HAPPY
   else
      if Mean Velocity > -0.35
         SAD
      else
         TENDER
```

With only eight subjects, the risk of over-fitting the data is very high, so the results in this paper are to be considered very preliminary. In case a smooth variation between emotions is desired, a Regression Tree can be used. Similar results can also be obtained using the Fuzzy Classifier described in [11].

3.3.3. Model evaluation

The data from the gestures performed after listening to the 16 musical excerpts were standardized with the means and standard deviations from the training set, and used to evaluate the model. A confusion matrix of the classification compared to the nominal emotion (that of the performance defined by the parameters described in Sec. 3.2) is shown in Tab. 1. There is a very clear separation between high activity (*happy* and *angry*) and low activity (*sad* and *tender*) emotions. The classification on the whole did not perform very well: most of the gestures were labeled as either *happy* or *tender*. This was partly expected in the case of the confusion between *sad* and *tender*, since it can be seen (Fig. 1) that there is almost no difference in the data (in fact, from informal conversations with the subjects it emerged that it was very difficult to actually express the difference between *sad* and *tender*). The classifier thus marked most of the gestures after a *sad* performance as *tender*. Less expected, because of the much clearer separation in the training data, was the fact that most of the gestures after an *angry* performance were identified as *happy*. It was visually observed by the authors that after listening to the music, less "extreme" gesture were performed compared to the case in which an *angry* gesture was explicitly asked. The incorrect classification of *angry* gestures can be also justified by the conversations with the subjects, who pointed out that there were very few really *angry* performances in the 16 clips. Therefore, it sounds more promising for future developments of the system to consider only emotional gestures which are not performed after listening to a musical clip,

since the idea is to sonify gestures using musical clips, i.e. music comes after the user's gesture, and not vice versa.

A rough analysis of the ratings further justifies the relatively poor performance of the classification. Among subjects there was a very high variance in the ratings of the different clips. This is in part a consequence of the small number of subjects. It can also be seen that the easiest emotion to identify was *happiness*. For many clips, *sadness* was confused with *tenderness*, and *anger* with *happiness*, similar to what happened with the Classification Tree. In one case, *tenderness* was confused with *happiness*. A strong influence on the rating of a performance came also from the intrinsic emotion expressed in the four melodies, which in certain cases was the opposite of the one expressed by the performance parameters, thus adding to the confusion.

4. CONCLUSIONS

A way to indirectly sonify emotional gesture data collected through an accelerometer was presented in this paper. The data are mapped to a set of performance rules that affect the tempo, sound level and articulation of a musical score, effectively changing the emotional expression of the music. Three different mappings were described. Two basic mappings, decided a priori by the authors, were felt by the users as being not intuitive. Thus, a data-driven mapping was designed by first collecting gesture data from eight test subjects, then extracting a number of features, and finally training a simple Classification Tree. The evaluation gave relatively poor results. This was partly expected from the observation of the rough accelerometer data, from informal conversations with the subjects, and after looking at the large variance among subjects in the emotion ratings given to the music they were supposed to represent with their gestures.

It is possible that the behavior of the users will adapt to the system when the classifier will give a real-time audio feedback, thus leading her to, for example, express *anger* in a more stereotypical manner. An evaluation of the real-time system is required to fully understand if the mapping is capable of effectively translating emotional gestures into a corresponding music performance. Furthermore, the small number of subjects used in this pilot experiment strongly reduced the statistical power of the ratings analysis and probably led to over-fitting in the training of the classifier.

Future work includes a new data collection with a larger number of subjects; the use of more sophisticated classifiers; the evaluation of the real-time system; a more thorough analysis of the ratings; the use of other techniques for identifying emotions, such for example stereotypical gestures, as described in [9].

5. ACKNOWLEDGEMENTS

This experiment was funded by the SAME project (FP7-ICT-STREP-215749), `http://sameproject.eu/`.

6. REFERENCES

[1] Rolf I. Godøy and M. Leman, Eds., *Musical Gestures: Sound, Movement, and Meaning*, Routledge, 2009.

[2] Patrik N. Juslin and John A. Sloboda, Eds., *Music and emotion: theory and research*, Oxford University Press, Oxford (UK), 2001.

[3] Anders Friberg, Roberto Bresin, and Johan Sundberg, "Overview of the KTH rule system for musical performance," *Advances in Cognitive Psychology, Special Issue on Music Performance*, vol. 2, no. 2-3, pp. 145–161, 2006.

[4] Roberto Bresin, "What is the color of that music performance?," in *Proc. of the International Computer Music Conference (ICMC2005)*, Barcelona (Spain), 2005, pp. 367–370.

[5] Thomas Hermann, "Taxonomy and definitions for sonification and auditory display," in *Proceedings of the 14th International Conference on Auditory Display (ICAD 2008)*, Brian Katz, Ed. ICAD, June 2008, ICAD.

[6] S. Vieillard, I. Peretz, N. Gosselin, S. Khalfa, L. Gagnon, and B. Bouchard, "Happy, sad, scary and peaceful musical excerpts for research on emotions," *Cognition & Emotion*, vol. 22, no. 4, pp. 720 752, 2007.

[7] Roberto Bresin and Anders Friberg, "Emotion rendering in music: range and characteristic values of seven musical variables," *CORTEX*, submitted.

[8] Tuomas Eerola, Anders Friberg, and Roberto Bresin, "Emotion perception in music: importance and additive effects of seven musical factors," *CORTEX*, submitted.

[9] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana, "Continuous realtime gesture following and recognition," *LNAI*, vol. 5934, pp. 73–84, 2010.

[10] K. Schneider and R. F. Zernicke, "Jerk-cost modulations during the practice of rapid arm movements," *Biological Cybernetics*, vol. 60, no. 3, pp. 221–230, 01 1989.

[11] Anders Friberg, Erwin Schoonderwaldt, and Patrik N. Juslin, "CUEX: An algorithm for extracting expressive tone variables from audio recordings," *Acta Acustica united with Acustica*, vol. 93, 2007.