PHYSICAL MOVEMENT AND MUSICAL GESTURES: A MULTILEVEL MAPPING STRATEGY

Diego Fenza, Luca Mion
CSC-DEI
University of Padua

Sergio Canazza, Antonio Rodà
MIRAGE
University of Udine

ABSTRACT

Musical interpretations are often the result of a wide range of requirements on expressiveness rendering and technical skills. Aspects indicated by the term expressive intention and which refer to the communication of moods and feelings, are being considered more and more important in performer-computer interaction during music performance. Recent studies demonstrate the possibility of conveying different sensitive content like expressive intentions and emotions by opportunely modifying systematic deviations introduced by the musician. In this paper, we present a control strategy based on a multi-layer representation with three different stages of mapping, to explore the analogies between sound and movement spaces. The mapping between the performer (dancer and/or musician) movements and the expressive audio rendering engine resulting by two 3D "expressive" spaces, one obtained by the Laban and Lawrence's effort's theory, the other by means of a multidimensional analysis of perceptual tests carried out on various professionally performed pieces ranging from western classical to popular music. As an example, an application based on this model is presented: the system is developed using the eMotion SMART motion capture system and the Eyesweb software.

1. INTRODUCTION

Today's sensor technologies and motion capture systems virtually make any kind of physically expression able to be detected and tracked. Computer-based sound generation is now a mature technology. This paper focuses on the control, rather than on synthesis of music, by means of body movements which can be performed by a dancer or even by a musician in artistic contexts, by a player in a mixed-reality environment for extreme-gaming or by a user of a medical-therapeutic system. The control is becoming more interesting both in a practical sense of technologies control, and in a more theoretical and general sense: what can a performer control, and to what extent? Thus, control has a double aspect: tools and strategies for obtaining a musical result, also representing a way to convey a musical expression to the listeners.

Expressive contents can be conveyed, among other, by means of the deviations in timing, in dynamics, in timbre, in tempo, which are not written in the score and are always introduced by the performer. They generally differ according to the musical genre, to the particular instrument and to the performer itself, thus different players can produce considerably differing performances indeed, even if using the same score: some studies [10], [9], [15] demonstrated that different performances of the same piece can communicate different expressive intentions. Most music performances involve expressive intentions from the performer's side, regarding what the music should "express" to the listeners. Consequently, interpretation involves assigning some kind of meaning to the music. Performer's intentions are captured by the listener as better as they do share a common coding of musical expression, and a possible aim could be examining how these intentions are captured. This approach was suggested by Seashore [14], who asserted that psychophysics relations between the performer and the listener are fundamental for the comprehension of the microstructures of the musical performance.

In order to conceptualize expressive intentions we can consider two different approaches [12]: categorical and dimensional. The former assumes that people experience expressive intentions as categories that are distinct from each other. It is typically used in emotion research to categorize basic emotions from which all other emotional states can be derived. Dimensional approach focuses on the identification of expressive intentions based on their locations in a low-dimensional space. Categorical representations may apply quite well to basic emotions, but much less so elsewhere; moreover, labels as such are very poor descriptions. These considerations encouraged us to use the dimensional approach to implement a mapping function between movements and musical expressive intentions.

The choice of a mapping strategy for the interactive control of a device for the real-time generation of sound objects (in this context, the "musical instrument" term is not intentionally used, for the sake of generality) comes from several considerations, amongst other: i) the application goal, ii) the availability of metaphors and/or structural relations for connecting the input and output domains. The aim of an application which generates sounds controlled by movements can be of two kinds: functional or artistic. By *functional*, we mean those applications which have the purpose to solve a well-defined problem, for instance the implementation of human-machine interfaces based

on multimodal communication [11], or the use in the therapeutic field for the rehabilitation of motor activities. In this case, the mapping has the main purpose to translate the information from the input to the output domain; sounds becomes a reinforcement of the motor gestures. In the case of *artistic* applications, the mapping design can be driven by many principles, which exclusively refer to the composer/artist's aesthetic sensibility; for instance, it is possible to use the sound to confirm or contradict the information implied in the motor gesture. In both cases, functional or artistic, the design of the mapping cannot overlook the knowledge of the structural relations between movement and sound gestures. The purpose of the system presented in this paper is to propose an architecture to explore the analogies between sound and movement spaces.

This paper is organized as follows: section 2 describes the system architecture; section 3 explains the strategy to map movements in musical gestures, introduces the multilevel approach chosen to split up the problem of expressive gesture analysis and mapping into different sub-problems, presents the movement analysis models (inspired to the theories of Laban and Lawrence [13]), and describes the abstract space we used to control the expressive audio rendering; section 4 presents a preliminary HW/SW implementation of the system, using the eMotion SMART motion capture system (www.emotion3d.com/smart/smart.html) and EyesWeb open software platform (www.eyesweb.org).

2. SYSTEM ARCHITECTURE

The system is composed by several units: i) a motion capture equipment, ii) a movement analysis tool, that calculate low and mid level cues starting from the motion capture data, iii) a mapping function, that maps movement cues to audio cues, iv) an audio processing engine, that generate real-time controlled sound objects. The system uses the streaming data coming out from the 3D-motion capture system eMotion SMART. SMART is an optical motion capture system, consisting of 6 cameras with IR light strobes and 1 power supply and synchronization unit. The movement analysis tool was implemented using the EyesWeb open software platform. EyesWeb is a graphical environment planned for the processing of multimedia data streams and the creation of audio/video interactive applications, in which the user can build up an application using a library of blocks that can be connected together into a patch. In our experiment we developed some new blocks that read data from SMART, extract movement parameters, perform functions of mapping on different spaces and process a recorded audio file by mean of VST plugins. We finally connected them in a patch for the real time interaction and data are sent from SMART system (server) to EyesWeb (client), respectively writing and reading a data flow in a memory area (dataport).

3. MAPPING STRATEGY

With the aim to realize a flexible and modular system, an architecture based on a three stages mapping was designed. A similar architecture has already been taken into account by [1] and [17], but the internal representation, the movement and audio cues, and the definition of the different stages are very different from those proposed in this paper. With reference to Fig. 1, the input signals coming from the acquisition system are processed calculating a vector of movement cues, which will be defined in section 3.1; the M1 function maps such vector in a three-dimensional space (3D movement expressive space) defined starting by the Laban's theories [13] (see 3.2); the M2 function maps a trajectory in the 3D movement expressive space to a trajectory in the 3D musical expressive space; finally, the M3 function maps a point in the musical expressive space in a vector of audio cues, which become control signals for the audio processing engine [8], [5].

3.1. Movement cues

In this work we refer to the multi-layer approach investigated by [2], in order to split up the problem of expressive gesture analysis and mapping into different sub-problems. This allows a flexible interpretation of the concepts in relationship to emotional, affective and sensitive processing. The movement cues can be grouped according to the three-layer model shows in Fig. 2.

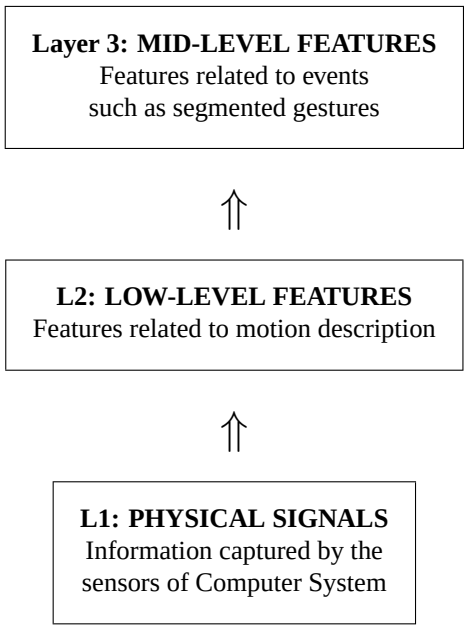

Figure 2. The conceptual framework, consisting of three layers, used for motion analysis

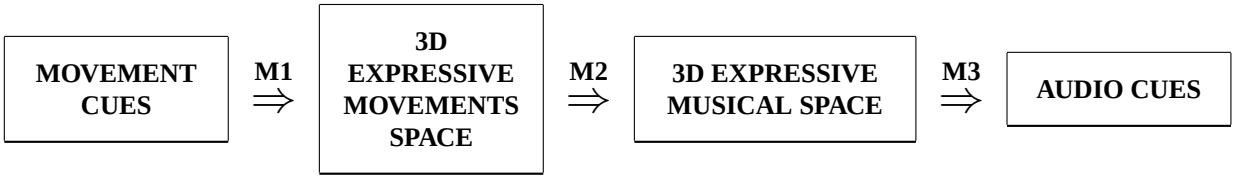

Figure 1. Multilevel mapping between movement and audio cues.

3.1.1. Layer 1

Layer 1 gets data from SMART and returns, for each frame, a matrix which contains the coordinates of every tracked point. Data flow is written at the operating sampling frequency of the capture system (50 Hz). A matrix with dimension $n \times 3$, where n is the number of points in each frame, is then written by the server every 20 ms

3.1.2. Layer 2

Layer 2 performs the cues extraction by means of statistical and signal processing techniques. The cues we extracted are: Quantity of Motion (QoM), Contraction Index (CI), Movement Length (ML), Straight trajectory Length (SL), and Directness Index (DI).

Being QoM proportional to the distance (for each point mass and sampling frequency are constants), we computed the $QoM(n)$ at frame n following equation 1

$$QoM(n) = \sum_{i=1}^{N_p} m_i \, |\mathbf{p}_i(n) - \mathbf{p}_i(n-1)| \qquad (1)$$

where N_p is the number of tracked points in the frame, m_i is the mass associated to the $i{-}th$ point, and $|\mathbf{p}_i(n) - \mathbf{p}_i(n-1)$ is the euclidean distance between the position of point $i-th$ in two successive frames.

The second cue CI is related to the body contraction or expantion, and it is computed considering the barycenter of the points in the frame $n-th$. We fixed the problem of not having a reference measure for the contracted position of the body by setting, for each point, a constant named MIN_DIST_i describing the minimum distance of the point $i-th$ from the barycenter in the most possible contracted position. Then we computed $\widetilde{CI}(n)$, given by equation 2). This value is inversely proportional to the contraction, so the default value of $\widetilde{CI}(n)$ is 0.

$$\widetilde{CI}(n) = \tfrac{1}{N_p} \sum_{i=1}^{N_p} |\mathbf{p}_i(n) - \mathbf{bar}(n)| - MIN\_DIST_i \qquad (2)$$

where $|\mathbf{p}_i(n) - \mathbf{bar}(n)|$ is the distance, at the frame $n-th$, between the point $i-th$ and the barycenter

It is more useful to consider the value of the ***CI(n)*** belonging to the range $[0, 1]$, with the meaning of maximum contraction when $CI(n) = 1$ and of maximum expansion when $CI(n) = 0$. We can so define $CI(n)$ as the equation 3.

$$CI(n) = \frac{1}{1 + \widetilde{CI}(n)} \qquad (3)$$

The other cues can be extracted from the trajectory motion: ML cames from the sum of all segments which join the points of a trajectory; SL is the measure of the segment which join the first and the last trajectory point. DI is the ratio of SL to ML; this parameter gets a deviation index of the movement from the direct trajectory, and it explains if the trajectory is direct or flexible; if DI is next to zero, we are considering a very large movement.

3.1.3. Layer 3

The main process of this layer is the segmentation during the motion and the pause phases: to perform this, we set a threshold on the QoM, so that the movement is detected after raising over the threshold. This yields some *motion bells* in the QoM diagram: every motion bell identifies a motion phase (see Fig. 3), related to the movement fluidity.

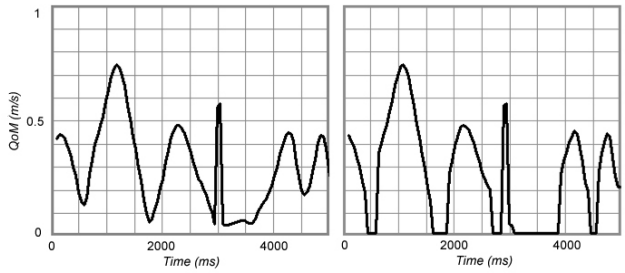

Figure 3. The *motion bells*

During our experiment we set several thresholds (absolute, relative to the space filled by the body and a threshold relative to the QoM itself). All the parameters showed in 3.1.2 can also be recomputed for the motion phase, in order to classify every single movement. In this way we compared, for example, the movement length for two different movements (see Fig. 4). In this layer we also computed the fluidity and impulsiveness index, which are directly connected to the Laban's theory. To define these indexes, we have to consider three distinct time periods: Movement time (t_m) is the time covered by the motion phase itself. Pause time (t_p) is the sum of the pause times before and after the movement. Total time (t_t) is given by the sum of t_m and t_p. Then we can define, for each segmented gesture, the fluidity as follows, assuming values in the range $[0, 1]$:

$$F = \frac{t_m}{t_t} \qquad (4)$$

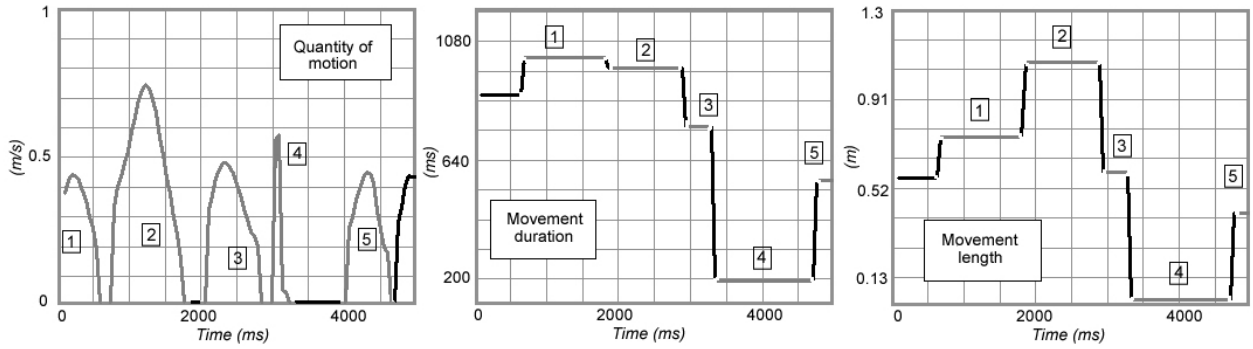

Figure 4. Movement classification. From the left, respectively the quantity of motion, the motion duration and the motion length are depicted. Movement 4 has a very tight motion bell, and the related motion duration (central graph) has a low value (about $200ms$). Movement 2 has a large motion bell and long movement (right).

To define the impulsiveness we introduce two new constants: `MIN_IMPULSIVE_TIME`, which is the minimum time duration for a movement; `MAX_DISTANCE`, representing the maximum distance between the positions of a point in two successive frames. The impulsiveness (I) is then given by the eq. 5:

$$I = \frac{Nm \cdot \texttt{MIN\_IMPULSIVE\_TIME}}{t_m} \cdot \frac{max\_mov\_len}{\texttt{MAX\_DISTANCE}} \quad (5)$$

where Nm is the number of movements considered, and max_mov_len is the maximum value of the average distances of the points in next frames, within the considered time interval. I assumes values between 0 and 1. Other cues belonging to the third level are the number of Gestures per Second (GPS) and the Movement Velocity (MV).

3.2. 3D Expressive Movements Space

The 3D Evpressive Movements Space used for the mapping is derived from the Laban and Lawrence's effort's theory. Following this theory, the expressive content of every physical movement is mainly related to the way of performing it, and it is due to the variation of four basic factors: time, space, weight and flow. The authors defined as basic efforts the eight combinations of two values (quick/sustained, flexible/direct and strong/light) associated with the first three factors. Each combination gives rise to a specific expressive gesture to which is associated an adjective, for instance a slashing movement is characterized by a strong weight, quick time and flexible space (i.e., a curved line). A 3D Space, named the Laban's cube, can be derived from the three categories Time, Space and Weight [13].

3.3. 3D Expressive Musical Space

To define a control system for the interactive control of musical expressiveness, it is necessary to understand the performer's expressive intentions, by means of perception measurements. Performances played according to different expressive intentions have been evaluated in listening experiments. Then, a low dimensional structure is derived by multivariate analysis of response data. In [3] and [4], the analysis of results led to three distinct axes. From a sonological point of view, the first axis sets rapid Tempo against slow; the second one is mainly correlated to Energy-related parameters as Intensity; the third axis is connected to the Brightness. Successful identification of the player's intention is demonstrated by the fact that each performance is placed near semantically related adjectives. An interpretation can be applied to this space. The first factor is related to Kinematics; the second factor is associated to Energy; the third factor is related to Brightness. Moreover, the KEB space proved to be effective in interactive control of expressivity in synthetic music performance [7], [8], [4].

4. SYSTEM IMPLEMENTATION

In this section, a preliminary implementation of this model will be presented. The SMART motion capture equipment detects and records the three-dimensional position in time of small passive non-invasive markers applied on the subject or object to be analyzed (see Fig. 5). In this system, markers were applied on six strategic parts of the subject's body: for each arm, one marker was placed at the wrist position, another one was placed close to the elbow and the third marker was sticked on the subject's shoulder. In order to analyse the movement features, the system uses the multilevel representation presented in section 3.1. The first layer is realized by two software modules reading data from the *SMART* file (or from the *dataport*). These modules return some matrixes as explained in section 3.1.1. Layer 2 consists of two modules too, to compute QoM and CI and to compute the parameters extracted from the motion trajectory (movement length, angles, velocity, etc..). Operation run by the third layer are computed by the module that in the layer 2 gives QoM and CI; this module performs the motion segmentation, and it is also controlled by one more input which describes whether there is a motion phase or not; in this way we can choose between the instantaneous modality or compara-

tively to the single motion phase (Fig. 4).

After analyzing the movements in the context of layers 1, 2 and 3, the system maps the parameters in the Laban's space by means of the M1 function (see Fig. 1), defined starting from the experimental results by Camurri and Volpe [16]. In particular, Number of Gestures per Second and Movement Velocity influence the Time axis, Quantity of Motion and Impulsiveness influence the Weight axis, Contraction Index and Directness Index influence Space axis. The following M2 function was defined hypothesizing an isomorphism between the expressive movement space (Laban's cube) and expressive musical space (KEB). In particular, from the definition of the spaces (section 3.2 and 3.3) derives a correspondence between Weight and Energy axes, Time and Kinematics axes, Space and Brightness axes. Although such a hypothesis do not have yet any in-depth experimental verification, some informal perceptive tests revealed a promising behavior. Finally, the M3 function was defined starting from the results reported by Canazza et al. [6], [7] and maps the KEB space into control signals for the audio rendering engine.

5. DISCUSSION

At the moment, a rigorous validation of the system presented in this paper was not carried out. However, from a series of informal tests, the following considerations emerge. The motion capture equipment provides entry data reliable and with very little noise; moreover, the data flow is suitable to be elaborated in time-real for the motion cues extraction. For against, the system presents a quite high cost and is hardly transportable in contexts outside the laboratory. The modular architecture allows to modify separately each of the mapping functions (M1, M2, M3), making an in-depth study of the various aspects related to the analogies between motor and musical gestures easier. Many issues deserve still further investigation. In particular: to evaluate the advantages of the motion capture equipment in comparison to systems based on a single telecamera, as that one presented in [18]; to evaluate the behaviour of the system in function of the number and the position of the markers on the performer's body; to evaluate if the set of the motion cues defined in section 3.1 is necessary and/or sufficient to represent the expressive qualities of the movement; to evaluate possible alternatives or improvements to the Laban' s cube and to the KEB Space; to study in a more depth way the mapping between the expressive movement space and the expressive musical space.

6. CONCLUSIONS

The system presented in this paper, thanks to his flexibility and modularity, is a starting point to explore the isomorphisms between physical and musical gestures. In this context, we considered the Laban's theory to analyze the movement and we used the KEB, a 3D expressive space obtained by means of a multidimensional analysis of perceptual tests, to control the audio rendering engine. In this way, we can control the music expressiveness by means of physical movement. The practitioners of this new art spring from many walks of life: academic researchers, musical performer and composers, dancers and choreographers, artistic designers, video game developers, interactive and media installation artists, teachers, and therapists (special needs, exercise, and relaxion).

7. ACKNOWLEDGEMENTS

This research was partially supported by the European Network of Excellence "Enactive Interfaces" (www.enactive.org). The authors would also like to thank Prof. Ruggero Frezza for useful discussions and for hosting our experiment to his lab.

8. REFERENCES

[1] Arfib, D., Couturier, J., Kessous, L., Verfaille, V. "Strategies of mapping between gesture data and synthesis model parameters using perceptual spaces", *Organized Sound*, vol. 7(2), pp. 127-144, 2002.

[2] Camurri, A., De Poli, G., Leman, M., Volpe, G. "A multi-layered conceptual framework for expressive gesture applications", *Proc. Workshop on Current Research Directions in Computer Music*, Barcelona, Spain, pp. 29-34, 2001.

[3] Canazza, S. "Analisi mediante test percettivi e modello dell'espressività del clarinetto nell'esecuzione musicale" *Master's Thesis*, Department of Information Engineering, University of Padova, 1996.

[4] Canazza S., De Poli G., Vidolin A. "Perceptual analysis of the musical expressive intention in a clarinet performance", In M. Leman (ed), *Music, gestalt, and computing, Studies in cognitive and systematic musicology*, Berlin, Heidelberg: Springer-Verlag, pp. 441-450, 1997.

[5] Canazza, S., De Poli, G., Drioli, C., Rodá, A., Zamperini, F. "Real-time morphing among different expressive intentions in audio playback", *Proc. of ICMC*, Berlin, Germany, 27 august-1 september, pp. 356-359, 2000.

[6] Canazza, S., De Poli, G., Drioli, C., Rodà, A., Vidolin, A. "Audio morphing different expressive intentions for Multimedia Systems", *IEEE Multimedia*, July-September, 7(3), pp. 79-83, 2000.

[7] Canazza, S., De Poli, G., Rodà, A., Vidolin, A. "An abstract control space for communication of sensory expressive intentions in music

Figure 5. Images from the implemented system. One of the wrist markers is depicted on the left. On the right the subject while testing the system; On the right side, a 3D plot of markers position and the partitioned QoM; These graphs are plotted in real time by the SMART interface.

performance", *Journal of the New Music Research*, 32(3), pp. 281-294, 2003.

[8] Canazza, S., De Poli, G., Drioli, C, Rodà A., Vidolin, A. "Modeling and Control of Expressiveness in Music Performance",(invited paper), *The Proceedings of the IEEE* vol. 92(4), pp. 286-701, 2004.

[9] Clarke, E. F. "Imitating and evaluating real and transformed musical performances", *Music Perception*, 10(3), pp. 317-341, 1993.

[10] Gabrielsson, A. "The performance of music", *The psychology of music*, New York: Academic Press, pp. 501-602, 1997.

[11] Hunt, A.D., Paradis, M., Wanderley, M. "The importance of parameter mapping in electronic instrument design", Invited paper for the *Journal of New Music Research*, SWETS, special issue on New Interfaces for Musical Performance and Interaction, eds. J. Paradiso & S.O'Modhrain, Vol. 32 No. 4, pp 429-440, 2003.

[12] Juslin, P. N. "Communicating emotion in music performance: A review and a theoretical framework", In *Music and emotion: Theory and research*, P. N. Juslin, & J. A. Sloboda (Eds.), pp. 305-333, New York: Oxford University Press, 2001.

[13] Laban, R., Lawrence, F. C. "Effort", *Macdonald & Evans Ltd.*, London, England, 1947.

[14] Seashore, H. G. "An objective analysis of artistic singer". In C. E. Seashore (Ed.) *Objective analysis of musical performance*, vol. 4, pp. 12-157. University of Iowa, 1937.

[15] Senju, M., Ohgushi, K. "How are the player's ideas conveyed to the audience?", *Music Perception*, 4(4), pp. 311-324, 1987.

[16] Camurri, A., Marazzino, B., Volpe, G. "Analysis of expressive gestures in human movement: the eyesweb expressive gesture processing library", *Proceedings of the XIV Colloquium on Musical Informatics (XIV CIM 2003)*, Firenze, Italy, 2003.

[17] Wanderley, M. M., Depalle P. "Gestural Control of Sound Synthesis", *The Proceedings of the IEEE* vol. 92(4), pp. 632-644, 2004.

[18] Camurri, A., Marazzino, B., Richetti, M., TImmers, R., Volpe, G. "Multimodal analysis of expressive gesture in music and dance performances", In A. Camurri and G. Volpe (eds.), *Gesture-based communication in human-computer interaction*, Berlin, Heidelberg: Springer-Verlag, pp. 20-39, 2003.