Towards an affective gesture interface for expressive music performance

Vassilios-Fivos A. Maniatakos
RepMus-IMTR
IRCAM-CentrePompidou
1, place Igor Stravinsky
75004 Paris, France
fivos.maniatakos@ircam.fr

Christian Jacquemin
AMI
LIMSI-CNRS & Univ. Paris 11
BP 133
91403 Orsay, France
christian.jacquemin@limsi.fr

ABSTRACT

This paper discusses the use of ‘Pogany’, an affective anthropomorphic interface, for expressive music performance. For this purpose the interface is equipped with a module for gesture analysis: a) in a direct level, in order to conceptualize measures capable of driving continuous musical parameters, b) in an indirect level, in order to capture high-level information arising from ‘meaningful’ gestures. The real-time recognition module for hand gestures and postures is based on Hidden Markov Models (HMMs). After an overview of the interface, we analyze the techniques used for gesture recognition and the decisions taken for mapping gestures with sound synthesis parameters. For the evaluation of the system as an interface for musical expression we made an experiment with real subjects. The results of this experiment are presented and analyzed.

Keywords

affective computing, interactive performance, HMM, gesture recognition, intelligent mapping, affective interface

1. INTRODUCTION

The shift of interest of Human Computer Interaction (HCI) towards emotions and social interaction resulted to intensive studies relative to *Affective Interfaces*: such are the interfaces that appeal to the emotional state of their users and allow to express themselves emotionally, by receiving information of emotional content and decoding it through appropriate techniques. In this work we argue that high-level gesture information can be revealing for the expressive diathesis or emotional state of the user. Furthermore, an interface that succeeds in decoding such information can be inspiring to use as a virtual music instrument. Using strategies 1) to decode high-level gesture information from the user 2) to link this information with its semantic meaning 3) to create intelligent correspondences between these semantics and music synthesis parameters, we have built and evaluated a music performance system for the ‘Pogany’ interface.

2. SCOPE AND MOTIVATION

‘Pogany’ is an affective, anthropomorphic (head), hand-manipulated interface designed at LIMSI laboratory [1]. The purpose of the designers was to offer a new medium of communication that can involve the user in an affective loop. Related work in terms of design and motivation can be found in [2] for a doll interface to emotional expression and in [3] for the analysis of voice expressivity through a hand-puppet interface.

What is investigated in our work is the appropriateness of an interface such as Pogany, apart from human communication, for musical creation and interaction purposes. The deeper scope of this work is to provide the user with a interactive performance system that captures expressivity. For Pogany, such a task sounds challenging a priori, basically for two reasons:

1. The familiarity of a user with the human face, either by view or touch, can help the user associate instrumental gestures for the manipulation of the interface with common hand gestures. Thus, such a music interface provides the performer with facilitated apprenticeship.
2. Particular gesture patterns may correspond to high-level expressive or emotional information. For instance, if we regard a real human face as the interface itself, and we somehow detect the facial expressions produced by the alteration of the face parts (nose, lips, etc), we can then directly have a link between these expressions and corresponding emotions [4]. In our case, visual feedback (with the form of an animated head) helps the user in creating a link between gestures in the vicinity of Pogany and emotions through the intermediate semantic level of face expressions. Apart from this type of indirect association between gestures and emotions, additional emotional information can occur by the type and the particular area of the contact that the user can have with the interface. For instance, when someone touches a face on the cheek, depending on the force used and the speed and suddenness of the gesture, this action could be each time attributed to contradictory emotional intentions: from expression of calmness and tenderness to inelegance and brutality, with an extreme variability such as the one that exists between a caress and a punch. Such emotions

would be a very interesting input in the design of a virtual music instrument, and what remains to do is the validation, classification and detection of such kind of emotions with the proper interface. Therefore Pogany, as a member of the affective interfaces family, seems to have a priori a major advantage against other interfaces in the context of music performance and interaction.

3. OVERVIEW OF POGANY INTERFACE

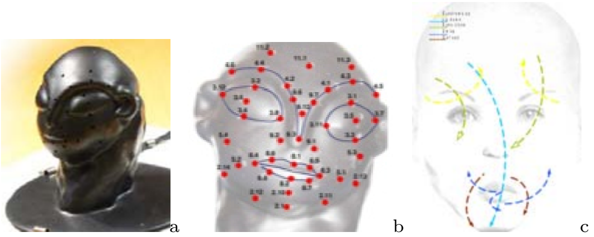

Figure 1: a)physical interface b)interaction holes (KeyPoints) c)types of meaningful gestures detected

'Pogany' is a head-shaped tangible interface for the generation of facial expressions through intuitive contacts or proximity gestures. The input to the interface consists of intentional or/and natural affective gestures. The interface takes advantage of camera-capture technology, passing a video stream to a computer for processing. A number of constraints mentioned in [1] gave to the interface the size of a joystick and the form shown in figure 1. The position of *KeyPoints*, small holes on the surface of the head used for finger position capturing, was inspired by the MPEG-4 control points. In a lit environment, passing over or covering these holes with the hands variates the luminosity level captured by a camera placed inside the facial interface. From each frame of the raw video image captured we analyze only the pixel blocks that correspond to KeyPoints and thus to gestural information on the vicinity of the head.

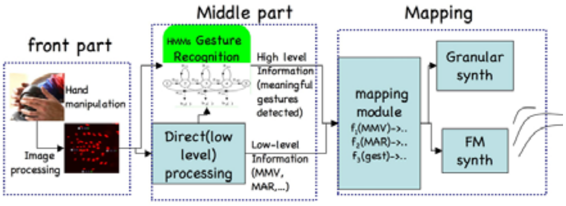

Figure 2: Architecture of Pogany music interface

In figure 2 we show an overview of our system. In the front part, we isolate the image pixel blocks associated with KeyPoint holes in each frame of the video. Then, in the middle part of the system, we process this information in order to extract the important features from gesture. These features are used either directly for mapping to music parameters or as an input to a second higher layer of processing (Gesture Recognition Module) employing HMMs. At the last part of the system and after the processing procedure, we map the processed data to a sound synthesis module that is responsible for producing the continuous sound feedback of gestural action.

3.1 Front-end: gesture capture

The front-end module of the system, is based on the use of a camera and a proper video-capture software interfacing to 'Virtual Choreographer' (VirChor) environment [5]. An image segmentation tool integrated in VirChor keeps only the important blocks from the image and finds the normalized mean luminocity value of the pixels that belong to each block. In this way we keep just one normalized value of light intrusion (called *alpha value*)for each of the pixel blocks that correspond to each KeyPoint.

$$alpha\ value = \frac{current\ luminosity}{luminosity\ at\ calibration\ time} \tag{1}$$

Alpha value is bounded between 0 and 1, with 0 corresponding to maximum light intrusion (that means no covering of the hole, thus zero activity) and 1 to minimum light intrusion (the hole is fully covered, maximum activation of the KeyPoint). The output of the front-end of the system consists of instantiations of a 43 float vector with a rate of 30 fr/sec, thus providing the gesture recognition core with a low dimensional vector instead of raw data of image format. Further information concerning the particular techniques used (image segmentation, calibration tool) can be found in [1], [8].

3.2 Middle part: gesture processing

The middle part of the system includes the processing-feature extraction unit and the gesture recognition module.

3.2.1 Gesture analysis-feature extraction

Here we extract useful features from gesture, such as energy and velocity.

Energy : Particular meaning for the mapping procedure in next stage has the definition of the energy of the signal that denotes activation in front of the interface. We call this multidimensional signal $\mathbf{X_t}$. In case we define energy as: $\mathbf{E_t} = \mathbf{X}_t^2$, where $\mathbf{E_t}$ is the temporal energy vector for the frame t=0, 1,..n. The normalized mean short time energy of the signal at frame t is:

$$\overline{E_t} = \frac{1}{N_{kp}} \sum_{j=1}^{N_{kp}} {X_{j,t}}^2, \tag{2}$$

where N_{kp} the total number of KeyPoint holes. As the signal does not take negative values, it is not wrong instead of energy to consider $\overline{M_t} = \sum_{j=1}^{N_{kp}} X_{j,t}$, where M_t the *Mean Magnitude Value per frame* (MMV), a metric for the activation of the KeyPoints of the interface.

Velocity : The velocity of the multidimensional signal is defined as:
$\mathbf{V}_t = \frac{\mathbf{X}_t - \mathbf{X}_{t-\delta t}}{\delta t}$, t=1,2,..n. We assume that $\mathbf{V}_0 = 0$;
If $\delta t = 1$, $t \geq 1$ the mean velocity value per frame t is:

$$\overline{V_t} = \frac{1}{N_{kp}} \sum_{j=1}^{N_{kp}} X_{j,t} - X_{j,t-1} \tag{3}$$

A useful measure to be used for gesture segmentation is the *Mean Activation Rate* (MAR) defined as:

$$MAR_t = \frac{1}{N_{kp}} \sum_{j=1}^{N_{kp}} |X_{j,t} - X_{j,t-1}|, \tag{4}$$

3.2.2 *Real-time Gesture Recognition Module*

The gesture recognition module is responsible for the identification of a ‘meaningful’ gesture or posture that the user addresses to the interface out of a continous stream of gesture data in real time. The difference between gestures and postures lies on the motion or motionlessness of the hand in front of the interface. Meaningful gestures (figure 1c) correspond to gestures with a particular significance that the system has been trained to recognize; classified on a high level, they function as expressivity-related commands that tend to modify the sound synthesis procedure in the form of modulation or interrupts. These gestures, in order to be distinguished from raw gesture data with higher success rate, demand permanent contact with the interface.

Inspired from our experiments for off-line isolated gestures based on HMMs presented in [6], we developed a real-time module for continuous gesture recognition. On the parallel, we were interested in keeping a high degree of expandability for the system; that means to let open future enhancements with multiple gestures, complex gestures and a large-scale gesture vocabulary.

HMM configuration : In our HMM models the number N of states for the HMM is set to 4, plus the two non-emitting states at start and at the end. We use a left-to-right-no-skips topology and an observation vector of a size of 43. The training of the system is based on the Baum-Welsh algorithm. For the recognition we employed a non-consuming Viterbi-like algorithm.

Segmentation for continuous gesture : An important issue for the recognition of the continuous gesture is segmentation. It is implemented in the activity detector module. This module is responsible for detecting predefined meaningful gestures and postures in raw gesture data (meaningless gestures and silent parts). This module makes use of the previously defined MMV and MAR metrics in combination with a number of constraints. We provide the core of the algorithm for a) separation of activity parts (gestures and postures) from silent parts (no activity in front of the interface) and b) separation of gesture from posture:

```
if MMV > thresh   then ‘activity’
else ‘silence’
if ‘activity’ then
  if MAR > thresh2 then ‘gesture’
  else  ‘posture’
```

MMV represents the general amount of activation in the vicinity of the interface. Therefore, it gives evidence or not for the existence of some kind of activity (gestural or postural) or, for values near zero, what for we call ‘gesture silence’. MAR expresses the speed of the gesture, therefore it is useful in separating gestural from postural activity. *thresh1* and *thresh2* are thresholds used to regulate the procedure relatively to light conditions. According to the output of the activity detector module described above, the system triggers or not the gesture and posture recognition and replies analogously according to the vocabulary of the meaningful types of gestural/postural activity it is trained to detect.

3.2.3 *Implementations*

In order to support the interface, we have implemented a variety of cooperating modules which we have integrated to VirChor rendering environment (image segmentation, gesture collection and data transformation algorithm, gesture detection module, HMM recognition core etc.). It is also worth to mention a module for visual feedback, in the form of an animated head for facial expressions: this permits the implicit link of user gestures with emotions arising from facial expressions. Finally, for the HMM core we used the HTKLib (library for the HTK toolkit for speech recognition)[7], adequately adapted to face with real-time recognition issues for gesture. Details for these modules, as well as a module for gesture intention recognition based on the Token-Passing algorithm (estimation for the type of gesture before it is completed), are described in [8].

4. MAPPING STRATEGIES

For the mapping module (see figure 2) we followed mixed direct and indirect strategies: the first concern low-level continuous information arising from direct gesture processing; the second refer to the semantic (high-level) information of meaningful gestures and postures. We linked this information with parameters from two types of synthesis: FM and Granular Synthesis (GS). In general, for low-level information we used one-to-one and for high-level information one-to-many mapping. Correspondences are shown in table 1.

Table 1: Mapping low & high level information to FM and GS parameters

	low level: Energy	low level: Velocity	High level: Gesture
FM	loudness	Modulation Index	Frequency Ratio
Granular Synthesis	loudness	time between grains	audio sample, grain duration, pitch transposition,...

4.1 Direct Mapping strategies

MMV and MAR metrics mentioned in the previous section serve as continuous parameters that adjust music parameters in the synthesis procedure.

4.1.1 *Mapping Energy*

The Magnitude Value per Frame (MpF) represents loudness. The function selected for this transformation was:

$$MpF(nMMV) = 1 - e^{-nMMV/a},\ 0 \leq nMMV \leq 1, \qquad (5)$$

where nMMV the normalized MMV in [0..1], a a parameter for the control of the gradient of $MpF(x)$. This parameter helps to adjust the radius of sensitivity around the interface. MpF is a conjunction of the need to quasi-linearize the distance factor and to preserve the additive effect of multiple finger haptic interaction in zero distance.

4.1.2 *Mapping Velocity*

MAR was defined as a metric for the speed of the gesture in front the interface. According to theory, the Modulation Index $MI = Am/Fm$ in FM is responsible for the brightness of the sound, as the relative strength of the different sidebands (which affects the timbre) is determined by the relationship between the modulator amplitude (Am) and the modulator frequency (Fm). Hence, we have set $MI = MAR/b$, where b a normalization factor which gives to the continuously changing value a meaningful -in musical terms- range [8]. The MAR metric was also adapted for granular synthesis, this time in order to control the time between grains.

4.2 Indirect mapping strategies

Gesture recognition acts in two levels of interest for the mapping procedure: First, at the level of gesture recognition; hand gestures with a number of frames varying from 5-70 frames (0.18-2.3 sec), are isolated and probable to get identified by the system. Second, at the level of posture recognition; in between gestures, recognizable or not by the system, whenever hands remain almost steady, the recognizer estimates the probability that this posture of the hands corresponds to one of the pre-learned postures.

The decisions we took concerned recognition for four principal gestures (figure 1c): 'eyes up', 'eyes close', 'smile', 'sad', that correspond to the emotions: 'surprise', 'suspicion', 'joy', 'sadness' respectively (for further details concerning the different types of gestures see [6]).

4.2.1 Mapping to FM

From FM synthesis practice, the Frequency Ratio produces harmonic sounds when it is a multiple of 1. On the contrary, with non-integer frequency ratio, inharmonic partials become more prominent. In our situation, facial emotions with a rather positive impact, such as 'joy' and 'surprise', were associated with values of harmonicity ratio that lead to a consonant result (an harmonic sound). On the other side, 'suspicion' and 'sadness', as emotions with mostly negative impact, were less likely to result to an harmonic spectrum. Thus, by mapping gesture infrormation to Frequency Ratio we gained control over the consonance of the resulting sound.

4.2.2 Mapping to GS

With appropriate selections of the audio material used for the grains of GS we can adequately control the nature of the sound to be representative of positive or negative emotional impacts. The parameters of the granular synthesis which participate on the mapping with gestures are the upper and bottom limits of the grain duration, the limits of transposition, the time between grains and of course, the location where the grains were extracted from. Whenever a meaningful gesture is recognized, above parameters of GS are affected relatively to their instant values.

5. EVALUATION OF THE INTERFACE

In order to evaluate our design selections for the interface, we organized an experiment with human subjects that interact musically with the interface. Context, preparation and conclusions arising from this experiment are described in this section.

5.1 Context

Discordance over the evaluation criteria and methods [8] do not provide a concrete evaluation process to follow. Under these circumstances, we decided to base our evaluation method on the axis set by Wanderley [9], adapted to the particularity of the 'Pogany' interface. Thus, the interface set-up for the experiment aimed to give clues for four main attributes: 1) time controllability, 2) sound controllability, 3) learnability, and 4) explorability. Furthermore, the evaluation process was properly adapted in order to provide an objective measure for judging the effect of high-level discrete gestural information to musical expressivity.

5.2 Preparation of the experiment

The experiment was divided in two sessions. Both sessions made use of the same synthesizer modules (FM and GS), and also shared the same mapping elements, as far as direct strategies are concerned. This means that we were motivated to create the 'loudness by distance (MMV)' and 'brightness by MAR' correspondences (described in 4.1) in order to drive the two synthesizers in both experiments.

The main difference concerns the indirect mapping strategies. In the second experiment we made use of indirect mapping exactly as described in the previous session. Whenever the user performed a meaningful gesture, this information was set to adjust a set of parameters inside the synthesizer, linearly and with a certain amount of delay.

On the contrary, in the first session the mappings of high-level information to the music parameters of the synthesizers were arbitrary. This means that after a meaningful gesture the change in parameters of FM and GS was not the one which corresponded to this particular gesture but an arbitrary selection from a sum of preset values of all gestures.

It is worth to mention that for the second experiment, an animated head was connected with the system in order to execute face-animation commands according to gestures over the interface. The recruitment of such a feedback was necessary in order to ensure that, even implicitly, the user assigns a set of actions to corresponding facial emotions, and thereafter to the corresponding sound feedback.

The system in the second experiment was trained to recognize four meaningful gestures which correspond to four basic emotions: joy, sadness, surprise and suspicion. Alteration of one of these 4 moods was triggering relative changes in the harmonicity ratio of the FM synthesis and the duration, pitch and grain source of the granular synthesis module. Additionally, five postures were recognized during session, which corresponded in five different types of activation in front of the main areas of the facial interface: the eyebrows, the eyes, the cheeks, the mouth and the nose. The activation of such cues was mapped to result to minor -in comparison with the primal gestures- change on the sound. The two symmetrical parts of the face were designed to give equal sound results.

In these two experiments we have tried to achieve global similarity on the sound quality, as well as their temporal evolution and duration variability: this would allow a fair comparison between other parameters that were different between the two sessions (indirect mapping).

The experiment : The experiment took place at LIMSI laboratory, Orsay. The light conditions during the experiment were physical (slightly non-homogenous).

The procedure was as follows: The subject received some explanations concerning the devices that he/she should use and the general concept of the experiment. Then she/he was given a short time to familiarize her/himself with the interface by observation and touch, without any kind of feedback. The next step was to get introduced to the procedure: during two sessions of 5 minutes each, the subject would be let to interact with the interface in every desirable manner. The subjects were encouraged to perform quick or slow

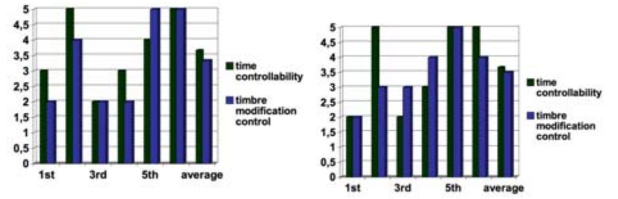

Figure 3: Evaluation of the subjects over controllability in time and sound evolution control for the (left) 1st and (right) 2nd session. Horizontal axis represents the subjects, and vertical axis the evaluation score (between 0 and 5)

movements, by distance or touch, in the front or the vicinity of the interface. Alternative modes of action were also proposed, such as tapping, caressing or hitting (slightly), using each hand separately or both hands simultaneously. After the end of the two sessions, the subjects were asked to fill a questionnaire related to the experiment.

Six subjects passed the music experiment with Pogany (five male and one female). Their age varied from 23 to 29 years old. All subjects had used before interfaces connected to a computer; two subjects have made use before of an interface for music over five times, three subjects less than five and for one subject it was the first relative experience. After interaction with the interface, all subjects answered to a set of questions concerning their experience. A number of these questions was focussing on eliciting their subjective view for the controllability potential of the setups, in terms of time, sound quality, sound modification capabilities and expressiveness. Other questions concerned the ability for recovering past gesture-sound dyads, and repetition of performed patterns during performance. Another set of questions focused on the easiness to explore new sounds in the given time, and the expectation for an hypothetical second chance. Finally, other questions referred to the visual feedback effect, the willingness of the subjects to experience the same or similar interfaces in the future, etc. Apart from the questionnaires, audio material was collected for each of the sessions of participants, as well as text records concerning the number of meaningful gestures activated for both sessions.

5.3 Results

Despite the limited number of participants, data gathered proved sufficient to provide important feedback and the base for a number of conclusions. Furthermore, it gave clues for the capability -or not- of the interface for expressive music creation.

In the domain of controllability, the participants found the quality of control in time and timbre adjustment more than satisfying (figure 3). In a range of 0 to 5 (with 0 corresponding to 'very bad' and 5 to 'excellent'), subjects evaluated the system with an average of 3.66 and 3.33 for time and timbre modification flexibility respectively for the first session. At the second session only the average value for the timbre modification increased slightly, while the temporal modification potential remained unchanged. Despite dependence of these values from factors such as the complexity of mapping and the degree of polyphony which were not the purpose in our situation, results show in general the strong acceptance of the interface as a virtual instrument. It is important to mention that the values between the two experiments show minimum differences. This seems reasonable, as the question mostly referred to the direct mapping strategies that are responsible for the modification of the most prominent parameters of a sound itself: loudness and brightness (for FM).

At this point, it was important to correlate the answers on the questionnaires with real data extracted from the performances. It was difficult to set objective evaluation criteria for the character of each performance. An interesting approach to this matter was to use a normalized MAR as a criterion for the kind of activity on the vicinity of the interface: High values of MAR make proof of high velocity in movements. Hence it was decided to calculate the zero crossing rates of $TMAR = m - MAR$, where m the estimated mean value for MAR during gesture activation. Variations of the MAR value for each session and user were recorded. After processing, the mean value of MAR was set to m= 0.7. In 2 we show the zero crossing rates for the Transposed MAR (TMAR).

Table 2: zero crossing rate of the TMAR

	1st	2nd	3rd	4th	5th	6th	average
TMAR	7	26	27	28	43	69	33.3

Making the comparison of table 2 with the graphs in figure 3, it is straightforward to understand that the subject 1 who claimed less control had the lowest TMAR score, this means that the amount of general activation rate (i.e the velocity of his gestures) was limited (with a value of 7 to a mean value of 33.3 among subjects). On the other hand, subjects 5 and 6 that ranked the system as very good or excellent had a definitely a more 'attacking' approach. It is also important to note that subject 1 had not any experience of a music interface before, a fact that gives the clue that previous experience with interfaces probably affects the learning curve and the easiness of manipulation, as well as the overall view on the effectiveness of such as system.

One of the most important issues of our work concerned the expressivity capabilities through the interface. This fact had had a straight impact on decisions taken concerning the configuration of each session. After the experiments, participants were asked in the questionnaire to give a judgement about the system expressive capabilities. The question was posed relatively between sessions, asking if there was one session in particular that help them more in expressing themselves. It was impressive that all but one subject found the second configuration better in expressing themselves, while the other subject found two sessions as of equal expressivity capabilities. This fact, in correlation with the minor divergence in evaluation of time and timbral control between sessions, pose an important issue for expressivity related excusively to the process of decoding gesture cues in a higher-level approach employed in the second session. It is also overwhelming that one of the subjects- the one that was statistically found to have the best score of the TMAR value- evaluated the level of control as being better in the first experiment, while in the same time confirmed the superiority of the second session in terms of expressivity.

In terms of learnability, five of the six subjects claimed that they definitely succeeded in learning new gestures throughout the little time they were given for manipulation, while the sixth-referred as 1st on statistics- also gave a positive answer but with less certainty. On the question if, even after the experiment, the subject can recall correspondences between gestures and resulting sounds all subject gave a positive answer, each time with more or less certainty. The opinion of the subjects on the matter was of great interest, as with their spontaneous thoughts they have underlined one of the most important issues for an interface: how to establish a learning curve that would not discourage amateurs from getting on with learning and in the same time set high limits for the perfection of performance and thus be intriguing for more experienced users to go on exploring the capabilities of the virtual instrument. Hence, as far as the term of learnability converges with the issues set by the term of explorability, it would be worth having a look at the statements of some of the subjects (1st, 5th and 6th respectively):

'Many difficulties encountered when trying to explore new sounds... difficulties to find a logic and patterns...'

'...For the manipulation some time is necessary to explore the possibilities but when it's done, it is very interesting to produce different sounds.'

'... However, the control on the second experiment was less effective, maybe due to that it demanded a higher degree of expertise gained through practice.' .

In a question asking for the subject's expectation concerning the exploration of new sounds in an hypothetical second chance with the interface all subjects have responded positively, as if the impression created to themselves is that there is still part of the potential of the interface not discovered yet. Some of the subjects underlined the importance of the visual feedback in the form of an animated head for the exploration of the sound capabilities of the interface. Concerning this kind of feedback, all subjects found it in all ways useful, also mentioning 'control' and 'logic' in the sound as factors of the creation where it can contribute.

About the general impression on the interface, the 1st subject was rather negative. He insisted in the problems he encountered in trying to understand how exactly it works. All the other subjects found the interface at least interesting. Although some subjects claimed not to have familiarity with the 'type' of music it produced, or even not to find it pleasant, this did not prevent them from attaining a good general impression:

'...sometimes it is noisy, but it's funny. I felt like playing (good or bad!) a music instrument...'

A subject underlined the constructive appropriateness of the 'Pogany' interface for such a scope:

'Touching the interface seems important and the contact/touch impression is quite nice...'

Finally, some of the subjects proposed types of usage where setups such as the one of 'Pogany' for music would prove particularly useful, such as for blind people. An inspiring point of view was also set from one of the subjects, mostly concerning intuitive purposes of tangible interfaces for music :

'With this interface, people have to guess how to touch it, to learn it by themselves...perhaps a 'traditional' instrument player, after practicing with an interface such as the head interface, will try to find other manners to play with his instrument and produce new sounds.'

6. CONCLUSIONS-FUTURE RESEARCH

The impressions we obtained from this experiment were encouraging at many different levels. Firstly, the high-level gestural information decoding module in the second session proved to be particularly useful in terms of expressivity of the user, as stated by all the subjects and confirmed by the equivalence of the two sessions in all other aspects of synthesis' global quality. Second, even through a non-complicated mapping, the general impressions for timbre modification and time precision were positive, as well as for the interface itself as a device. Third, the interface succeed in providing sufficient conditions for learning patterns and exploring new gestures, with a priority in the advanced users learning curve. Finally, even not proved from the particular experiment, the decisions concerning the expandability options of the setup that were left open during architectural design (such as the option for the interface to be trained for complex gestures) were not discouraged by the results of the experiment.

Recent results showed that the use of an interface for music within an affective protocol could be beneficial. In the future we will focus on consolidating our results with further experiment and artistic performance use cases. In this framework it is worth to also deal with technical issues concerning the interface: robustness, increased sensibility and enhanced multimodal techniques, instability under difficult light conditions, latency etc.. Enhancements within pure recognition issues could also help to improve the overall performance of the interface.

Finally, for an affective interface such as 'Pogany', even if the visual head animation feedback implicitly creates correspondences between users emotions and sound results, a study of relative research in psychology field (such as a model for touching parts of the body) is more than imperative. However, such a model is difficult to evaluate, due to the polyparametric nature of actions of touch among people and the social factor effect. Nevertheless, this would surely help create a solid base for the semantic space to be linked to gestural information.

7. REFERENCES

[1] Jacquemin, C. 'Pogany: A tangible cephalomorphic interface for expressive facial animation', ACII '2007, Lisbon, Portugal, 2007.

[2] Paiva, A., Andersson, G., Hook, K., Mourao, D., Costa, M., Martinho, C.'Sentoy in FantasyA: Designing an affective sympathetic interface to a computer game', Personal Ubiquitous Comput. 6(5-6) (2002) 378389.

[3] Yonezawa, T., Suzuki, N., Mase, K., Kogure, K., 'HandySinger: Expressive Singing Voice Morphing using Personified Hand-puppet Interface ' , NIME 06, Paris, 2006.

[4] Ekman, P., Friesen, W.V.: 'Facial action coding system: A technique for the measurement of facial movement'. Consulting Psychologists Press, Palo Alto, CA, USA,1978.

[5] http://virchor.sourceforge.net

[6] Maniatakos, F., 'Affective interface for emotion-based music synthesis', Sound and Music Computing conference SMC07, Leykada, Greece, 2007.

[7] htk.eng.cam.ac.uk/

[8] Maniatakos, F., 'Cephalomorphic interface for emotion-based musical synthesis', ATIAM Master Thesis, UPMC Paris 6 & IRCAM, LIMSI-CNRS, Orsay, France, 2007.

[9] Wanderley, M., Orio, N. 'Evaluation of input devices for musical expression; borrowing tools from HCI', Computer Music Journal, 26(3):6276, 2002.