HAL
open science

Towards the design of a natural user interface for performing and learning musical gestures

Edgar Hemery, Sotiris Manitsaris, Fabien Moutarde, Christina Volioti, Athanasios Manitsaris

To cite this version:

Edgar Hemery, Sotiris Manitsaris, Fabien Moutarde, Christina Volioti, Athanasios Manitsaris. Towards the design of a natural user interface for performing and learning musical gestures. Procedia Manufacturing, 2015, 3, pp.6329-6336. 10.1016/j.promfg.2015.07.952 . hal-01256374

HAL Id: hal-01256374

https://minesparis-psl.hal.science/hal-01256374v1

Submitted on 14 Jan 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Manufacturing 3 (2015) 6329 – 6336

Procedia MANUFACTURING

6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015

Towards the design of a natural user interface for performing and learning musical gestures

Edgar Hemery[a,*], Sotiris Manitsaris[a], Fabien Moutarde[a], Christina Volioti[b], Athanasios Manitsaris[b]

[a]*Centre for Robotics, MINES ParisTech, PSL Research University, France*
[b]*Multimedia Technologies and Computer Graphics Laboratory, Univeristy of Macedonia, Greece*

Abstract

A large variety of musical instruments, either acoustical or digital, are based on a keyboard scheme. Keyboard instruments can produce sounds through acoustic means but they are increasingly used to control digital sound synthesis processes with nowadays music. Interestingly, with all the different possibilities of sonic outcomes, the input remains a musical gesture. In this paper we present the conceptualization of a Natural User Interface (NUI), named the Intangible Musical Instrument (IMI), aiming to support both learning of expert musical gestures and performing music as a unified user experience. The IMI is designed to recognize metaphors of pianistic gestures, focusing on subtle uses of fingers and upper-body. Based on a typology of musical gestures, a gesture vocabulary has been created, hierarchized from basic to complex. These piano-like gestures are finally recognized and transformed into sounds.

© 2015 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of AHFE Conference

Keywords: Human factors; Gesture recognition; Natural user interface; User experience; Musical interface; Interactive design; Ergonomics

* Corresponding author.
E-mail address: edgar.hemery@mines-paristhech.fr

2351-9789 © 2015 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of AHFE Conference
doi:10.1016/j.promfg.2015.07.952

1. Introduction

As far as we can trace back cultural heritage in various art's forms, it has been fundamental to capture, record and reproduce what our senses perceive. Similarly to the eye, a camera fixes series of images and similarly to the ear, the microphone transduces acoustical vibrations to electrical signals. Sensors, emerging from different technological fields allow us to capture "human" information, and help creating bridges between different research fields such as computer vision and music. Here, we investigate specific human expert gestures and attempt to recognize and model them by using several 3D cameras fused into a unique structure, named the *Intangible Musical Instrument* (IMI). Inspired by pianistic technics, we use a typology of "piano-like" gestures, organized hierarchically from basic to complex ones, in order to create our own metaphors on the IMI.

This paper is structured as follows. First, we present related works covering typologies of musical gesture vocabulary, interactive musical systems, gesture to sound mapping techniques and we will introduce some vision-based sensors. Second, we propose our methodological approach for capturing fingers and upper body gestures, keeping in mind musicological and ergonomic considerations. Hence, we will see how our methodology provides, through a gesture to sound mapping, a way to learn expert musical gestures as well as to perform them and compose with them.

2. Related work

2.1. Typology of musical gestures

A preliminary study on musical gestures is necessary to discern what parts of the body are active and what parts are not, in the process of producing sounds with an instrument. For this matter, we based ourselves on Delalande's categorization of effective, accompanist and symbolic gestures [1]. *Effective* gestures – or Instrumental Gestures in Cadoz' lexicon [2] are necessary to mechanically produce a sound. This category can also be split into subdivisions such as *excitation* and *modification* type of sound-producing gesture [3]. An example is the pressing of a key on a keyboard – called *fingering*. *Accompanist* gestures (or sound facilitating in [3]) are not involved directly in the sound production, but are inseparable from it. They can be subdivided in *support, phrasing* and *entrained* gestures. This type of gesture is as related to the imagination as to the effective production of the sound. *Figurative* gestures, also referred as *symbolic* gestures, are not related to any sound producing movements; they only convey a symbolic message. They can also be seen as communicative gestures in a performer-to-performer context (e.g. a band rehearsing) or performer-to-perceiver (e.g. concert). This typology is fundamental to develop a methodological approach for capturing and model musical gestures. This knowledge helps us to build our targets and to know a priori what features we wish to extract and model from vision sensors data.

2.2. Gestural control of sound

Mapping gesture to sound is the procedure by which, one correlates the gesture input data with the sound control parameters. In order to implement the gesture-sound mapping procedure we need first to decide, which gesture characteristics or *features* and sound synthesis variables we are going to use. We present here a quick overview of the essential strategies of mapping, namely *explicit* and *implicit*.

Direct or *explicit* mapping refers to an analytical function that correlates output parameters with input parameters. The explicit mapping (also called *direct* mapping) can create a direct correlation between the fingers and the production of the note. Similarly with bijective functions, there is a one-to-one correspondence between gesture parameters, such as the position of fingertips in one physical dimension, and one characteristic of the sound such as the pitch for instance. [4]

Indirect or *implicit* mapping can be seen as a "black box" between input and output parameters. The desired behavior of this mapping is specified through machine learning algorithms that require a training phase or purposely designed as stochastic. For example, an analysis of gesture features based on Hidden Markov Models (HMM) allows estimating the most likely temporal sequence with respect to a template gesture [5,6,7,8,9]. HMMs capture the temporal structure of gesture and sound and their variations, which occur between multiple repetitions. The

model is also used to predict in real-time the sound control parameters associated with a new gesture – allowing the musician to explore new sounds by moving more freely.

2.3. Natural Interfaces for Musical Expression

Interactive systems allowing performing with body gestures have appeared in the 90's thanks to motion capture sensors, which allow 3D gesture tracking, mapped onto MIDI parameters of sound synthesizers. The first glove transforming hands and gestures into sounds were created for a performance at the Ars Electronica Festival 1991. Ironically called "the Lady's Glove" [10], it was made of a pair of rubber kitchen gloves with five Hall effect transducers glued at the tip of fingers and a magnet on the right hand, varying voltages were fed to a Forth board and converted into MIDI signals. Preceded with the Digital Baton of the MIT Media Lab in 1996, a major progress in musical interfaces came with the inertial sensors such as accelerometers and gyroscopes, placed in contact with body motion capture sensors or held in hand (cf: The USB Virtual Maestro using a WiiRemote [11], MO Musical Objects [12]). Dozens of interesting projects for virtual dance and music environment using motion capture have been presented over the last decade of NIME (New Interfaces for Musical Expression) conferences.

Thanks to more recent technological breakthrough, gestural data can be obtained with computer vision algorithms and depth cameras. Computer vision is a branch of computer science interested in acquiring, processing, analyzing, and understanding data from images and videos. Video tracking systems are ideal for musical performances since they allow freedom in body expression and are not intrusive; as opposed to motion capture devices. Therefore, a musical interface – or instrument – which draws gestural data from vision sensors, feels natural for the user's experience point of view, provided that the gesture to sound mapping is intuitive and has a low latency response.

2.4. Vision based sensors

We present here two types of vision-based sensors, which we used in our research. As this technological field is growing fast, we could not explore all the existing sensors possibilities; however, the sensors we choose are well documented, largely spread, low cost and fit our requirements.

The first type of sensor is the Microsoft Kinect depth camera. Originally created for video gaming purposes, it had an important impact in many other fields such as sensorimotor learning, performing arts and visual arts to name a few. Equipped with a structured light projector, it can track the movement of whole body of individuals in 3D. The first version of the Kinect delivers a fairly accurate tracking of the head, shoulders, elbows and the hands, but not the fingers. For this matter, we are interested in a second type of depth camera, the Leap motion. This camera works with two monochromatic cameras and three infrared LEDs. Thanks to inverse kinematics, it provides an accurate 3D tracking of the hand skeleton, with more than 20 joints positions and velocities per hand. The Leap motion has a lateral field of view of 150°, a vertical field of view of 120°. The effective range extends from approximately 25 mm to 600mm above the camera center (the camera is oriented upwards) [13]. Leap motion is known for being precise at pointing accurately and fast, and to be one of the best commercial sensors for close range use [14].
The Microsoft Kinect, has a 43° vertical field of view, 57° lateral field of view and an effective range from 0.4 to 4 meters in near range mode [15]. It is also one of the rare vision sensor – if not the only – to track people in either standing or sitting poses.

3. Overview of the Intangible Musical Instrument

We present here the design of the instrument along with a short explanation on the elements positions and purposes. There are three sensors: two Leap motions and one Kinect. Once placed on their slots on the IMI, the Leap motions field of view cover the whole surface of the table and a volume above it. The Leap motions are centered in the halved parts of the surface while one Kinect is placed in front and slightly above the prototype table as displayed on figure 1(a) and 2. Additionally, the whole structure can be lifted up of lowered down according to the musician height. The height can also be adjusted so as to play seated (e.g. for learning scenario) or standing up. (e.g. performance context). A space behind the Plexiglas is dedicated to hold a laptop and a Kinect. The body skeleton obtained from the fusion of the three sensors is depicted in figure 2(b). For information, the skeleton fusion intends

to fuse the skeleton coming from the two Leap motions and one Kinect into a single fused skeleton by coupling the palm joints from the different sensors together.

A table made of Plexiglas table is placed approximately 10 cm above the two Leap Motions, where the sensors field of view cover the area best. The whole instrument is articulated around this table, which serves of frame of reference for the fingers. It also constitutes a threshold for the activation of the sound: one triggers sounds by fingering the table's surface.

Therefore, the Plexiglas plate delimits the framework of interaction of the IMI. The gestural interaction is not limited to this 2D surface but to a volume up to 30 cm above the Plexiglas. This tracking space provided by the IMI is colored in grey in figure 2. The tracking space also serves as "bounding box", delimiting the field of view of the sensors in which the data is robust and normalized.

The boundary fixed by the table's surface eases the repetition of a type of gesture. This conclusion arises from the difficulties of gesture repetitions observed in "air-instrument" [16], where the movement is done in an environment with no physical point of reference. In this respect, it is a profitable constraint to add a table since it enables the user to intuitively place his/her hands at the right place and helps repeating similar gestures.

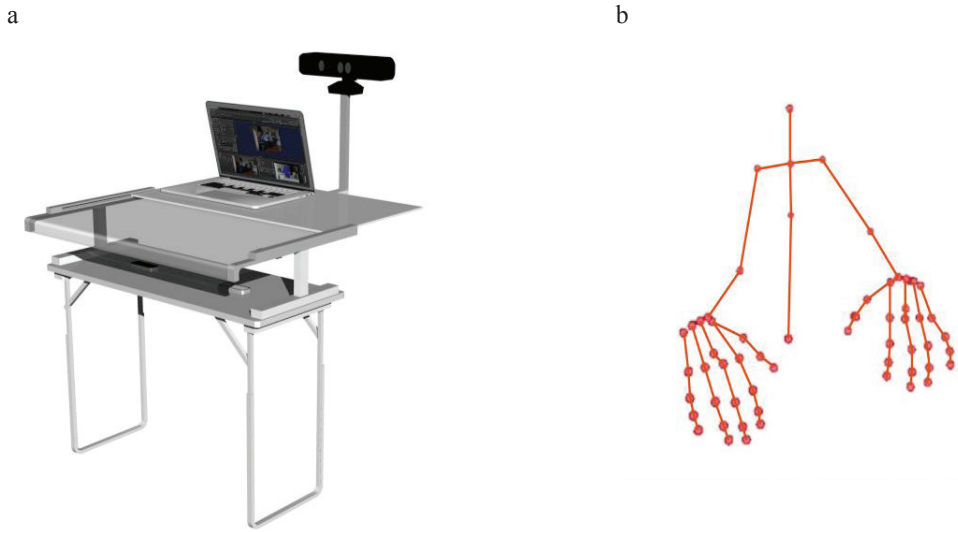

Fig. 1. (a) The Intangible Musical Instrument prototype; (b) Fused skeleton obtained from the IMI.

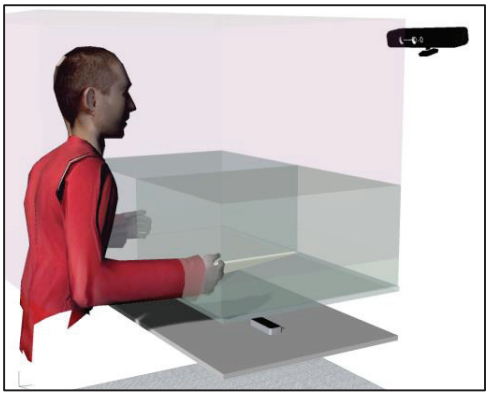

Fig. 2. Interactive surface and space in the Intangible Musical Instrument.

4. Methodology

4.1. Hierarchical metaphors of musical gesture

The primary function of interface is to be used for gesture analysis and extraction of specific gestures. As introduced in section 2.1, there is an existing typology of musical gestures and we use it so as to precise what gestures we wish to extract and model. As we saw, there are many ways to categorize musical gestures. For instance, the *effective* gesture implicitly encompasses many gestural characteristics such as fast, slow, agitated, calm, tense, relaxed to name a few, which have their equivalent in musical terminology (legato, staccato, piano, forte, etc.). This led us to build a hierarchical structure of musical gestures, to help us in the gesture to sound mapping development. This hierarchical structure can be seen in figure 3. Hence, the sensors record and extract gesture features that are organized in low, mid and high level feature as a musician performs on the instrument – following the musical characteristics shown in the pyramid.

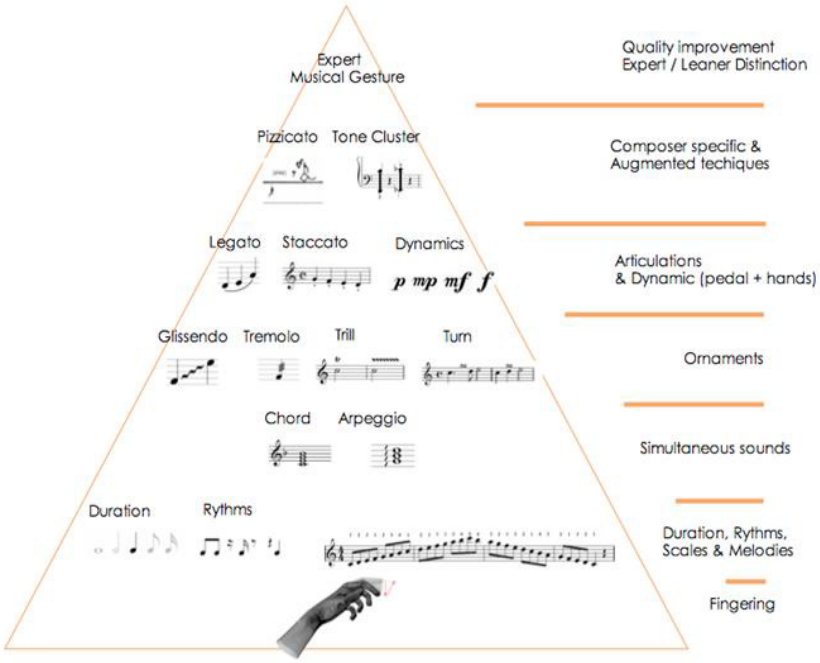

Fig. 3. Hierarchical representation of musical gestures.

4.2. Conceptualizing the musical instrument

The design of the instrument is done according to what the gesture to sound mapping (introduced in section 2.2) allows. The algorithms and heuristics - incorporated in the term “mapping” - are the core of the interface as they set the rules on how the learner/performer plays and what type of sound s/he triggers. The algorithms created to articulate the mappings are built on the hierarchical metaphors that we proposed in section 4.1.

The first heuristic for the design of the surface interface is the division of the table’s surface in several zones, (represented by the colored squares in figure 4). The key idea here is to cover a range of notes, without the need to be extremely precise while fingering on the table since the latter is flat and transparent. Therefore, a zone (either blue, red or green) corresponds to a set of fives notes (e.g.: EFGAB), where each note corresponds to one finger. In a zone, there is no need to move the hand’s position in order to play different notes as long as the palm is above one of the zones. Each finger is tracked and has a fixed ID associated to it. Therefore, when the player touches a zone

with one finger – say right index in "normal zone", s/he will play the note G. If the same finger were in centered zone, it would play the note F. Each hand covers three zones so one player can cover six zones in total, which correspond to almost 3 octaves (from F3 to C5).

The hand's centroid being associated with a 10 cm wide zone, it has a certain flexibility and position tolerance. Having such a system allows the player not to worry too much about fingertips positions on the surface of the table but to focus on other parameters such as the velocity and trajectory prior to contact, which constitute the expressivity of the movement.

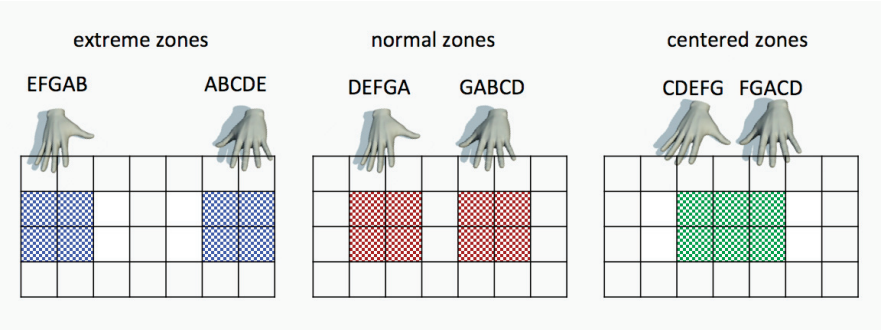

Fig. 4. Table's zones.

4.3. Fingering model for explicit mapping

Secondly, we are interested in building the dynamics, the articulation and the duration metaphors, inherent in the fingering. This led us to the decomposition of the fingering in several phases so as to extract information about the trajectory and the temporality of each part. This representation is in four phases: *Rest, Preparation, Attack & Sustain*, inspired by the PASR (*Preparation, Attack, Sustain, Release*) model [17]. Segmenting the fingering into four essential phases, we observe distinct features for each phase (see figure 5). In rest position, the hand and fingertips are relaxed on the table. In preparation, one of several fingers lift upwards. In attack, one or several fingers lower down at the table's level. In sustain, the fingertip(s) press(es) against the table.

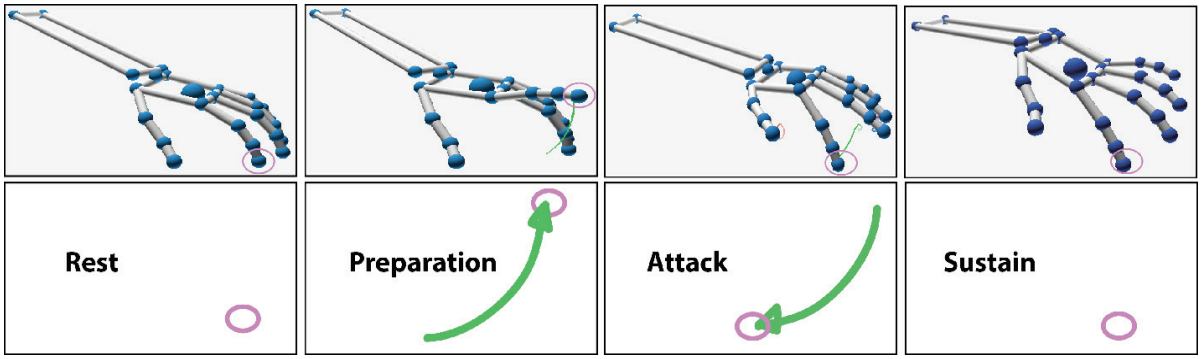

Fig. 5. Rest, Preparation, Attack, Sustain Model.

The RPAS model enables us to model the gesture in details, providing us information on the duration, the trajectories and the speed of each phase. This information can be used in real-time for the mapping. The preparation time (the time spent in preparation phase) along with the attack velocity can be used to express the dynamics of the sound. After some fine-tuning, we modeled a simple logarithmic function that transforms the velocity of the attack to express the dynamics naturally – as one would intuitively expect when pressing a key. It allows the system to be

reactive and sensitive to small velocity variations and to reach an asymptote when the attack gets stronger, acknowledging for to the dynamics limits of the instrument. The sustain time enables the musician to make a note last for a determinate duration. Finally the rest position enables the player to stop the sound. This way, one can play notes with an intended duration and dynamics.

4.4. Implicitly mapping upper body gestures to sound

The algorithm based on implicit mapping replays sound samples at various speeds according to the gesture performed in real-time by the upper body (head, arms and vertebral axis). Audio time stretching/compressing and re-synthesis of audio can be performed using a granular sound synthesis engine. Particularly, it is based on a method called mapping by demonstration or temporal mapping developed by Jules Francoise at the IRCAM [17]. This mapping works by associating a sound to a template gesture and links temporal states of a sound with temporal states of the template gesture. This technique requires a preliminary step in which the expert creates the gesture model by training the system with the expert gestures. The system allows to choose any pre-recorded sound or to produce one and to bind the gesture model to it.

5. Implementation and use-cases

5.1. Performing and composing music with gestures

The IMI allows for extraction of specific gestures. As a musician performs on the device, the sensors stream low-level gesture features (3D coordinate data) into our program written in the real-time programming environment Max/MSP, which are then transformed into higher perceptual level features (e.g. dynamics, articulation, duration). These features are finally transformed into sounds via a mapping and the user is free to use different type of sound synthesis engines for the actual sound production. As we created gestural metaphors based on “piano-like” gestures, the system is well adapted to piano sounds. Additionally, the table surface contributes to keep the experience intuitive for anyone who experiences with the IMI – as pianistic gestures are generally well known by the general public.

However, it is important to precise that the IMI is not a virtual replacement of the piano, but opens the way for the adaptation of keyboard instruments paradigm to the a new digital area, keeping the physical gestures inherent in music practice at an expert and natural level. The variety of sounds the instrument can produce is equivalent to most synthesizers, but the way the musician interacts with it is totally unique, which makes the interface a powerful tool for both performing and composing electronic music.

5.2. Learning musical gestures

In a learning scenario, while the learner performs the expert gestures, s/he attempts to get close enough to the gesture model so as the sound is played back at its original speed. The resulting sound is the feedback given to the learner in order to adjust his/her gestures to the expert’s one. In this fashion, a beginner can learn pianistic technics.

Another option of the system is to “augment” musical scores – constituting a Tangible Musical Heritage (TCH) – providing visual annotations on the expert gestures, showing apprentices how to move their fingers, arms and shoulders and how to perform expressively. These indications can be displayed on a screen in addition to the sonic feedback resulting from gesture-to-sound mapping. The research in the future will focus on the design of the augmented music score and its visualization in the 3D platform.

6. Conclusion

In summary, the methodological conceptualization of a Natural User Interface supporting learning, performing and composing with gestures is completed, together with its first version implementation, facilitating, thus, the access and transmission of the musical ICH. The Intangible Musical Instrument is a new type of musical interface,

using computer vision, which frees people to wear or hold any invasive equipment so as not to undermine the expressivity of gestures. The IMI is especially conceived to transmit the multi-layer musical Intangible Cultural Heritage to the general public, allowing people to perform on it and compose music by using gestures. From a technical point of view, the first version of IMI is a unified interface framework for all the three levels: a) musical gesture recognition, b) implicit and explicit mapping sound to gestures, c) sound synthesis.

The IMI has been designed to contribute to the preservation, transmission and renewal of the Intangible Cultural Heritage to next generations in terms of expert gestural knowledge of composers and musicians. Our future work will include further developments of both explicit and implicit mapping, improvements of the design and visual feedback

Acknowledgements

The research leading to these results has partially received funding from the European Union, Seventh Framework Programme (FP7-ICT-2011-9) under grant agreement n° 600676.

References

[1] Delalande, F.,. La gestique de Gould: éléments pour une sémiologie du geste musical G.Guertin. G. Gould, ed., Courteau, Louise. 1988
[2] Cadoz, C. & Wanderley, M., Gesture-music. Trends in gestural control of music, 2000, p.71–94.
[3] R. I. Godøy and M. Leman, Musical Gestures: Sound, Movement and Meaning, Ed., Routledge, 2009.
[4] D. Arfib, J. M. Couturier, L. Kessous, and V. Verfaille. 2002. Strategies of mapping between gesture data and synthesis model parameters using perceptual spaces. Org. Sound 7, 2 ,August 2002, p. 127-144.
[5] Françoise, J., Gesture--sound mapping by demonstration in interactive music systems. Proceedings of the 21st ACM international conference on Multimedia - MM '13, 2013, p. 1051–1054.
[6] Robots and Interactive Multimodal Systems (Springer Tracts in Advanced Robotics, Vol. 74, pp. 127–142). Berlin, Heidelberg: Springer
[7] Françoise, J. & Schnell, N. & Bevilacqua, F. Gesture – based Control of Physical Modeling Sound Synthesis: a Mapping-by-Demonstration Approach. Proceedings of the 21st ACM international conference on Multimedia (MM'13), Barcelona, Spain, pp.447–448. Oct 2013,
[8] Françoise, J., Realtime Segmentation and Recognition of Gestures using Hierarchical Markov Models. Available at: http://articles.ircam.fr/textes/Francoise11a/index.pdf. , 2011
[9] Françoise, J., Schnell, N. & Bevilacqua, F. A multimodal probabilistic model for gesture--based control of sound synthesis. Proceedings of the 21st ACM international conference on Multimedia - MM '13, Barcelona, Spain, (2013) p. 705–708.
[10] Rodgers, T. (2010). Pink Noises: women on electronic music and sound. Duke University Press
[11] Nakra, T.M. et al. The UBS Virtual Maestro : an Interactive Conducting System. 2009
[12] MO Musical Objects | interlude project. Available from: http://interlude.ircam.fr/wordpress/?p=229 [Accessed January 13, 2015].
[13] Anon, Kinect for Windows Sensor Components and Specifications. Available at: https://msdn.microsoft.com/en-us/library/jj131033.aspx [Accessed April 9, 2015].
[14] API Overview — Leap Motion C# SDK v2.2 documentation.
[15] Available at: https://developer.leapmotion.com/documentation/csharp/devguide/Leap_Overview.html [Accessed April 8, 2015].
[16] Stimulant | Depth Sensor Shootout: Kinect, Leap, Intel and Duo.
[17] Available at: http://stimulant.com/depth-sensor-shootout/ [Accessed April 9, 2015].
[18] Rolf Inge Godøy1, Egil Haga1 and Alexander Refsum Jensenius1, "Playing 'Air Instruments': Mimicry of Sound-Producing Gestures by Novices and Experts", Gesture in Human-Computer Interaction and Simulation, University of Oslo, Department of Musicology, Oslo, Norway, ISBN 978-3-540-32624-3. , February 2006
[19] Françoise, J., Caramiaux, B. & Bevilacqua, F. A Hierarchical Approach for the Design of Gesture-to-Sound Mappings. Proceedings of Sound and Music Computing (SMC), Copenhagen, Denmark., 2012