A Knowledge-based, Data-driven Method for Action-sound Mapping

Federico Visi, Baptiste Caramiaux, Michael Mcloughlin

A Knowledge-based, Data-driven Method for Action-sound Mapping

Federico Visi
Universität Hamburg, DE, EU
ICCMR, Plymouth University
Plymouth, UK, EU
mail@federicovisi.com

Baptiste Caramiaoux
McGill University,
Montreal, Canada
IRCAM, Paris, France, EU
baptiste.caramiaux@ircam.fr

Michael Mcloughlin
ICCMR
Plymouth University,
Plymouth, UK, EU
michael.mcloughlin@plymouth.ac.uk

ABSTRACT

This paper presents a knowledge-based, data-driven method for using data describing action-sound couplings collected from a group of people to generate multiple complex mappings between the performance movements of a musician and sound synthesis. This is done by using a database of multimodal motion data collected from multiple subjects coupled with sound synthesis parameters. A series of sound stimuli is synthesised using the sound engine that will be used in performance. Multimodal motion data is collected by asking each participant to listen to each sound stimulus and move as if they were producing the sound using a musical instrument they are given. Multimodal data is recorded during each performance, and paired with the synthesis parameters used for generating the sound stimulus. The dataset created using this method is then used to build a topological representation of the performance movements of the subjects. This representation is then used to interactively generate training data for machine learning algorithms, and define mappings for real-time performance. To better illustrate each step of the procedure, we describe an implementation involving clarinet, motion capture, wearable sensor armbands, and waveguide synthesis.

Author Keywords

Mapping, Sonic Interaction Design, Human-centered Machine Learning, Multimodal Data

ACM Classification

H.5.5 [Information Interfaces and Presentation] Sound and Music Computing — Methodologies and Techniques
H.5.2 User Interfaces — User-centered design

1. OVERVIEW

This method is a tool for sound designers, composers, and performers. It allows them to use data describing action-sound couplings collected from a group of people to generate multiple complex mappings for sound interaction. In other words, this is a technique for *sampling music-related embodied knowledge* from a group of people and designing sonic interactions based on this information.

The structure of the method is outlined in Fig. 1. The main steps of the procedure are:

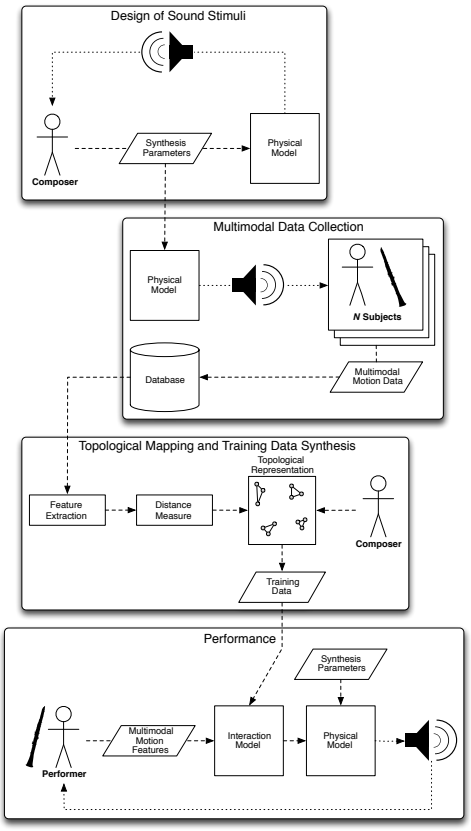


Figure 1: Method structure and workflow.

- Design the sound stimuli using the synthesis engine that will be employed in performance.
- Present the sound stimuli to the group of participants, asking them to move as if they were producing the sounds with the instrument they are given and collect multimodal data during each performance.
- Extract features from the multimodal data and define a topological representation of the performances.
- Select a point in the topology to generate the corresponding data and train a machine learning model for real-time interaction with the generated data and the synthesis parameters used for producing the sound stimulus.
- Use the resulting mapping for composition and performance.

The last two steps can be reiterated to generate new mappings from the same dataset.

NIME'17, May 15-19, 2017, Aalborg University Copenhagen, Denmark.

A salient feature of the method is that data can be obtained from a vast a group of people and – at the same time – individual idiosyncrasies and commonalities are preserved and mapped. This information can then be used for interaction design.

We will describe the steps of the procedure individually, both in general terms and by illustrating how the method was implemented using a clarinet, a synthesis engine based on a flute physical model, motion capture, and armbands with EMG and IMU/MARG sensors.

2. BACKGROUND AND MOTIVATION

The main motivation behind this method is to make use of the music-related movement knowledge of a group of people to define motion-sound mappings for live interaction. The resulting mappings take advantage of the *ecological knowledge of action-sound couplings* of the group of people that participated to the multimodal motion data collection.

A topological representation of the motion data aims at providing an interpretation of what is shared and what is idiosyncratic among the participants, thus allowing researchers and performers to take into account commonalities and individualities when generating the training data for the machine learning model. The interactive method based on this representation allows training data to be generated that either preserves certain peculiarities of individual subjects, or is based on features shared by many participants. This gives more control over the transparency and intuitiveness of the resulting movement-sound mapping. This is particularly desirable for expressive applications such as music performance, where idiosyncrasies and non-obvious mappings could be deliberately employed for expressive purposes.

From a theoretical point of view, a central assumption is that studying the relationships between gestures and sound help us understand how movement contributes to structure our experience of music [12] and that the ecological knowledge about action-sound couplings guides our perception of artificially created action-sound relationships [15]. Thus, the design of action-sound mappings would benefit from action-sound couplings that have similar properties and are part of this ecological knowledge.

This method employs procedures and techniques for data collection and analysis analogous to those typically adopted in experiments of music cognition and systematic musicology. An example is the study by Godøy et al. [13], who analysed video recordings of people mimicking piano-playing movements while listening to musical excerpts. This was done to explore the ability of listeners with different musical backgrounds to reproduce the geometry and the dynamics of movements related to piano performance. Teixeira et al. [25] employed similar techniques to evaluate the gesture consistency of a group of clarinetists over several performances. In our case, the experimental procedures serve a different purpose, namely the creation of gesture-to-sound mapping informed by action-sound couplings. The main goal of this method is to obtain data describing how a group of people associate instrumental movements to certain musical sounds and observe shared and individual features of this movements. This data can be seen as *a snapshot of the ecological knowledge* that a group of individuals has of certain action-sound couplings related to the musical instrument they are 'performing' the sound stimuli with. Rather than performing statistical analyses aimed at corroborating a hypothesis or exploring recurrent patterns in music-related movement, here the data is used for defining movement-sound mappings that take advantage of specific information regarding how a group of people embody clarinet performance movements. Unlike the research carried out by Godøy et al. [13], non-experts also use a real instrument. This is done in order to obtain movements that are constrained by the physical affordances of the clarinet.

Still, the purpose of this method is not to substitute interaction design choices based on intuition with decisions informed by quantitative data. Rather, these techniques are aimed at providing a method for interpreting and utilising movement information for musical interaction. In fact, this approach may also be used in conjunction with other mapping strategies. To attempt an analogy with common electronic music production techniques, this method can be seen as a way of *sampling movement knowledge* from a group of individuals. This information can then be manipulated and repurposed, as it is common practice with audio samples. The size and composition of the group this information is sampled from can also be considered a factor that can be deliberately manipulated by the composer. For example, one might be interested in working with movement data collected from a small group of individuals of specific ethnicity, gender, age group, etc. as opposed to analysing large databases that describe movements of a vast population. This method aims at preserving and exploring variability, and is influenced by human-centred approaches to machine learning [10]. The following sections describe how to implement the method, which is informed by interdisciplinary scientific studies and designed to serve artistic and interaction design purposes.

3. DESIGN OF SOUND STIMULI

The set of sound stimuli to present to the participants during data collection is designed by the composer/performer by recording and editing sequences of synthesis parameters. Playing back these sequences allow the physical model to generate the desired sound.

In this implementation, the sound stimuli were designed and synthesised using a flute model algorithm based on waveguide synthesis [24, 23]. It was originally designed by Bessell [1] and used in his piece *Ophidian*. Parameter sequences were deliberately designed to obtain some examples of various articulations that can be achieved using the flute model. It should be noted that obtaining sounds that closely resembled those produced by a real flute was not our goal. Rather, we aimed at obtaining sounds that preserve some timbral qualities of wind instruments but go beyond what a conventional wind instrument can achieve in terms of pitch, tone, and dynamics.

The sound stimuli were synthesised in Max. Nineteen parameters of the physical model were exposed for control, therefore each sound sequence is made up of nineteen parameter envelopes. The audio output of the physical model and all the synthesis parameters were recorded in separate tracks of a MuBu container [11] and then saved as SDIF files. Doing so allows re-synthesis of the stimuli in real time using the parameter data and storing recorded audio for reference. Six stimuli of length between 10 and 28 seconds were eventually selected.

It is worth mentioning that different strategies can be adopted to record the synthesis parameters of the stimuli: from manually designing each parameter envelope with a graphical editor to using other controllers to perform and record parameter modulations.

4. MULTIMODAL DATA COLLECTION: MATERIAL AND METHODS

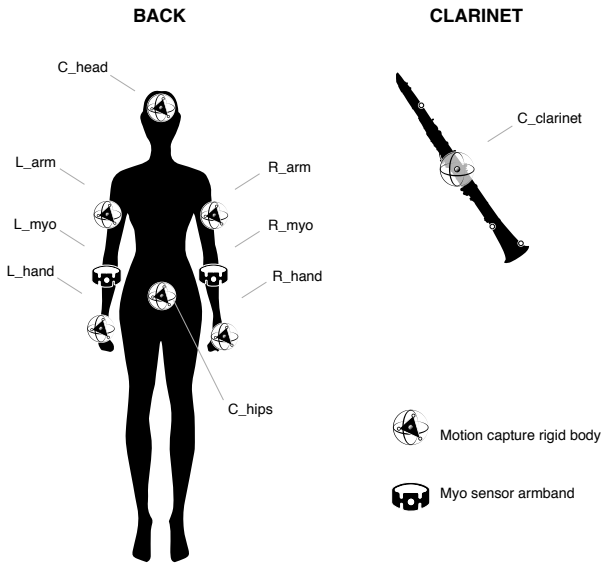


Figure 2: Multimodal configuration: locations of the rigid bodies and the Myo sensor armbands. The clarinet was fitted with three single markers and then defined as a single rigid body with the pivot point at the centre.

Once the parameter sequences for the synthesis of the stimuli are defined, individuals from a group are asked to *mime a performance* of each sound using the musical instrument of choice. Motion, IMU, and EMG data is recorded during each trial. This produces a multimodal database containing data describing the performance movements each participant associated to each stimulus aligned to the synthesis parameters used to generate the sound.

4.1 Apparatus

The movement of the participants were recorded using a multimodal set-up involving a six-camera optical motion capture system (OptiTrack Flex 3) and two Myo sensor armbands. The motion capture system was used to track seven rigid bodies, each one constituted by three or four reflective markers. The locations of the rigid bodies were as follows: head, left upper arm, right upper arm, left hand, right hand, sacral wand (hips), and clarinet (see Fig. 2).

The 6 Degrees of Freedom (6DoF) data (3D position coordinates and orientation quaternions) was streamed to Max via Open Sound Control (OSC) using a custom MATLAB script. The participants also wore two Myo armbands, one on each forearm. The devices streamed IMU/MARG 9DoF data and EMG data over a dedicated OSC port. The IMU data is constituted by 3D acceleration, 3D angular velocity, and 3D orientation Euler angles, while the EMG data has eight channels per armband. The sound stimuli were re-synthesised in real time during the recording session using the previously recorded parameters and were played back via a pair of Genelec 8020C loudspeakers.

4.2 Task and Data Collection Procedure

Each participant was informed about the purpose of the data collection and asked to wear the Myo armbands and the rigid body markers as described in section 4.1.

The participant was then given a clarinet fitted with three reflective markers and with the reed removed. The embouchure of the clarinet was protected with a layer of food grade cling film, which was replaced after every individual session. This was done for hygiene purposes and in order to allow each participant to comfortably use the embouchure.

For each of the six stimuli, each participant was first asked to carefully listen to the sound and imagine the movements they would do if they were to perform that sound using the clarinet. This listening phase could be repeated however many times the participant wanted. The participants were allowed to rehearse the movements while listening to the sound in order to find the movements and actions that, in their opinion, best matched the idea of performing that sound using the clarinet. After having sufficiently familiarised with the sound and decided the movements, the participant was asked to mimic a performance along the sound for three times. Participants were also instructed to perform each stimulus consistently (i.e. trying to perform the same performance movements they devised during the listening phase the best way they can throughout the three takes). In order to help the participant to start the performance synchronised with the sound, each stimulus playback was introduced by a four-beat count in at 120 bpm tempo. During each take, all the data from the rigid bodies motion tracking and the EMG and IMU data of the Myo armbands were recorded in a single, multitrack MuBu container in Max. The synthesis parameters and audio of each stimulus were also recorded in the same container synchronously, and so was the click track that produced the count in before the stimulus. Sensor and control parameters were sampled at 50 Hz, while audio at 44.1 kHz. All the performances were also filmed using a Canon DSLR.

4.3 Participants

Eight participants took part in the data collection phase of this project (7 male, 1 female, aged 24-53, average age: 31, SD of age: 9.0), which took place at the ICCMR Studio, Plymouth University, in June 2016. All the participants had some musical background.

It this important to point out that in this context whether the sample is representative of a larger population is not of major concern. In fact, as pointed out in section 2, selecting a biased sample could be done deliberately in order to obtain datasets that yield peculiar mappings.

5. DATA PROCESSING AND DISTANCE MEASURE

Following the data collection phase, the content of the multimodal database is processed and analysed. This section describes the steps necessary to represent the performance movements for each sound stimulus as a point in a feature space. This is done by selecting and extracting motion features from the data and adopting a distance measure to locate each trajectory on the feature space. The distance relationships between all the movements the participants performed in response to a single stimulus define a map: a gestural topology of the movement reactions to that sound. This representation is useful to understand differences and similarities between each performance in relation to the selected features. Points clusters in regions of the feature space would indicate that a group of participants have performed similar movements, whereas outliers might suggest a more idiosyncratic performance.

5.1 Feature Selection and Extraction

We then selected the locations of the body and the motion features on which we wanted to focus on for creating the mappings for real time performance. In this implementation, the features selected are the Quantity of Motion of the clarinet rigid body and the envelope of the mean absolute value of the right arm EMG data.

Quantity of Motion (QoM) is a motion feature widely

used in the study of body movement in music [14] and it is also employed for detecting affective states and emotion [20]. Fenza et al. [5] define Quantity of Motion (QoM) as the sum of Euclidean distances between successive points in a time window. Rotation angles can also be considered in order to obtain a descriptor that takes advantage of 6DoF information:

$$6DoFQoM(t) = \sum_{k=0}^{N-1} \beta_1 \mid \|q_{t-k}\| - \|q_{t-k-1}\| \mid + \\ +\beta_2 \mid \|p_{t-k}\| - \|p_{t-k-1}\| \mid \quad (1)$$

where $\|q\|$ is the norm of the orientation quaternion, and β_1 and β_2 are weights to balance the contributions of translational and rotational motion data. The values for each frame are summed over a time window of length N samples.

The EMG feature used in this implementation is the mean of the absolute values of all the eight EMG channels of the armband. This value can be considered an index of the overall muscular activity in the forearm. It is calculated as follows:

$$MAV = 1/N \sum_{k=1}^{N} |X_k|. \quad (2)$$

X is the vector with the EMG data and N is equals to the size of X, which in this case is equal to 8 since the Myo has eight EMG channels.

5.2 Distance Measure and Feature Space

A distance measure needs to be adopted in order to locate each performance in a two-dimensional feature space defined by the selected features. We used Dynamic Time Warping (DTW) to measure the distance between the feature vectors of each take. DTW returns the smallest distance between trajectories if warped, therefore it accounts for the fact that sequences might shift slightly forward or backward in time. It is widely used for time series analysis and classification tasks [22, 21] and for real-time gesture recognition [9].

As described in section 4.2 above, the participants performed along each stimulus three times. Fig. 3a shows the locations in the feature space of all the performances along stimulus 5. The locations of the points were obtained by placing the mean of all the performances at origin of the axes and using DTW to calculate the distances of each point from the mean. In Fig. 3a, the trials performed by the same participant are displayed with the same colour. By connecting the respective takes and filling the resulting triangular area we obtain a visual representation of the consistency of each participant across the three trials. The circle inside each triangle is the centroid obtained by averaging the locations of the three performances. Areas of the feature space with higher concentration of points suggest shorter distances between participants and therefore more similarity in the selected features.

6. TOPOLOGICAL REPRESENTATION AND SYNTHESIS OF TRAINING DATA

Van Nort et al. [26] describe the topological perspectives of adopting a holistic conceptual approach to mapping between control and sound parameters. In particular, they focus on "functional properties related to a mapping's geometric and topological structure in the case of continuous, many-to-many mappings". Rather than focusing on the interconnections between individual parameters, a functional view of parameter mapping is concerned with the structural properties of the set of input and output parameters. These properties determine a *mapping topology*, which defines "the nature of the continuity, connectedness, and boundary definition in the mapping association (or associations) between control and sound sets." [26, p. 7].

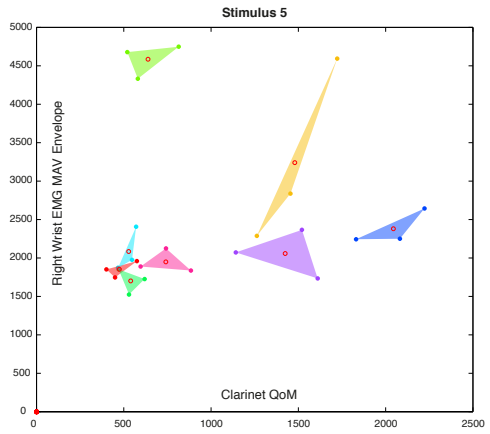
In this case, the topology defined using the distance measure described in the previous section is used to synthesise the training data for the machine learning model that will be used for real-time interaction. In order to represent each participant as a single point in the feature space, the features describing the three performances along the stimulus were averaged. The resulting time series and corresponding distances in the feature space are exported from MATLAB and loaded in Max.

Using the radial basis function interpolation [8] Max tool RBFI, each participant is represented as the centre of a Gaussian kernel [18] in the feature space. The largest graph on the top left corner of Fig. 3b shows the Gaussian kernels obtained from the same data used to plot the triangles shown in Fig. 3a. Each point in this feature space corresponds to a temporal feature set obtained by continuously interpolating between the features extracted from the performances of the participants. This is done by using the distances from the selected point to calculate the contribution (weight) of the feature set of each participant at that point of the feature space. The two graphs on the left (blue and red) show the interpolated features corresponding to the point in the feature space selected using the cursor (white cross).

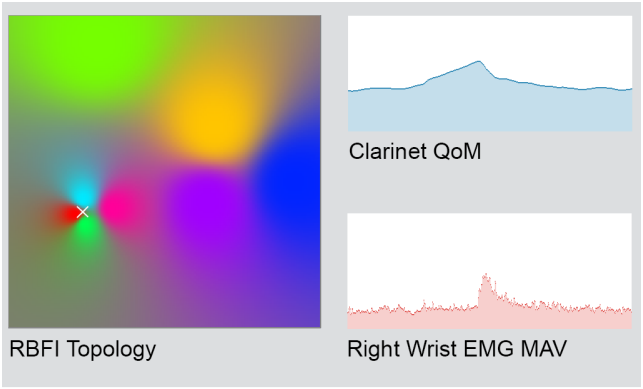
From a practical point of view, this interactive display of the data allows to intuitively create a feature set that can be used in conjunction with the synthesis parameters of the stimulus to train a machine learning model for real-time interaction. Displaying the data of each participant as a location in a feature space has the purpose of communicating certain topological qualities of the data. The distance relationships between the Gaussian kernels give higher-level information about how the participants moved to the sound that is useful for defining the interaction. For example, placing the cursor close to a cluster with several participants would result in training data that is closer to how the participants in that cluster performed along the stimulus. From this perspective, clusters can be considered as different 'styles' of performance movements along the same sound stimulus. Positioning the cursor away from clusters and closer to outliers would instead result in training data that is more *idiosyncratic*: representative of how a single individual reacted to and performed along the sound stimulus. Clusters – on the other hand – suggest that a group of participants performed the sound stimulus with a movement that has certain *shared* features. Ostensibly, moving closer to a highly-populated cluster would result in training data that would produce mappings that are more transparent and intuitive for the population of that cluster. Conversely, data closer to less populated areas and outliers would instead lead to less predictable interactions. However, the characteristics of the resulting mappings also depends on the chosen features and the machine learning algorithm used.

7. INTERACTION MODEL AND PERFORMANCE

The data generated by selecting a point in the feature space paired with the synthesis parameters of the corresponding sound stimulus can then be used to train a machine learning model for real-time interaction. Various algorithms can be adopted to do this. Some of the established supervised learning methods that make use of multiple examples to



(a)



(b)

Figure 3: (a) Two-dimensional topological representation of the performances along stimulus 5. The vertices of each triangle correspond to the three performances of each participants. (b) Screenshot of the user interface in Max. Each Gaussian kernel in the RBFI graph (left) corresponds to a participant. The location of the kernels corresponds to the location of the triangles of the same colour in Fig. 3a. The white cross is the cursor used to select a point in the feature space. The two graphs on the right side show the interpolated features corresponding to the selected point.

recognise gesture classes are based on Hidden Markov Models (HMM) [16] and Support Vector Machines (SVM) [19]. In particular, hybrid methods based on gesture templates and statistical recognition are employed for real-time continuous control using a limited number of training samples [2].

Machine learning is also widely used for implementing interactive approaches to gesture-sound mapping. Fiebrink et al. [6] use supervised learning to build a training dataset from the gestures users perform along a musical score. Caramiaux et al. [3] use a perception-action loop as a design principle for gesture-sound mapping in digital musical instruments. Françoise et al. [7] employ HMM to conjointly model control and synthesis parameters.

In this instance, we used the Max implementation of the Gesture Variation Follower (GVF)[1] [4]. The training data obtained with the procedure described in the previous section is used as a gesture template for GVF. In performance, the two features used for analysing the movements of the participants are calculated in real time and fed to GVF. The synthesis parameters that produced the stimulus the participants performed along to are loaded in the physical modelling patch. During performance, the GVF continuously outputs the temporal alignment with the gesture template. This information is used to move through the temporal dimension of the synthesis parameters of the sound stimulus, thus mapping the movements of the performer to sounds generated by the physical model.

For this implementation, GVF was chosen over other algorithms because – by modelling the temporal information of the gesture template – it is able to detect when the movement described in the template is performed backwards. Moreover, it allows for continuous interaction with the physical model without needing to define the beginning and end of a gesture. These characteristics allow us to obtain different sounds and articulations by interacting with the synthesis parameter space defined by the sound stimulus. Gesture-sound mapping can be easily redefined by repositioning the cursor on another region of the topological representation and generate new training data. This allows to interactively explore the different mappings that the feature space defined by the movement of the participants affords.

[1] `https://github.com/bcaramiaux/ofxGVF`

The sound palette can be expanded beyond what can be achieved using the parameters of a single sound stimulus. This is done by repeating the procedure described in this and the previous sections for the other stimuli, thus generating additional templates for GVF paired with the synthesis parameters of the corresponding stimuli.

In performance, the amount of rigid bodies and sensor armbands worn by the performer can be reduced to the ones that are necessary for extracting the selected motion features in real time. This results in a less cumbersome performance setup. For performing with the two features selected in this example, only the right Myo armband and the clarinet's rigid body markers would be necessary. However, the data from the other rigid bodies and sensors is not superfluous, as it can be stored and used for other compositions based on the same instrument family and for generating other mappings based on other motion features and locations of the body.

8. DISCUSSION

The motion features selected for the clarinet example are relatively simple. Using more sophisticated features and more complex synthesis engines could in principle lead to more complex motion-sound interactions. The same procedure can be applied using features that involve more body locations or full-body motion descriptors. This implementation is limited to two features also for usability purposes. This allows for a clear representation of the performances in a two-dimensional feature space and a simple graphical user interface can be employed to select a point in the topology and generate training data. Working with three or more features is also possible, in which case a different way of presenting the topology of the performance data should be adopted. A 3D representation is certainly a straightforward solution but more complex, multidimensional relations could be represented using alternative methods such as topological networks [17]. We have described an instance were marker-based motion capture is used. However, the method is hardware agnostic, and can be implemented also using cheaper and more portable sensor technologies.

If movement affects our experience of music and therefore can be used as a musical feature, data describing music-related movement contains information that can aid musical

composition and performance. This method was designed to harness multimodal data sets to generate sonic interactions based on shared embodied knowledge.

9. ADDITIONAL AUTHORS

Additional authors: Eduardo Miranda (Plymouth University, UK, EU, email: eduardo.miranda@plymouth.ac.uk).

10. REFERENCES

[1] D. Bessell. Ophidian and the Uncanny Valley. *International Computer Music Conference Proceedings*, 2011, 2011.

[2] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana. Continuous realtime gesture following and recognition. In *In Embodied Communication and Human-Computer Interaction, volume 5934 of Lecture Notes in Computer Science*, GW'09, pages 73–84. Springer Verlag, Berlin, Heidelberg, 2010.

[3] B. Caramiaux, J. Françoise, N. Schnell, and F. Bevilacqua. Mapping Through Listening. *Computer Music Journal*, 38(3):34–48, sep 2014.

[4] B. Caramiaux, N. Montecchio, A. Tanaka, and F. Bevilacqua. Adaptive Gesture Recognition with Variation Estimation for Interactive Systems. *ACM Transactions on Interactive Intelligent Systems*, 4(4):1–34, dec 2014.

[5] D. Fenza, L. Mion, S. Canazza, and A. Rodà. Physical movement and musical gestures: a multilevel mapping strategy. In *Proceedings of Sound and Music Computing Conference*, Salerno, Italy, 2005.

[6] R. Fiebrink, P. R. Cook, and D. Trueman. Play-Along Mapping of Musical Controllers. In *Icmc 2009*, pages 61–64, 2009.

[7] J. Françoise, N. Schnell, and F. Bevilacqua. A multimodal probabilistic model for gesture-based control of sound synthesis. *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, pages 705–708, 2013.

[8] A. Freed. Visualizations and Interaction Strategies for Hybridization Interfaces. In K. Beilharz, B. Bongers, A. Johnston, and S. Ferguson, editors, *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 343–347, Sydney, Australia, 2010.

[9] N. Gillian, B. Knapp, and S. O'Modhrain. Recognition Of Multivariate Temporal Musical Gestures Using N-Dimensional Dynamic Time Warping. In A. R. Jensenius, A. Tveit, R. I. Godoy, and D. Overholt, editors, *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 337–342, Oslo, Norway, 2011.

[10] M. Gillies, R. Fiebrink, A. Tanaka, J. Garcia, F. Bevilacqua, F. Nunnari, A. Heloir, W. Mackay, S. Amershi, B. Lee, N. D'Alessandro, J. Tilmanne, T. Kulesza, and B. Caramiaux. Human-Centered Machine Learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3558–3565, 2016.

[11] D. Glowinski, G. Gnecco, S. Piana, and A. Camurri. Expressive Non-verbal Interaction in String Quartet. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 0:233–238, 2013.

[12] R. I. Godøy. Gestural Affordances of Musical Sound. In R. I. Godøy and M. Leman, editors, *Musical gestures: Sound, movement, and meaning*. Routledge, 2010.

[13] R. I. Godøy, E. Haga, and A. R. Jensenius. Playing 'Air Instruments': Mimicry of Sound-producing Gestures by Novices and Experts. In S. Gibet, N. Courty, and J.-F. Kamp, editors, *Gesture in Human-Computer Interaction and Simulation, 6th International Gesture Workshop*, volume 3881 of *LNAI*, pages 256–267. Springer, Berlin Heidelberg, 2006.

[14] R. I. Godøy, E. Haga, and A. R. Jensenius. Playing " Air Instruments ": Mimicry of Sound-Producing Gestures by Novices and Experts. pages 256–267, 2006.

[15] A. R. Jensenius. *Action-sound: Developing methods and tools to study music-related body movement*. Ph.d. thesis, Universitetet i Oslo, 2007.

[16] F. B. Jules Françoise, Baptiste Caramiaux. A hierarchical approach for the design of gesture-to-sound mappings. *Smc'12*, pages 1 – 8, 2012.

[17] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, and G. Carlsson. Extracting insights from the shape of complex data using topology. *Scientific reports*, 3:1236, feb 2013.

[18] A. Momeni and D. Wessel. Characterizing and Controlling Musical Material Intuitively with Geometric Models. In M. M. Wanderley, R. McKenzie, and L. Ostiguy, editors, *Proceedings of the 2003 Conference on New Interfaces for Musical Expression (NIME-03)*, pages 54–62, Montreal, QC, Canada, 2003.

[19] K. Nymoen, K. Glette, S. A. Skogstad, J. Torresen, and A. R. Jensenius. Searching for Cross-Individual Relationships between Sound and Movement Features using an SVM Classifier. In K. Beilharz, B. Bongers, A. Johnston, and S. Ferguson, editors, *Proceedings of the 2010 Conference on New Interfaces for Musical Expression (NIME 2010)*, number Nime, pages 259–262, Sydney, Australia, 2010.

[20] S. Piana, A. Staglianò, A. Camurri, and F. Odone. A set of full-body movement features for emotion recognition to help children affected by autism spectrum condition. In *IDGEI International Workshop*, Crete, Greece, 2013.

[21] S. Salvador and P. Chan. FastDTW : Toward Accurate Dynamic Time Warping in Linear Time and Space. *Intelligent Data Analysis*, 11:561–580, 2007.

[22] P. Senin. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, pages 1–23, 2008.

[23] J. Smith. Physical Modeling Synthesis Update. *Computer Music Journal*, 20(2):44–56, 1996.

[24] J. O. Smith. Physical modeling using digital waveguides. *Computer Music journal*, 16(4):74–91, 1992.

[25] E. C. Teixeira, M. a. Loureiro, M. M. Wanderley, and H. C. Yehia. Motion Analysis of Clarinet Performers. *Journal of New Music Research*, 44(2):97–111, apr 2015.

[26] D. Van Nort, M. M. Wanderley, and P. Depalle. Mapping Control Structures for Sound Synthesis: Functional and Topological Perspectives. *Computer Music Journal*, 38(3):6–22, sep 2014.