x2Gesture: how machines could learn expressive gesture variations of expert musicians

Christina Volioti
MTCG Lab, Department of Applied Informatics, University of Macedonia,
156 Egnatia Street, GR-540 06,
Thessaloniki, Greece,
christina.volioti@uom.edu.gr

Sotiris Manitsaris
Centre for Robotics, MINES ParisTech, PSL Research University,
60, Boulevard St-Michel,
75272, Paris, France
sotiris.manitsaris@mines-paristech.fr

Eleni Katsouli
MTCG, Department of Applied Informatics, University of Macedonia,
156 Egnatia Street, GR-540 06,
Thessaloniki, Greece,
katsouli@uom.edu.gr

Athanasios Manitsaris
MTCG Lab, Department of Applied Informatics, University of Macedonia,
156 Egnatia Street, GR-540 06,
Thessaloniki, Greece,
amanitsaris@uom.edu.gr

ABSTRACT

There is a growing interest in 'unlocking' the motor skills of expert musicians. Motivated by this need, the main objective of this paper is to present a new way of modeling expressive gesture variations in musical performance. For this purpose, the 3D gesture recognition engine 'x2Gesture' (eXpert eXpressive Gesture) has been developed, inspired by the Gesture Variation Follower, which is initially designed and developed at IRCAM in Paris and then extended at Goldsmiths College in London. x2Gesture supports both learning of musical gestures and live performing, through gesture sonification, as a unified user experience. The deeper understanding of the expressive gestural variations permits to define the confidence bounds of the expert's gestures, which are used during the decoding phase of the recognition. The first experiments show promising results in terms of recognition accuracy and temporal alignment between template and performed gesture, which leads to a better fluidity and immediacy and thus gesture sonification.

Author Keywords

expert gesture, expressive variations, musical performance, confidence bounds, gesture sonification, fluidity, immediacy

ACM Classification

H.5.2 [Information Interfaces and Presentation] User Interfaces – Interaction Styles, H.5.5 [Information Interfaces and Presentation] Sound and Music Computing.

1. INTRODUCTION

Gesture constitutes a component of human expression. It can also be characterized as a self-contained part of music. A musical performance is a sequence of expressive gestures that encapsulate both theoretical knowledge and practical motor skills. Each musical performance is unique due to expressivity, since for a given musical excerpt, interpretations can vary greatly, depending on the performer or even on expression that the performer has each time s/he plays the same piece [11].

Recently, research on the capturing and recognition of musical gestures has become very appealing. Many researchers and musicians have developed interfaces that use machine learning algorithms and aim at recognizing not only the cinematic aspects of the gesture [4][17], but also measurable parameters about expressivity [12]. From a machine learning point of view, there is usually an important compromise to make between a fast, or a rich training of the model. There are musical interfaces that are based on one-shot learning [4][12][17], in which the system requires only one training example instead of large data sets; thus, the training time is greatly reduced but significant limits are put on the modeling of expressive variations of the same gesture. Thus, the modeled information is less rich than when using large data sets. Moreover, within a sensory-motor learning context, it is important to identify precisely the tolerance between the executions of an expert performer in order to provide meaningful feedback to the learner. Therefore, the mathematical description of how an expressive gesture is being performed, along with the modeling of its variations are becoming crucial research topics.

Our approach is based on the concept that expressiveness is an intended gestural variation, which should be taken into account when modeling the gesture. In one of our previous work, Manitsaris et al. [22] has proposed a way to model offline gestural know-how in craftsmanship. As an extension of this work, we propose x2Gesture, which aims at recognizing musical expert gestures in real-time taking also into account the expressive variations. This is accomplished by implementing a) the existing work which models expert motor skills, and b) machine learning algorithms for real-time expert gesture recognition. Finally, our proposed methodology can support a unified user experience for both *learning* of expert musical gestures and *performing* musical gestures.

This paper is structured as follows: firstly, we review the state of the art (SoA) concerning machine learning algorithms that are used for gesture recognition (Section 2). Then our methodological approach (Section 3) and its implementation in two case studies (Section 4) are described. Finally, we conclude with our first evaluation results (Section 5).

2. RELATED WORK

2.1 Expert musical gestures

Firstly, we shall define some terms, which are key to our methodological approach. The term 'musical gestures' lies in the intersection between observable actions and mental

NIME'16, July 11-15, 2016, Griffith University, Brisbane, Australia.

representations [13]. A good definition of this, taken from Hatten (2003) is [19]: ‘musical gesture is biologically and culturally grounded in communicative human movement. Gesture draws upon the close interaction (and inter-modality) of a range of human perceptual and motor systems to synthesize the energetic shaping of motion through time into significant events with unique expressive force’.

When we refer to expressive gesture, what do we mean? According to [6], ‘expressiveness is conveyed by a set of temporal and spatial characteristics that operate more or less independent from the denotative meanings of those gestures’. The notion of expressivity measures *how* the expert gesture is performed. Hence, *how* an expressive gesture is performed is equally as important as *what/which* expressive gesture is performed [18].

By using the term ‘expert gestures’, we mean that performers have mastered their gestural skills. For example, they are those gestures that require years of training and practice before performers are able to perform them. Although this kind of expert has acquired high-level motor skills, expressive variations may occur between the different musical interpretations, even unconsciously. In order to control and measure expressive variations, some researchers use the ‘neutral performance’ as a reference [7], which is the performance played without any specific expressive intention. Alternatively, the mean of all the performances was taken as a reference [23].

2.2 Machine learning algorithms

Machine learning algorithms, such as those based on Hidden Markov Models (HMMs) [20], Dynamic Time Warping (DTW) [1], Hierarchical Hidden Markov Models (H-HMMs) [15], Sequential Monte Carlo technique [12] etc., are widely used for gesture recognition systems for continuous interaction. [2][3][4] successively developed a system based on a hybrid model between HMMs and DTW, called Gesture Follower (GF), for both continuous gesture recognition and following, between the template or reference gesture, and the incoming or performed gesture (template-based method). It can learn a gesture from a single example (one-shot learning), by associating each template gesture to a ‘state’ of a hidden Markov chain [5]. During the performance, a continuous estimation of parameters is calculated in real-time, by providing information for the temporal position of the performed gesture. Time alignment occurs between the template and the performed gesture, as well as offering an estimation of the time progression within the template in real-time.

One limitation of HMMs is that observations are produced at the frame level, and as a consequence they do not support the transitions between segments [15]. Therefore, [14][15] developed a system based on H-HMMs with two levels for real-time gesture segmentation and recognition. Similarly to GF, it adopts a template-based method and implements one-shot learning. The system is trained with a single pre-segmented gesture, which is annotated by the user. Each segment is associated with a high-level state (segment state), which generates the sub-models of the signal level (lower level), encoding the temporal evolution of the segment [14][16].

The aforementioned methodologies and research approaches do answer the question of what/which gesture is performed, but not how expressive gesture is performed. [12] further extended the research by proposing a template-based method which implements a Sequential Monte Carlo technique. Its main advantage is that the recognition system, named Gesture Variation Follower (GVF), is being adapted to gesture expressive variations in real-time. Specifically, in the learning phase only one example per gesture is required. Then, in the performing phase, time alignment is computed continuously and expressive variations (such as speed, size, etc.) are estimated between the template and the performed gesture [10][12].

2.3 Conclusions from SoA and Motivation

Leveraging the above, we can conclude that the majority of algorithms answer the question of what/which gesture is performed, or how it is performed, or both. Furthermore, in most cases, a parameter is implemented, measuring how much the performance is allowed to be different from template gestures [25]. Additionally, the users can control the degree of generalization of the model to ensure a robust estimation of their performed gestures with this parameter [15]. In GF and GVF, this parameter is called *tolerance* [8][25] and in [15] which is based on H-HMMs, *variance offset*. The main advantage of this parameter is that if its value is low, the system will be more robust and will recognize gestures with more accuracy. If it is set high, the system will be less reliable, due to the fact that the model will be too general and it will lead to overlaps between classes [15][25]. However, the main drawback is that the value of this parameter remains fixed during the performance of the gesture. This leads to the possibility that the system might fail to recognize some variations *within* the gesture, because it might require a slightly higher or slightly lower value of this parameter. Moreover, there is an impact on the time alignment between template gesture and performed gesture, which can vary importantly, thus reducing the immediacy and fluidity of the gesture sonification.

An additional conclusion from the literature review is that, the purpose or end-use of the implementation of algorithms is for installations, performances or even entertainment. But what happen in the case of the educational and learning process? Can the existing algorithms successfully recognize expressive gesture variations between expert and learner’s performances? For this reason, our proposed methodology deals with the know-how transmission between expert and learner. Moreover, we propose *confidence bounds*, instead of fixed values of tolerance and variance offset, which are derived from expert gesture performance [22] and can dynamically and more precisely recognize the variations that occur *within* the learner’s performance (performed gesture) in relation to the expert’s performance (template gesture). Apart from the learning the scenario, the proposed methodology gives also the possibility to the user to perform his/her own musical gestures and control sound parameters.

3. MODELING AND RECOGNITION

In the proposed methodology, the goal is not simply to train, recognize and sonify expert musical gestures, but by exploiting the existing methodologies and adding the parameter of confidence bounds, to develop a system that will be able to recognize expressive variations that take place within the gesture performance.

3.1 Expert operational model

The first step was to model expert gestural know-how in the case of the piano. This was accomplished by capturing expert musical gestures while the expert performed specific musical gestures on the piano. Then, expert gestural analysis was conducted. The purpose of using the State Space estimation methodology was two-fold: a) in order to model expert musical gestures, we built an operational model that describes how expert gestures are performed; and b) in order to develop a system that will be able to recognize more accurately the variations that might occur within the learner’s performance, we extracted the confidence bounds, based on the iterations of the same expert musical gesture, from the expert operational model [22].

The general specification of the State Space presentation of vector Y_t is given by the following dynamic system [21]:

$$\boldsymbol{Y}_t = \boldsymbol{\beta}_t + \boldsymbol{Z}_t\boldsymbol{a}_t + \boldsymbol{\varepsilon}_t \qquad (1)$$
$$\boldsymbol{a}_{t+1} = \boldsymbol{\gamma}_t + \boldsymbol{W}_t\boldsymbol{a}_t + \boldsymbol{\eta}_t \qquad (2)$$

where:

- $\boldsymbol{Y}_t$ is a n×1 vector, which can refer to as the signal or observation equation (1)
- $\boldsymbol{a}_t$ is an m×1 vector of possibly unobservable state variables, which can be referred to as the state or transition equation (2)
- $\boldsymbol{\beta}_t, \boldsymbol{Z}_t, \boldsymbol{\gamma}_t$ and $\boldsymbol{W}_t$ are conformable vectors and matrices
- ε_t and η_t are vectors of mean zero, Gaussian disturbances

Following equations (1) and (2), in our case the functional version of the expert operational model, presenting the gestures of the right hand with respect to dimension X (RHX), is as follows:

$$\boldsymbol{RHX}_t = \boldsymbol{Z}_{1t}\boldsymbol{a}_1 + \boldsymbol{Z}_{2t}\boldsymbol{a}_{2t} + \boldsymbol{\varepsilon}_{1t} \qquad (3)$$
$$\boldsymbol{a}_{2t} = \delta_1\boldsymbol{a}_{2t-1} + \boldsymbol{\eta}_{1t} \qquad (4)$$

where:

- $\boldsymbol{Z}_{1t} = [\boldsymbol{I}\ \ \boldsymbol{RHZ}_{t-1}\ \boldsymbol{RHY}_{t-1}\ \boldsymbol{LHX}_{t-1}]$, $\boldsymbol{Z}_{2t} = [\boldsymbol{RHX}_{t-1} - \boldsymbol{RHX}_{t-2}]$
- **I** = unit vector, $\mathbf{Z}_\mathrm{t} = [\mathbf{Z}_{1\mathrm{t}}\ \mathbf{Z}_{2\mathrm{t}}]$, $'$ = transposition, and
- $\boldsymbol{a}_1' = [a_{10}\ a_{11}\ a_{12}\ a_{13}]$, $\boldsymbol{a}_{2\mathrm{t}}$ and δ_1 are parameters to be estimated.

Analytically, the equations to be estimated are as follows:

$$RHX_t = a_{10} + a_{11}RHZ_{t-1} + a_{12}RHY_{t-1} + a_{13}LHX_{t-1} + a_{2t}(RHX_{t-1} - RHX_{t-2}) + \varepsilon_{1t} \qquad (5)$$
$$a_{2t} = \delta_1 a_{2t-1} + \eta_{1t} \qquad (6)$$

In our piano case study, we mostly focused on the gestures made by playing with two hands. Thus the complete operational model has two sets of equations: three right hand equations (RHX_t, RHY_t and RHZ_t), and three left hand equations (LHX_t, LHY_t and LHZ_t).

Having estimated the system of equations (5) and (6), the expert operational model is dynamically simulated and the dependent variables are forecasted. Consequently, the estimated forecast standard error is derived according to:

$$RHX\_forecast\ se_t = s\sqrt{1 + RHX_t'(Z_t'Z_t)RHX_t} \qquad (7)$$

where s = standard error of the estimated equation.

Then, we calculated the *confidence zone* for each musical gesture, including *confidence bounds* (a higher and a lower bound). The equations of the higher (8) and lower (9) bound referring to right hand are the following, where RHX_f_t is the forecasted data series at discrete time t, and RHX_se_t is the forecasted standard error:

$$RHX\_high_t = RHX\_f_t + RHX\_se_t \qquad (8)$$
$$RHX\_low_t = RHX\_f_t - RHX\_se_t \qquad (9)$$

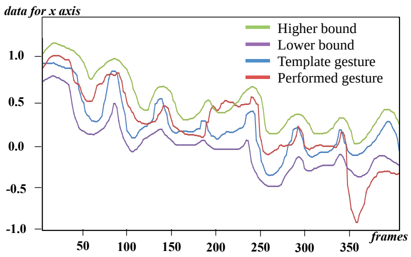

Figure 1. Confidence bounds of the expert musical gesture.

If the performed gesture is between these confidence bounds (Figure 1) during the whole performance, this means that we can successfully take into consideration the expressive variations that occur between the template and performed gesture. We further generalize this methodology by implementing the confidence bounds and using them with machine learning algorithms, in order to recognize expressive variations that might take place between the learner's (performed gesture in Figure 1) and expert's (template gesture in Figure 1) performance in real-time, in order to improve both the recognition results and the gesture sonification.

3.2 Implementation

x2Gesture is based on GVF library [1] [10][12][26], and implements the State Space and Particle Filter algorithm. The state elements are the gesture characteristics, which are for example, the time progression of the performed gesture (temporal alignment), the relative speed, the scaling coefficient (size) and the angle of rotation (orientation). The transition function is linear, relying on a Gaussian noise [9] and the observation function is the distance between the adapted template gesture and the performed gesture [10].

The algorithm includes two phases: the learning (or training) and the following (or recognition) phase. x2Gesture is first trained with a single expert example per gesture along with an audio file (pre-recorded sound). This process is repeated until the system is trained with all the template gestures, which are mapped to the respective sounds. Thereafter, in the following phase, the learner or performer imitates in real-time the same expert gesture. For each performed musical gesture, x2Gesture selects the appropriate confidence bounds, which correspond to the performed gesture. At the same time, the model aligns the incoming gesture onto the template gesture, estimating also the gesture variations [10][26]. Moreover, the system resynthesizes a plausible imitation of the original (expert) sound in real time according to the learner's gesture performance, by using the granular sound synthesis engine. The better the recognition results are, the better the gesture sonification and the re-synthesis of the sound will be.

The added value in the recognition system, as it is already mentioned, is the implementation of the confidence bounds. In this way, during the recognition, the system can prevent numerical errors that might happen due to expressive variations, and as a result, confidence bounds could improve the gesture classification and therefore the gesture sonification. This happens because confidence bounds are extracted from the expert operational model and they are not a fixed number selected by the user during the learning process or musical performance.

4. CASE STUDIES

For the evaluation of x2Gesture we organized two case studies: a) a learning scenario of expert musical gestures and b) a performance with musical gestures by using Intangible Musical Instrument (IMI) [24]. IMI setup is a construction made of Plexiglas, shaped so as to look like a table on which the learner and/or performer can put his/her hands and perform musical gestures. In both case studies, three musical gestures were included in the musical vocabulary (Table 1): a) G_1: ascending scale performed in legato style, b) G_2: descending arpeggio performed in staccato style, and c) G_3: a musical excerpt from a famous Greek song.

Table 1. (a) G_1: ascending scale, (b) G_2: descending arpeggio and (c) G_3: a musical excerpt from a Greek song

(a)	(b)	(c)
Slow – 72 bps (adagio)	*Slow* – 80 bps (andante)	*Slow* – 72 bps (adagio)
Normal – 100 bps (andante)	*Normal* – 112 bps (moderato)	*Normal* – 100 bps (andante)
Fast – 116 bps (moderato)	*Fast* – 126 bps (allegro)	*Fast* – 116 bps (moderato)

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

All gestures have duration approximately 10-15 seconds and each user was asked to repeat each gesture five times. Apart from that, the user repeated each musical gesture in two different rhythms, slow and fast (Table 1).

In order to capture in real-time the musical gestures, two inertial sensors (Animazoo IGS-150 [2]) were used. These sensors are gyroscopes, providing XYZ axis rotations. Also they were placed on user's two hands, and specifically on wrists.

4.1 Case study I: Learning

In the learning scenario, 7 users were participated, one from whom was the expert pianist and the rest 6 were the learners. The purpose was to capture the expert pianist while performing the expert musical gestures on the piano (Figure 2 (a)). For each expert musical gesture, one iteration was selected as the reference gesture. Then, the rest iterations have been aligned and timely warped based on the reference gesture, using the DTW technique. Therefore, all the iterations of the same gesture transformed into having the same duration. These transformed data were averaged per variable and the result was used in the estimation of the expert operational model and in the extraction of confidence bounds, following the steps, which are described in Section 3.1.

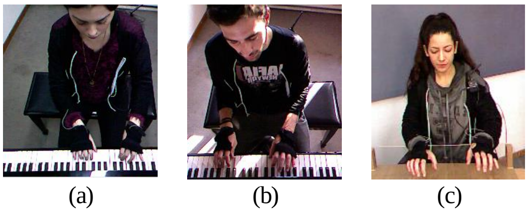
(a) (b) (c)

Figure 2. Different roles of users: (a) expert, (b) learner and (c) performer.

Subsequently, x2Gesture was trained with the three template gestures (reference). In the recognition phase, each one of the six learners performed the same expert musical gestures on the piano (Figure 2 (b)) five times. Their gestural data were captured in order to evaluate the recognition results of the model, as well as the accuracy and reliability of the confidence bounds.

4.2 Case study II: Performing

In the second case study, 6 performers were participated in total. For each performer the expert operational model and the confidence bounds were extracted. Moreover, apart from their gestural data, the sound that was produced was also recorded. Therefore, in the training phase, both reference gesture of each performers and the respective sound were given as input. In the recognition phase, each performer (Figure 2 (c)) performed the same musical gestures by using IMI, in order to resynthesize the pre-recorded sound in real-time.

5. EVALUATION

The goal of the experiment is to assess the recognition accuracy of x2Gesture, which implements the confidence bounds, comparing it also with established techniques, such as GF and GVF. The evaluation method that was used is called 'jackknife', or 'leave-one-out' approach. The basic idea is leaving out one or more observations at a time from the sample set. Practically, the database contains observations from five iterations of three musical gestures. Five distinct datasets have been created for each iteration of performed gesture. Therefore for each jackknife iteration, one dataset is left out to train the model M_i per musical gesture G_i and the rest four are used for testing. Two metrics were also used to evaluate the recognition accuracy: a) Precision, which takes into account the false recognitions and b) Recall, which takes into account the missed recognitions.

5.1 Evaluation of case study I

For the evaluation of the learning scenario, jackknife method was used only in expert's data. The aim is to evaluate the accuracy of the expert operational model and confidence bounds. Table 2 presents the results that x2Gesture gave for the five jackknife iterations, as well as the values of Precision and Recall per G_i.

Table 2. x2Gesture: Precision and Recall per expert gesture

		Maximum likelihoods			
		M_1	M_2	M_3	*Recall*
Observations	G_1	20	-	-	**100%**
	G_2	-	20	-	**100%**
	G_3	-	-	20	**100%**
	Precision	**100%**	**100%**	**100%**	

Because the results seemed to be perfect, we repeated the same experiment with GF and GVF. Therefore, Table 3 shows briefly the values of Total Precision and Total Recall per expert gesture from recognition with GF and GVF.

Table 3. GF and GVF: Precision and Recall per expert gesture

		GF		**GVF**	
		Precision	*Recall*	*Precision*	*Recall*
Observations	G_1	100%	100%	95%	100%
	G_2	100%	100%	100%	100%
	G_3	100%	100%	100%	95%
	Total	**100%**	**100%**	**98%**	**98%**

The high recognition results that x2Gesture, GF and GVF gave, can be explained by the fact that the expert pianist was very dedicated and focused on the expert performance of musical gestures. This resulted in not occurring expressive variations, even unconsciously, between the different iterations of musical interpretations. The tolerance that was used for these tests was 0.1 for both GF and GVF.

In order to complete the evaluation of the learning scenario, learners have to imitate the same expert musical gestures on the piano. The specific dataset contains: 6 learners * 3 musical gestures * 5 iterations = 90 gesture examples. The value of tolerance that was selected was 0.2 for GF and 0.1 for GVF. These tolerance values were the result of many experiments, as they gave better results for these specific musical gestures in comparison with smaller or larger tolerance values.

After having trained the system with the expert's template gesture (reference), the data from the learners' performances were given for recognition. The recognition results are presented in Table 4:

Table 4. GF, GVF and x2Gesture: expert – learners

	GF		**GVF**		**x2Gesture**	
	Precision	*Recall*	*Precision*	*Precision*	*Precision*	*Recall*
G_1	59%	57%	70%	53%	100%	70%
G_2	79%	37%	45%	43%	65%	37%
G_3	53%	83%	53%	70%	48%	83%
Total	**64%**	**59%**	**56%**	**55%**	**71%**	**63%**

According to Table 4, we can conclude that from the comparison of recognition percentages, x2Gesture gives better results than the others. These results are consistent to what we expected, and confirm the hypothesis that the recognition results can be improved with the implementation of confidence bounds. Moreover, the results confirm that confidence bounds can dynamically and more precisely recognize the variations that might occur within the learner's performance and expert's performance.

5.2 Evaluation of case study II

In the performance case study with the use of IMI, all 6 performers execute the musical gestures five times. As it is mentioned, during the

[2] http://synertial.com/

performance they were also asked to perform the gestures either slower or faster. The dataset for this case study includes per user: 3 musical gestures * 5 iterations (which contain data from slow, normal and fast speed) = 15 gesture examples. For the case study II, the value of tolerance that was selected was 0.1 for both GF and GVF.

x2Gesture was trained per user with the three template gestures along with the pre-recorded sounds. Then, in the recognition, x2Gesture selected the appropriate confidence bounds, according to the performed gestures, and resynthesized the sound in real-time, by using the granular sound synthesis engine. The recognition results per performer and per algorithm are shown in Table 5:

Table 5. GF, GVF and x2Gesture: performer – performer

		GF		GVF		x2Gesture	
		Precision	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*
User 1	G_1	75%	75%	68%	85%	64%	80%
	G_2	82%	90%	76%	65%	71%	75%
	G_3	83%	75%	61%	55%	79%	55%
	Total	**80%**	**80%**	**68%**	**68%**	**71%**	**70%**
User 2	G_1	100%	100%	100%	100%	100%	100%
	G_2	100%	100%	100%	100%	100%	100%
	G_3	100%	100%	100%	100%	100%	100%
	Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
User 3	G_1	100%	100%	95%	95%	87%	65%
	G_2	100%	100%	100%	100%	100%	90%
	G_3	100%	100%	95%	95%	67%	90%
	Total	**100%**	**100%**	**97%**	**97%**	**85%**	**82%**
User 4	G_1	71%	50%	78%	90%	95%	95%
	G_2	54%	35%	95%	90%	91%	100%
	G_3	55%	90%	89%	80%	100%	90%
	Total	**60%**	**58%**	**87%**	**87%**	**95%**	**95%**
User 5	G_1	100%	100%	95%	100%	100%	100%
	G_2	100%	100%	100%	100%	100%	100%
	G_3	100%	100%	100%	95%	100%	100%
	Total	**100%**	**100%**	**98%**	**98%**	**100%**	**100%**
User 6	G_1	100%	100%	95%	100%	100%	100%
	G_2	100%	100%	100%	95%	100%	100%
	G_3	100%	100%	100%	100%	100%	100%
	Total	**100%**	**100%**	**98%**	**98%**	**100%**	**100%**
	Grand Total	**90%**	**90%**	**91%**	**91%**	**92%**	**91%**

In the last row of Table 5, grand total from all performers are presented. If we interpret the table according to the last row, x2Gesture gives the highest results (with GVF and GF to follow).

Alternatively, if we interpret the results per performer, we can conclude that GF gives better recognition results than the others, while x2Gesture and GVF come after. This can be explained by the fact that in four out of six performers GF gives 100% in Precision and Recall, while x2Gesture in three and GVF in one. However, the majority of the percentages per performer from x2Gesture and GVF are really close to 100% (i.e. 98%, 97%, etc.), which means that the model did not manage to recognize correctly one or two gestures.

5.3 Evaluation on recognition stability and time

At this point, it is important to highlight an additional advantage of the implementation of confidence bounds. Figure 3 presents the *time progression* of the recognized G_3 from user 3 (case study II). Time index '0' is the beginning of the gesture and time index '1' is the end of the gesture. Compared to the other two algorithms, x2Gesture is more stable during the recognition process and faster than the others, because the system recognizes correctly G_3 from the 1st frame, resulting to the increase of the maximum likelihood that refer to G_3.

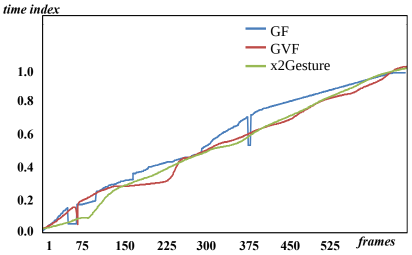

Figure 3. Gesture progression through the temporal alignments of GF, GVF and x2Gesture.

This can be also confirmed by the Figure 4(c), which presents the maximum instant likelihood. Therefore, the gesture sonification is more fluid and immediate because the new synthesized signal is much closer to the template sound. GVF becomes stable after 112 frames. Figure 4(b) shows the latency before G_3 takes the maximum likelihood. GF recognizes correctly G_3 after 145 frames, as it seems to oscillate between G_3 and G_1. The maximum likelihoods along with their transitions between gestures are presented in Figure 4(a). Although, in the end all three algorithms recognize correctly G_3, the production of the sound differs in three algorithms.

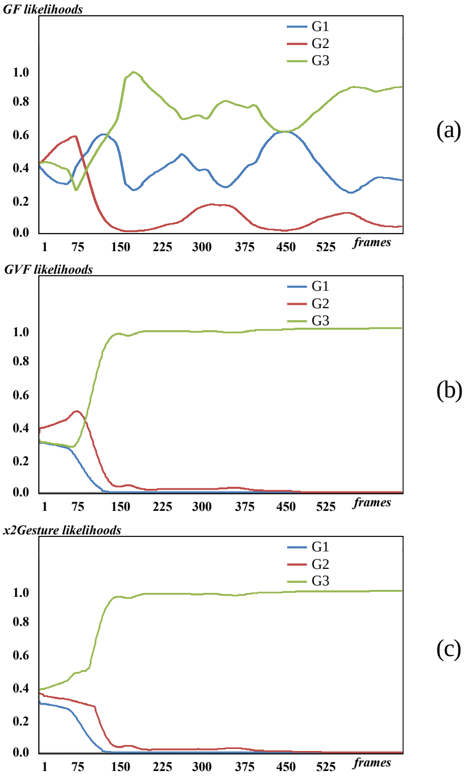

Figure 4. Instant likelihoods per frame using (a) GF, (b) GVF and (c) x2Gesture.

Additionally to the above specific example in which x2Gesture recognizes the right musical gesture faster than the other two algorithms, Table 6 presents the average time that each algorithm succeeds to recognize each musical gesture correctly.

Table 6. Average time that GF, GVF and x2Gesture need to recognize each gesture correctly

	GF		GVF		x2Gesture	
	Mean	*St.Dev.*	*Mean*	*St.Dev.*	*Mean*	*St.Dev.*
G_1	8,43	7,27	2,25	1,65	2,23	1,53
G_2	1,94	3,17	3,25	2,56	1,65	1,73
G_3	0,71	1,34	2,46	2,43	0,96	0,68

In order to evaluate the response time in real time, the database from case study I (expert – learners), was used. According to the small values of mean and standard deviation for each gesture, it can be further confirmed that x2Gesture can recognize faster and more stable the musical gestures without oscillating between all three musical gestures. However in G_3, GF has smaller mean value than x2Gesture, but larger standard deviation. This can be interpreted by the fact that, although GF has recognized more G_3 in comparison to x2Gesture (Precision in Table 4), the values of time that GF has taken the highest instant likelihoods for G_3 varied with each other more (max. time value 5,02 sec. and min. time value 0,11 sec.) than in x2Gesture (max. time value 3,01 sec. and min. time value 0,13 sec.).

6. CONCLUSION AND PERSPECTIVES

Summarizing, we propose the 3D gesture recognition engine 'x2Gesture', which has been especially designed to address the needs of both learning the expert musical gestures and live performing through gesture sonification. Moreover, the proposed modeling of the expressive variations and the output confidence bounds, led to higher recognition accuracy even in multi-user use-cases, by taking into consideration the expressive variations that might occur. Furthermore, the first evaluation results prove that there is a more fluid and immediate temporal alignment with the correct gesture.

Our future work is to generalize our methodology in order to be used in a variety of different disciplines, by creating connections between them. For example, to combine music with mathematics, or physics, or drawing, etc. Expressivity and creativity will be the core of these interdisciplinary musical performances.

7. ACKNOWLEDGMENTS

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

We wish to express our sincere appreciation to all musicians who took part in this experiment. Also, we would like to thank Gavriela Senteri for her contribution in the validation of the models.

8. REFERENCES

[1] F. Bettens, and T. Todoroff. Real-time dtw-based gesture recognition external object for max/msp and puredata. *In Proc. of the SMC 2009 Conference*, 30, 35, 2009.

[2] F. Bevilacqua, F. Guédy, N. Schnell, E. Fléty, and N. Leroy. Wireless sensor interface and gesture-follower for music pedagogy. *In Proc. of the NIME'07*, New York, 2007, 124-129.

[3] F. Bevilacqua, R. Muller, and N. Schnell. MnM: a Max/MSP mapping toolbox. *In Proc. of the NIME'05*, Vancouver, Canada, 2005.

[4] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana. Continuous realtime gesture following and recognition. *In Proc. of the 8th International Conference on Gesture in Embodied Communication and Human-Computer Interaction*, Bielefeld, Germany, 2009.

[5] A.F. Bobick, and A.D. Wilson. A state-based approach to the representation and recognition of gesture. *IEEE TPAMI*, 19, 12, 1997, 1325-1337.

[6] A. Camurri, G. De Poli, M. Leman, and G. Volpe. A multi-layered conceptual framework for expressive gesture applications. *In Proc. of the International MOSART Workshop*, Barcelona, Spain, 2001.

[7] S. Canazza, G. De Poli, C. Drioli, A. Roda, and A. Vidolin. Modeling and control of expressiveness in music performance. *In Proc. of the IEEE*, 92, 4, 2004, 286-701.

[8] B. Caramiaux. Études sur la relation geste–son en performance musicale. *Ph.D. Thesis*, Pierre and Marie Curie University (Paris 6), France, 2011.

[9] B. Caramiaux. Motion Modeling for Expressive Interaction: A Design Proposal using Bayesian Adaptive Systems. *In Proc. of the MOCO'14*, Paris, France, 2014.

[10] B. Caramiaux. Optimising the Unexpected: Computational Design Approach in Expressive Gestural Interaction. *In Proc. of the CHI Workshop on Principles, Techniques and Perspectives on Optimization and HCI*, Seoul,Korea,2015.

[11] B. Caramiaux, M. Donnarumma, and A. Tanaka. Understanding Gesture Expressivity through Muscle Sensing. *ACM Trans. on Computer-Human Interaction*, 2, 6, 2015.

[12] B. Caramiaux, N. Montecchio, A. Tanaka, and F. Bevilacqua. Adaptive Gesture Recognition with Variation Estimation for Interactive Systems. *ACM TiiS*, 4, 4, 2015.

[13] F. Delalande. *La gestique de gould: Élements pour une sémiologie du geste musical*. In: G. Guertin (Eds.), Glenn Gould Pluriel, Québec: Louise Courteau, 1988, 85–111.

[14] J. Françoise. Gesture-Sound Mapping by Demonstration in Interactive Music Systems. *In Proc. of the 21st ACM MM'13*, Barcelona, Spain, 2013, 1051-1054.

[15] J. Françoise. Motion-sound mapping by demonstration. *Ph.D. Thesis*, Pierre and Marie Curie University, France, 2015.

[16] J. Françoise, B. Caramiaux, and F. Bevilacqua. A Hierarchical Approach for the Design of Gesture-to-Sound Mappings. *In Proc. of the CMC Conference*, Copenhagen, Denmark. 2012.

[17] J. Françoise, N. Schnell, R. Borghesi, and F. Bevilacqua. Probabilistic Models for Designing Motion and Sound Relationships. *In Proc. of the NIME'14*, London, UK, 2014.

[18] N. Gillian. Gesture Recognition for Musician Computer Interaction. *Ph.D. Thesis*, Queen's University Belfast, UK, 2011.

[19] R.S. Hatten. Musical Gesture: Theory and Interpretation. *Course note*s, Indiana University, 2003, http://www.indiana.edu/~deanfac/blfal03/mus/mus_t561_9824.html (Accessed 5 January 2016).

[20] P. Kolesnik, and M.M. Wanderley. Implementation of the Discrete Hidden Markov Model in Max/MSP Environment. *In Proc. of the FLAIRS*, 2005, 68-73.

[21] S. J. Koopman, N. Shephard, and J. A. Doornik. Statistical algorithms for models in state space using SelfPack 2.2. *Econometrics Journal*, 1, 1998, 1-55.

[22] S. Manitsaris, A. Glushkova, E. Katsouli, A. Manitsaris, and C. Volioti. Modelling Gestural Know-how in Pottery Based on State-space Estimation and System Dynamic Simulation. *Procedia Manufacturing*, 3, 2015, 3804-3811.

[23] B.H. Repp. Diversity and commonality in music performance: an analysis of timing microstructure in schumann's "traumerei". *Journal of the Acoustical Society of America*, 92, 1992, 2546-2568.

[24] C. Volioti, E. Hemery, S. Manitsaris, V. Tsekouropoulou, E. Yilmaz, F. Moutarde, and A. Manitsaris. Music Gestural Skills Development Engaging Teachers, Learners and Expert Performers. *Procedia Manufacturing*, 3, 1543-15, 2015.

[25] B. Zamborlin, F. Bevilacqua, M. Gillies, and M. D'inverno. Fluid gesture interaction design: Applications of continuous recognition for the design of modern gestural interfaces. *ACM TiiS*, 3, 4, 2014, 1-30.

[26] A.V. Zandt-Escobar, B. Caramiaux, and A. Tanaka. PiaF: A Tool for Augmented Piano Performance Using Gesture Variation Following. *In Proc. of the NIME'14*, London, UK, 2014.