Exploring Stylistics of Expert Gestures and Singing Voice with Motion Capture and Reactive Statistical Mapping

Project Proposal for the eNTERFACE 2014 International Workshop

Principal Investigators

Nicolas d'Alessandro[1], Joëlle Tilmanne[1], Sotiris Manitsaris[2,3,4]

Team Candidates

R. de Charette[2], E. Hemery[2], E. Coupeté[2], C. Volioti[5], A. Glushkova[2,4,5], M. Astrinaki[1], A. Moinet[1], H. Cakmak[1], T. Ravet[1], R. Ben Madhkour[1], S. Laraba[1]

Affiliations

[1]numediart Institute, University of Mons *(Belgium)*, [2]Robotics Lab, MINES ParisTech *(France)*, [3]Realtime Musical Interactions, IRCAM *(France)*, [4]Lab of Rural Space, University of Thessaly *(Greece)*, [5]Multimedia Tech. and Computer Graphics Lab, University of Macedonia *(Greece)*

Abstract

For several years, the need for accurate and real-time gesture following and recognition techniques is increased in the context of natural user interaction. Indeed the development of affordable markerless motion capture techniques – such as depth cameras – has allowed to design new kinds of applications where the continuous tracking and understanding of gestural behaviors and styles may greatly improve usability and user experience. Particularly the fields of intangible heritage preservation and performing arts nowadays require going beyond the simple ideas of recording and explicit mapping. In this project, we aim at prototyping a playing booth for the use case of piano playing. An instrumented and motion-captured keyboard will allow the analysis of expert piano playing in real-time and the recognition of specific gestures, but also of the ongoing *playing style*, i.e. an distance measure with existing teaching schools and emotions. Our real-time stylistic gesture recognition relies on advanced HMM-based techniques that we have exported from the speech processing research. When the player can be positioned in a stylistic space, we aim at using this information to drive a singing synthesizer that also exposes stylistic information, in the form of singing techniques. Our platform will allow the testing of various *stylistic mappings*.

Objectives

The main objective of this eNTERFACE project is the sketching of an original user experience (UX) where a given performer can discover the stylistic dimensions of his/her own expert gestures in real-time and use these detected styles to explore another body of techniques, through a stylistic mapping interface. Particularly, we want to design a new *digital musical instrument* that accurately tracks and recognizes the inherent styles of piano playing gestures in real-time and uses this stylistic information to drive a style-enabled singing synthesizer. This vision has been shaped by many past eNTERFACE projects involving motion capture, real-time gesture recognition and innovative approaches towards statistical generation and mapping [1,2,3]. For this workshop, we decided to focus on three aspects:

Objective 1: Real-time capture of expert piano gestures

Although motion capture technologies exist for quite a while, it is only recently that it has been highly popularized and democratized in many research fields, thanks to the emergence of very inexpensive 3D cameras like the Kinect, the Leap Motion or the PMD CameraBoard. In this project, we want to combine long-range and short-range 3D cameras and extract upper body and hand skeletons from the depth maps in real-time. Skeleton extraction techniques have been mainly validated for functional and gaming motion. Therefore we need to adapt these algorithms to the case of a piano performer sitting at his/her keyboard and executing very specific and sophisticated hand and body gestures.

Objective 2: Real-time stylistic recognition of expert piano gestures

Once a stable upper body plus hands skeleton can be acquired, we want to investigate the real-time recognition of some specific gestures and styles. In this project, the *real-time following* and the *stylistic recognition* are the key factors. Indeed gesture recognition – as a matter of converting a stream of continuous motion data into a series of class labels – is going on for quite a long time [4]. Though, achieving this task in real-time, i.e. as the gesture is being performed, is quite new and only a few pieces of software can handle it, but for very simple use cases. Moreover the inherent styles of those gestures, i.e. the potential meaningful variants in their executions, are often discarded. In this project, we want to explore the idea of *projecting* a given performer in a given *stylistic space*, i.e. track the gestures that are executed and the most likely execution styles in real-time.

Objective 3: Real-time stylistic generation of singing voice

From objectives 1 and 2, we will be able to project a given piano performer into a multi-dimensional stylistic space, e.g. *76% French school*, *54% sad*, etc. Our third objective is to combine this ongoing stylistic information with some low-level gesture data coming directly from the motion capture (e.g. keyboard actions) and to use these modalities in order to drive a real-time stylistic singing synthesizer. Indeed our past eNTERFACE projects have brought the MAGE system to life, a reactive statistical parametric synthesizer. Since 2013, MAGE can work with any kind of modeling data and has a built-in model interpolation framework, which is particularly suitable for stylistic synthesis [3]. Moreover, we have recorded a large singing database, exposing well-known singing styles like *operatic*, *belting*, etc. Therefore we want to train a MAGE singing voice with those styles and see how we can *stylistically map* the recognized piano playing styles with the singing styles.

Background

Over the ten last years, an important amount of motion capture techniques have emerged. However most of these techniques, such as inertial suits[1] or optical markers tracking[2], did remain expensive, cumbersome and often experimental. More recently, the democratization of *depth cameras* – like the Microsoft Kinect – has considerably changed the scope of markerless motion capture research. Indeed the massive dissemination of these sensors gave many new research groups the chance to jump in this field and then provide various resources (databases, software, results) to the scientific community [5]. This technological breakthrough has brought motion capture into new application domains, like health [6] and the arts [7].

Depth cameras stream a *point cloud*, i.e. a collection of 3D points corresponding to the tangible envelope of what is seen by the camera. One key technology to make these point clouds more usable is the fitting of *skeletal models* in order to parameterize the inner structure of human limbs in the 3D scene. Particularly algorithms following Shotton's technique have become the best approach [8]. Shotton's algorithm uses a reversed approach to train its skeleton-fitting algorithm. Indeed it generates a huge collection of available poses for the skeletal model, then generates 3D models of the body envelope for these poses, and finally generates fake point clouds from these 3D models with color-coded points to track the joints. A similar approach has recently been used to track fingers of the hand [9].

The field of gesture recognition is very active for the last 20 years. It has been pushed forward by great progress that was made in speech and handwriting recognition [4]. However the idea of performing the recognition of the gesture in real-time, i.e. before the gesture is actually finished, is much newer. It coincides with the significant shift that happened in interaction design over the last few years, in which natural user interaction has greatly matured and we can now consider more complex functionalities attached to full-body gestures [10]. Among the existing techniques for real-time gesture recognition, we can find descriptive approaches like FUBI [11] or distance-based techniques like short-term Dynamic Time Warping (DTW) [12]. Nevertheless the most promising research direction is statistical modeling, and more particularly Hidden Markov Models (HMMs). IRCAM's *Gesture Follower* (GF) appears to be the state of the art in real-time HMM-based gesture recognition [13]. GF derives a giant HMM from a single utterance of the gesture and compute missing probabilities from a priori knowledge. Then, at runtime, it achieves an ongoing forward accumulation of the probabilities of the model and estimates the ongoing *likelihood*. This technique provides, for any input gesture, a fluctuating likelihood and the estimated time progression in the gesture, for all the captured gestures.

During eNTERFACE 2013, our team has investigated the problematic of gesture recognition from a different viewpoint [3]. Indeed we rely on an exploratory research by Tilmanne *et al.* which aimed at adapting various HMM-based modeling and synthesis techniques from the speech research to motion data [14]. We have explored the use of HTK-trained models (left-to-right context-independent HMMs) and developed specific decoding techniques based on short-term Viterbi algorithms [15]. We also have integrated *continuous stylistic information* in the mapping, by decoding average gestures (which gives better recognition results) and then use the full covariance matrix to retrieve and map the stylistic influence [16].

[1] MetaMotion IGS-190: http://www.metamotion.com/gypsy/gypsy-gyro.htm

[2] NaturalPoint OptiTrack: http://www.naturalpoint.com/optitrack

HMMs have also been greatly used for synthesis, with the work of Tokuda *et al.* in speech [17]. For the last few years, our team has enabled the *reactive synthesis of HMM-based trajectories* for speech, but also for motion data, with MAGE [3].

Technical Description

In this project, we aim at bringing together a new digital musical instrument that extends the straightforward gesture-to-sound mapping with a higher-level approach based on *gesture and sound stylistics*. From the past eNTERFACE workshops and other projects, we have already gathered a significant expertise in new instrument making [2], motion capture [1], statistical modeling of stylistic motion and statistical generation of vocal sounds [3]. The development of this new prototype will consist in consolidating various core functionalities and assembling them in a coherent processing pipeline. Such a consolidation will be benchmarked by *live-testing* our system among a very innovative and demanding use case: a pianist exploring the stylistic space of a virtual singer through the ongoing styles detected in his/her own musical gestures. The architecture of our platform will respect a MVC design:

- Controller: back-end analysis routines to capture body and hand motion;
- Model: middleware for statistical recognition and stylistic mapping tasks;
- View: front-end for rendering the reactive stylistic singing voice.

This part of the proposal gives more details on the different technologies that are envisioned and gives the main research and development axes that will be followed in order to build the new system. We also give greater insights about the devices, environments and prototyping strategies that will be aligned in this project. Finally we also describe the project management that will be deployed.

In this section, most of the following text refers to module names that are depicted in Figure 1. We also highlight the workpackages of the project and introduce a naming convention (WP_N) that will be reused in the section where the schedule is described. Workpackages split the research in homogeneous aspects.

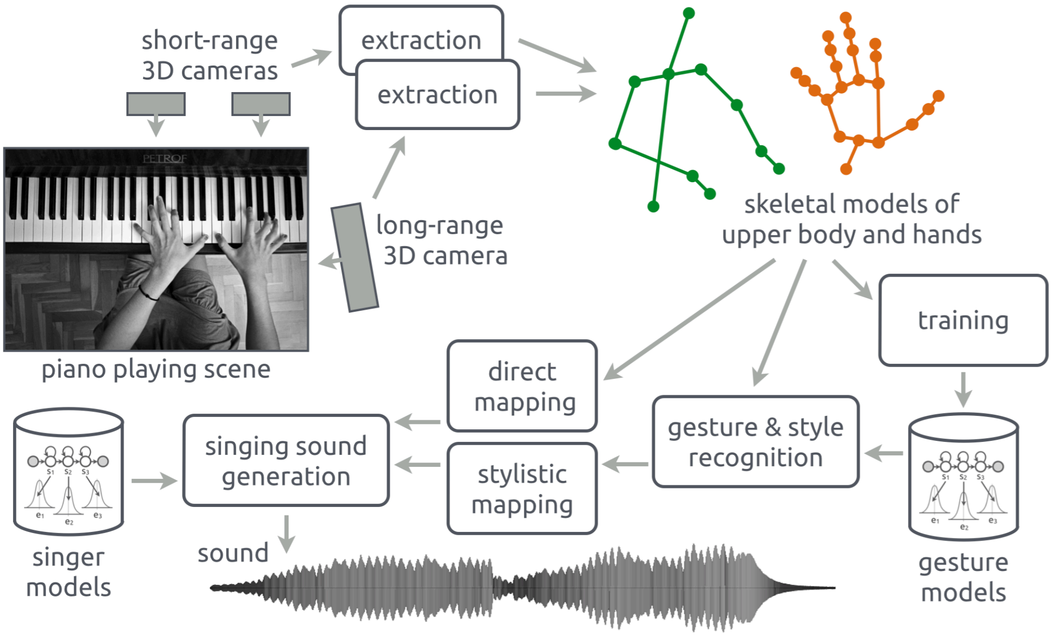

Figure 1 – Processing pipeline for the piano playing use case.

Research and Development Axes

The new digital instrument developed in this project will be composed of five main components, identifying the groups of researchers working on these problems: the extraction of skeletal models based on depth map images, the HMM training algorithm, the real-time recognition of gestures and styles, the development of mapping strategies and finally the reactive generation of stylistic singing sounds:

1. Extraction of Skeletal Models (WP_1)

As described in the Background section, the Shotton's training algorithm for fitting a skeletal model over a depth map is the key technology that enables the use of low-cost 3D cameras for motion capture. However such an algorithm requires a huge amount of training data, i.e. the skeletal models aligned with artificial depth maps reconstructed from 3D envelopes of the human body. These databases are generated *procedurally* for some specific target behaviors, like walking or playing games [8]. Therefore the tracking is far less efficient for other kinds of gestures. Moreover the adaptation of Shotton's technique for the hand postures is still at an early stage.

In this workpackage, we want to improve the existing algorithms for the extraction of skeletal models of the upper body and the hands, specifically to the use case of the piano playing gestures. The strategy of this task relies on the recording of a large synchronized *depth map + high-fidelity mocap* database, prior to the workshop. This database will contain the aligned streams of depth map data coming from short-range (CamBoard) / long-range (Kinect) 3D cameras and data from more accurate systems like the IGS inertial glove / suit or the OptiTrack marker-based tracking. This aligned and annotated contents will help researchers of this group to train various Shotton-like model-fitting algorithms and provide a nuanced comparison of their design choices based on the available high-fidelity ground truth.

2. HMM Training on Motion-Captured Data (WP_2)

It has been clear over the last few years that complex context-based HMM training techniques have exciting applications far beyond the field of speech processing. In our research group, we have particularly explored the use of HMMs for motion modeling. Our fundamental assumption is to consider that there are many similarities between speech and motion, and therefore, the significant collection of statistical modeling tools available in for speech – with well-known toolboxes like HTK[3] and HTS[4] – could advantageously be used with motion data. Particularly it allows us to incorporate the notion of *style* in our modeling approach. A proof of concept has recently been made with the work of Tilmanne *et al.* on stylistic human-like walk synthesis [14].

In this workpackage, we want to extend the current research in stylistic motion modeling to the use case of piano playing. We know that piano playing gestures are subject to a very strong and rigorous *grammar* (or *ontology*), in which some very specific styles can be expressed. We are very interested in piano teaching "schools" – like French or Russian approaches – and we would

[3] HTK Speech Recognition Toolkit: http://htk.eng.cam.ac.uk

[4] HTS HMM-Based Speech Synthesis System: http://hts.sp.nitech.ac.jp

like to know if these influences can be statistically modeled as motion styles, which could be retrieved, visualized and compared later on.

3. Real-time Recognition of Gestures and Styles (WP_3)

The field of gesture recognition is active for quite some time. As described in Background, many approaches have been used towards the classification of motion data based on statistical models. However the idea to infer recognition decisions at the time of the gesture – i.e. not waiting that the gesture is finished to take the decision – seems quite newer and there are not many toolboxes offering this *real-time* feature. Ircam's *Gesture Follower* [13] proposes an efficient, though quite limited, solution based on deriving the state space from one occurrence and accumulating likelihood without decoding. Recently our group has explored various HTK/MLPACK-based short-term Viterbi decoding techniques with encouraging results.

In this workpackage, we want to continue the development of our real-time gesture decoding algorithms based on advanced HTK-based motion models. More specifically, we want to develop an approach towards real-time style recognition. Indeed our gait reconstruction prototype at eNTERFACE 2013 has shown that it was possible to continuously track the constituting styles of a given gesture – in that case, a human step – through full covariance and use this information efficiently to drive synthesis. We want to generalize this approach and develop a style-tracking algorithm for piano gestures.

4. Design of Style-Based Mapping Strategies (WP_4)

The development of a new musical instrument also has to accommodate the question of design. This workpackage carries the question *"what to do with the detected styles and what do they mean?"* Indeed the information that emerges from classification algorithms is often hard to interpret. Before moving to the next step (WP_5), we want to bring this information to the user in a meaningful way. For instance, we speculate that not all the joints of the extracted skeletal models might have the same influence in the achievement of the gesture and we would like to be able to specify such top-down information to our system. We also want to iteratively determine the rules that enable the proper fusion between direct mapping (based on low-level data) and stylistic mapping (based on detected styles).

5. Reactive Stylistic Singing Synthesis (WP_5)

Over the last decade, the field of HMM-based speech synthesis has literally exploded. Our research group has been active in this field for a while, with the particular contribution of proposing real-time and reactive parameter generation algorithms, with a software toolbox called MAGE[5]. During eNTERFACE 2013, this toolbox has been extended so to accommodate any kind of training data. During this workshop, we also recorded a large database of singing encountering various singing styles, as described in The *Complete Vocal Technique*[6] and this database has been annotated.

[5] MAGE Platform for Performative Synthesis: http://numediart.org/mage

[6] Complete Vocal Institute: http://completevocalinstitute.com

In this workpackage, we want to put together the first version of a reactive stylistic singing synthesizer, i.e. a MAGE-based singing synthesizer which gathers two key functionalities. On the one hand, the ability to reproduce intelligible sung phrases from phonetic labels in real-time, as it is now currently possible for speech and motion data. On the other hand, we want to use MAGE reactive model interpolation for applying the composition of singing styles, as they will be mapped (WP_4) from motion styles (WP_3).

Prototyping Cycle

As in any HCI application development, the team workflow is made of iterations between various phases, including research & development (described above) but also updating the case study on which we are working (WP_6) and validating our UX and results in front of a panel of external observers (WP_7). Updating the case studies will consist in constantly revisiting the scenarios on which we are working. The other aspect is the evaluation of the overall UX (piano playing booth) and the system outputs (synthetic sounds) by external listeners/observers. We wish to demonstrate our ongoing prototype to as many eNTERFACE researchers as possible and establish a first informal benchmarking of successful strategies.

Facilities and Equipment

The team will essentially work with available devices brought by the participating labs. Obviously we will bring our own laptops. Moreover, we will try to bring several acquisition systems, Kinects, LeapMotions, PMD CamBoards, and maybe the OptiTrack and the IGS hardware. We will also bring a good MIDI keyboard. The only pieces of equipment that we would eventually require would be several extra speakers and secondary monitors that we might not be able to transport on site. We would also require some space for setting up our piano playing booth.

Project Management

The whole project will be supervised by Nicolas d'Alessandro, Joëlle Tilmanne and Sotiris Manitsaris. They should stay on the site of the workshop for the whole period. Based on the subscribed participants, sub-teams will be gathered around the specific workpackages of the project. The methodology that is promoted in this project aims at staying flexible and adapt to our successive prototyping cycles. We will work with guidelines inspired by various Agile techniques, such as organizing scrum meetings or collectively defining the development backlog.

Project Schedule

In this part we gather the various workpackages that have been highlighted in the technical and set them down on a one-month schedule, plus some extra tasks:

- **WP_1 – Extraction of Skeletal Models**: adaptation of existing skeleton extraction techniques to the case of piano playing, for body and hands
- **WP_2 – HMM Training on Motion-Captured Data**: adaptation of our existing HMM-based training on stylistic motion data to the case of piano playing
- **WP_3 – Real-time Recognition of Gestures and Styles**: development of real-time gestures & styles recognition algorithms based on trained HMMs
- **WP_4 – Design of Style-Based Mapping Strategies**: design of an authoring interface for the creation of specific rules and ontologies in the mapping

- **WP$_5$– Reactive Stylistic Singing Synthesis**: creation of a stylistic singing voice for MAGE with real-time control of the style based on the mapping
- **WP$_6$– Iteration on the Overall Playing UX**: inline reassessment of the digital musical instrument scenarios that we use in our case study
- **WP$_7$– Assessment of UX and Results**: organization of external observation regarding our overall UX and the synthetic results produced by our system
- **WP$_8$– Reporting and Publishing**: general dissemination tasks

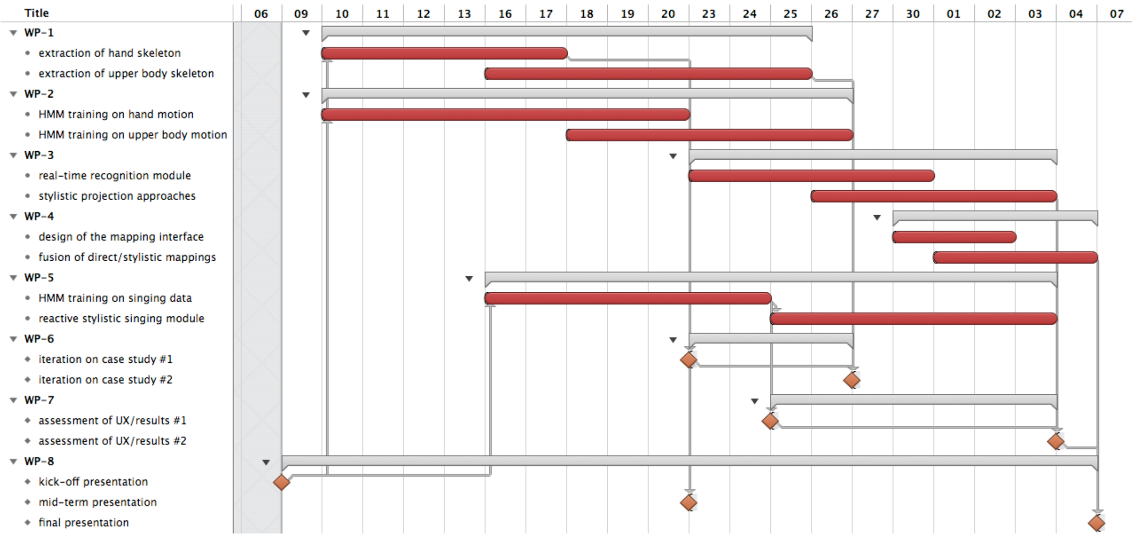

Figure 2 – Scheduling of the project (workpackages and milestones).

Deliverables and Benefits

In this part we describe what are the main deliverables and benefits that the team will provide at the end of the workshop:

- software component for the fitting of skeletal models on depth maps
- database of trained HMMs modeling the hands and upper body gestures
- software component for the recognition of hand and upper body gestures
- database of trained HMMs modeling various styles for the singing voice
- new interactive setup corresponding to the piano-playing use case
- synthesis examples of real-time hand/body-controlled singing
- open sessions during the workshop for welcoming observers
- a scientific report, distributed in the required format

Team Profile

Project leaders: N. d'Alessandro (real-time systems, singing voice, HCI), J. Tilmanne (motion capture, motion statistics), S. Manistaris (computer vision, HCI).

Team proposed: R. de Charette (computer vision), E. Hemery (computer vision), E. Coupeté (computer vision, HCI), C. Volioti (computer vision), A. Glushkova (machine learning), M. Astrinaki (software design, voice), A. Moinet (software design, voice), H. Cakmak (motion capture, machine learning), T. Ravet (software design, motion capture), R. Ben Madhkour (motion capture), S. Laraba (computer vision).

Collaborators that we are looking for: As described in the project schedule, this workshop will need pretty advanced software developers for most of the time. The first half of the month will be more oriented towards data analysis and statistical modeling, as the second half will require expertise in real-time applications. The second half of the project will also involve more testing of the synthesis results and human-computer interaction properties of our software. Therefore for this second half, we are also looking for HCI or Cognitive Sciences profiles.

References

[1] J. Tilmanne *et al.*, *"A Database for Stylistic Human Gait Modeling and Synthesis,"* Proceedings of the eNTERFACE Summer Workshop on Multimodal Interfaces, pp. 91-94, 2008.

[2] M. Astrinaki *et al.*, *"Is This Guitar Talking or What?"* Proceedings of the eNTERFACE Summer Workshop on Multimodal Interfaces, pp. 47-56, 2012.

[3] N. d'Alessandro *et al.*, *"Towards the Sketching of Performative Control with Data,"* [to appear in] Proceedings of the eNTERFACE Summer Workshop on Multimodal Interfaces, 2013.

[4] S. Mitra and T. Acharya, *"Gesture Recognition: A Survey,"* IEEE Trans. on Systems, Man and Cybernetics, C: Applications and Reviews, vol. 37, n° 3, pp. 311-324, 2007.

[5] Z. Zhang, *"Microsoft Kinect Sensor and Its Effects,"* IEEE Multimedia, vol. 19, n° 2, pp. 4-10, 2012, DOI: 10.1109/MMUL.2012.24.

[6] E. E. Stone and M. Skubic, *"Evaluation of an Inexpensive Depth Camera for Passive In-Home Fall Risk Assessment,"* International Conf. on Pervasive Tech. for Healthcare, pp. 71-77, 2011.

[7] Y. Kim, M. Lee, S. Nam and J. Park, *"User Interface of Interactive Media Art in a Stereoscopic Environment,"* Lecture Notes in Computer Science, vol. 8018, pp. 219-227, 2013.

[8] J. Shotton *et al.*, *"Real-Time Human Pose Recognition in Parts from Single Depth Images,"* Communications of the ACM Magazine, vol. 56, n° 1, pp. 116-124, 2013.

[9] A. Dapogny *et al.*, *"Towards a Hand Skeletal Model for Depth Images Applied to Capture Music-Like Finger Gestures,"* Intl. Sym. on Computer Music Multidisciplinary Research, 2013.

[10] Y. Wu and T. S. Huang, *"Vision-Based Gesture Recognition: A Review,"* Springer LNCS: Gesture-Based Comm. In Human-Computer Interaction, vol. 1739, pp. 103-115, 1999.

[11] F. Kistler, B. Endrass, I. Damian, C. Dang and E. André, *"Natural Interaction with Culturally Adaptative Virtual Characters,"* Journal of Multimodal User Interfaces, pp. 1-9, 2008.

[12] S. Dixon, *"Live Tracking of Musical Performances Using Online Time Warping,"* Proceedings of the 8th International Conference on Digital Audio Effects, pp. 1-6, 2005.

[13] F. Bevilacqua *et al.*, *"Continuous Realtime Gesture Following and Recognition,"* Springer LNCS: Gesture in Embodied Communication and HCI, vol. 5934, pp. 73-84, 2010.

[14] J. Tilmanne and T. Dutoit, *"Continuous Control of Style and Style Transitions Through Linear Interpolation in HMM-Based Walk Synthesis,"* Springer LNCS, vol. 7380, pp. 34-54, 2012.

[15] J. Bloit and X. Rodet, *"Short-Term Viterbi for Online HMM Decoding: Evaluation on a Real-Time Phone Task Recognition,"* Proc. of IEEE ICASSP, pp. 2121-2124, 2008.

[16] T. Hueber, G. Bailly and B. Denby, *"Continuous Articulatory-to-Acoustic Mapping Using Phone-Based Trajectory HMM for a Silent Speech Interface,"* Proc. of Interspeech, 2012.

[17] K. Tokuda *et al.*, *"Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis,"* Proc of. IEEE ICASSP, pp. 1315-1318, 2000.