The problem of musical gesture continuation and a baseline system

Charles Bascou
gmem-CNCM-marseille
first.last@gmem.org

Valentin Emiya
Aix-Marseille University
CNRS UMR 7279 LIF
first.last@univ-amu.fr

Mathieu Laurière
Paris Diderot University
UPMC CNRS UMR 7598 LJLL
mlauriere@math.univ-paris-diderot.fr

ABSTRACT

While musical gestures have been mapped to control synthesizers, tracked or recognized by machines to interact with sounds or musicians, one may wish to continue them automatically, in the same style as they have been initiated by a performer. A major challenge of musical gesture continuation lies in the ability to continue any gesture, without a priori knowledge. This gesture-based sound synthesis, as opposed to model-based synthesis, would open the way for performers to explore new means of expression and to define and play with even more sound modulations at the same time.

We define this new task and address it by a baseline continuation system. It has been designed in a non-parametric way to adapt to and mimic the initiated gesture, with no information on the kind of gesture. The analysis of the resulting gestures and the concern with evaluating the task raise a number of questions and open directions to develop works on musical gesture continuation.

1. THE PROBLEM OF MUSICAL GESTURE CONTINUATION

1.1 Musical gesture

From traditional acoustic instruments to modern electronic musical interfaces, gesture has always been a central problematic in musical performance. While acoustic instruments have to be continuously excited by energy impulsed by the performer's gestures, electronic instruments produce sounds without any mechanical input energy, which can last as long as electricity flows. In such electronic instruments, gestural interfaces have been used to control sound synthesis parameters. In [1], electronic instruments are defined by two components – the gestural controller and the sound production engine – and by the mapping of parameters between them.

The electronic performer can now deal with multiple layers of sound, mixed together as tracks on traditional digital audio workstation. Even if gestural control can be at a high level in the music system architecture – *e.g.*, on a mixing desk –, we often use several electronic instruments with performance-dedicated control strategies and interfaces. As they can't necessarily be played simultaneously, here comes the idea to design a system that continues a gestural behavior, primarily inputed by the performer on a particular instrument, and then automatically continued, while the performer can focus on other instruments.

Another motivation of such systems is the ability to define complex sound parameter modulations by gesture. From very simple Low Frequency Oscillators to chaotic systems, modulation methods are often parametric. One can use simple periodic/stochastic function or linear combination of these functions. This leads to very complex and rich results in terms of dynamics and movements but with a real pain on tweaking parameters. Indeed, these systems propose a lot of parameters, with complex interactions, making them really difficult to control intuitively. The idea to define modulation by gesture comes quite straightforward. Such a data-driven approach, as opposed to model-based parametric systems, leads to a system that could analyze an input gesture by means of its temporal and spatial characteristics, and then continue it *à la manière de*.

1.2 Problem characterization

Let us imagine an electronic instrument controlled by a tactile tablet. Consider gestures as isolated 2D strokes related to the contact of a finger on that tablet. An example of such a gesture is represented in black in Figure 1, together with a possible continuation of this gesture, in gray. This setting will be used throughout the proposed study and extensions to other settings will be discussed.

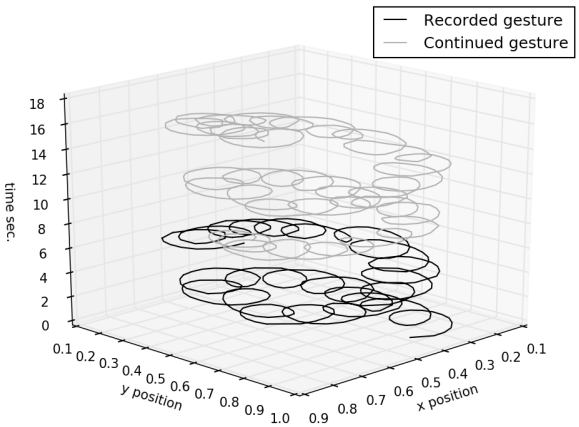

Figure 1. Example of a continuation (gray) of a performed 2D gesture (black). Space is on the x and y axes while time is on the z axis.

This work was partially supported by GdR ISIS project Progest and by the French National Research Agency (ANR), with project code MAD ANR-14-CE27-0002, Inpainting of Missing Audio Data. The authors would like to thank Pr. Franois Denis (LIF) for his fruitful scientific inputs.

We can formalize the problem with a traditional machine learning scheme composed of a learning phase (user performing the gesture) followed by a predicting phase (machine continuing the gesture). A gesture is a sequence of positions $\mathbf{p}(t) \in \mathbb{R}^2$ in cartesian coordinates for a single finger recorded at N discrete times t with $1 \leq t \leq N$. The goal of gesture continuation is to extrapolate a given gesture by generating the positions after the end of the available recording, *i.e.*, to estimate $\mathbf{p}(t)$ at times $t > N$.

The key issue is to study to which extent one may continue any gesture, with no *a priori* knowledge on the properties of the gesture. In particular, we want to avoid any categorization of gestures that would lead to, *e.g.*, parametric models of specific gestures. For instance, we are not interested in tracking periodic gestures to generate perfect loops, since the musical result would be too simplistic and is already well-used by say live-looping techniques, or to have a predefined list of allowed gesture patterns, which would dramatically reduce the performer's freedom. On the contrary, one may want, for instance: to capture the variability in the gesture of the performer, including when it is periodic ; to be able to continue aperiodic gestures that look like random walks; to reproduce the main characteristics of the gesture, including, at the same time, oscillating or random components – even if such structures do not appear in the sequence of positions, but in the velocity space for instance. Consequently, musical gesture continuation is not a well-posed problem. This important aspect should be considered when designing continuation systems as well as evaluation frameworks in order to keep in mind the ultimate goal – processing any gesture – and to avoid excessive simplification of the problem.

1.3 Related tasks

Part of the problem of musical gesture continuation is obviously related with the sound generation and mapping strategies involved in the electronic instrument in use. Indeed, gestures are completely dependent on the audio feedback, involving the need to study relations between sound and movement as in [2]. We chose for now to use a fixed reference electronic instrument, to work in the gesture domain only, and to tackle this question in future works.

Gesture continuation differs from other tasks that involve either gestures or continuation in music. In the gesture analysis field, gesture recognition [3, 4, 5] relies on reference gestures that are available beforehand and may be used to follow and align various media (sound, musical score, video) in live performance. Such reference gestures are not available in the generic gesture continuation problem. In [6, 7], the authors propose a system that can continue musical phrases and thus improvise in the same style. It works at a symbolic level (discrete segmented notes or sounds) and its application to continuous data (ie. gesture time series) is not straightforward.

1.4 Outline

This paper is organized as follows. In section 2, we propose a baseline system that has been designed to continue any arbitrary gesture, in the spirit of the open problem described above. A large place is dedicated to the evaluation of the results in section 3, including questions related to the evaluation methodology. We finally discuss a number of directions for this new problem in section 4.

2. A BASELINE SYSTEM BASED ON K-NEAREST NEIGHBORS REGRESSION

The proposed baseline system for musical gesture continuation is based on a simple regression scheme, as presented in section 2.1. In order to be able to continue any gesture, the system relies on the design of feature vectors discussed in section 2.2. The choice of the prediction function is finally detailed in section 2.3.

2.1 Overview and general algorithm

The proposed approach relies on the ability, at each time $t > N$, to generate the move $\boldsymbol{\delta}(t) \in \mathbb{R}^2$ from the current position $\mathbf{p}(t)$ to obtain the next one as $\mathbf{p}(t+1) = \mathbf{p}(t) + \boldsymbol{\delta}(t)$, by considering the past and current positions

$$\mathbf{x}(t) \triangleq \left[\mathbf{p}(t-\tau^o), \ldots, \mathbf{p}(t)\right] \tag{1}$$

where τ^o is a predefined memory length.

The proposed system is depicted in Algorithm 1. It relies on learning a prediction function (lines 1-7) which is then used to predict the moves at times $t \geq N$ (lines 8-13).

At each time t during both learning and prediction, a feature vector $\mathbf{v}(t)$ is computed from the current data point $\mathbf{x}(t)$ (lines 2-3 and 9-10).

The recorded gesture provides examples $(\mathbf{v}(t), \boldsymbol{\delta}(t))$ of the mappings between a feature vector $\mathbf{v}(t)$ and a subsequent move $\boldsymbol{\delta}(t)$ for $t \in \{1+\tau^o, \ldots, N-1\}$. Such a training set $\mathcal{S}$ is built at line 6. The prediction function $f_{\mathcal{S}}$ is obtained from $\mathcal{S}$ at line 7 by a supervised learning step. Once the prediction function is learned, gesture continuation is obtained in an iterative way for times $t \geq N$, by applying $f_{\mathcal{S}}$ to the feature vector $\mathbf{v}(t)$ in order to obtain the subsequent move (line 11) and the next position $\mathbf{p}(t+1)$ (line 12).

2.2 Feature extraction

In order to be as generic as possible, we consider simple features based on position, speed and acceleration along the gesture. Two options are proposed, in sections 2.2.1 and 2.2.2.

2.2.1 Instantaneous features

The simplest option consists in setting the characteristic memory length τ^o to 2 so that the point $\mathbf{x}(t) = \begin{bmatrix}\mathbf{p}(t-2)\\ \mathbf{p}(t-1)\\ \mathbf{p}(t)\end{bmatrix}$ considered at time t is composed of the last three positions. The feature vector is then defined as

$$\mathbf{v}^{\text{inst}}(t) \triangleq \begin{bmatrix}\mathbf{p}(t)\\ \mathbf{p}(t) - \mathbf{p}(t-1)\\ \mathbf{p}(t) - 2\mathbf{p}(t-1) + \mathbf{p}(t-2)\end{bmatrix} \in \mathbb{R}^6,$$

by concatenating the current position $\mathbf{p}(t)$, the instantaneous speed $\mathbf{p}(t) - \mathbf{p}(t-1)$ and the instantaneous acceleration $\mathbf{p}(t) - 2\mathbf{p}(t-1) + \mathbf{p}(t-2)$ computed in a causal way from point $\mathbf{x}(t)$.

Algorithm 1 Gesture continuation

Input(s):
recorded gesture $(\mathbf{p}(t))_{1\leq t\leq N}$
prediction length L
Output(s):
predicted gesture $(\mathbf{p}(t))_{N+1\leq t\leq N+L}$

Supervised learning on recorded gesture
1: **for** $t \in \{1+\tau^o, \ldots, N-1\}$ **do**
2: build point $\mathbf{x}(t) \leftarrow (\mathbf{p}(t-\tau^o), \ldots, \mathbf{p}(t))$
3: build feature vector $\mathbf{v}(t)$ from $\mathbf{x}(t)$
4: set move $\boldsymbol{\delta}(t) \leftarrow \mathbf{p}(t+1) - \mathbf{p}(t)$
5: **end for**
6: build training set $\mathcal{S} \leftarrow \{(\mathbf{v}(t), \boldsymbol{\delta}(t))\}_{t=1+\tau^o}^{N-1}$
7: learn regression function $f_{\mathcal{S}}$ from $\mathcal{S}$

Prediction
8: **for** $t \in \{N, \ldots, N+L-1\}$ **do**
9: build point $\mathbf{x}(t) \leftarrow (\mathbf{p}(t-\tau^o), \ldots, \mathbf{p}(t))$
10: build feature vector $\mathbf{v}(t)$ from $\mathbf{x}(t)$
11: estimate move $\boldsymbol{\delta}(t) \leftarrow f_{\mathcal{S}}(\mathbf{v}(t))$
12: set next position: $\mathbf{p}(t+1) \leftarrow \mathbf{p}(t) + \boldsymbol{\delta}(t)$
13: **end for**

2.2.2 *Finite-memory features*

Instanteous features may provide insufficient information to predict the next position, as it will be experimentally demonstrated (see section 3). An alternative choice is proposed by extending the memory length τ^o and by considering information at J different past lags t_j in the range $\{0, \ldots, \tau^o - 2\}$. We define the finite-memory feature vector

$$\mathbf{v}^{\text{mem}}(t) \triangleq \left[\mathbf{v}^{\text{inst}}(t-t_j)\right]_{j=0}^{J} \tag{2}$$

as the concatenation of several instantaneous feature vectors $\mathbf{v}^{\text{inst}}(t-t_j)$ taken at past times $t-t_j$ within the finite memory extension. In order to exploit the available information while limiting the feature vector size, the finite-memory data is sampled on a logarithmic scale by setting:

$$J \triangleq \lfloor \log_2(\tau^o - 2) + 1 \rfloor$$
$$t_0 \triangleq 0 \text{ and } t_j \triangleq 2^{j-1} \text{ for } 1 \leq j \leq J$$

where $\lfloor . \rfloor$ denote the floor function.

2.3 Prediction function

The desired prediction function maps a feature vector to a move in $\mathbb{R}^2$. Learning such a function from a training set $\mathcal{S}$ is a regression problem for which many well-known solutions exist, from the most elementary ones – *e.g.*, K-nearest neighbors regression, kernel ridge regression, support vector regression – to the most advanced ones – *e.g.*, based on deep neural networks. Since comparing all those approaches is not in the scope of this paper and since we target real-time learning and prediction, we use one of the simplest ones. The resulting system may serve as a baseline for future works and any other regression method may replace the proposed one in a straightforward way.

We use a K-nearest neighbors (KNN) regression approach based on a predefined number of neighbors K and the euclidian distance as the metric d in the feature space. Learning the regression function $f_{\mathcal{S}}$ from the training set composed of labeled examples $\mathcal{S} = \{(\mathbf{v}(t), \boldsymbol{\delta}(t))\}_{t=1+\tau^o}^{N-1}$ simply consists in storing $\mathcal{S}$ for further neighbor search. The KNN regression function is given by the Algorithm 2 and is used at line 11 in Algorithm 1. It first selects indices $(k_1, \ldots, k_K)$ of the K nearest neighbors of the current feature among the feature vectors $(\mathbf{v}_1, \ldots, \mathbf{v}_{N_{\mathcal{S}}})$ of the training set; and then define the predicted move $\boldsymbol{\delta}$ as the average of the related moves $(\boldsymbol{\delta}_{k_1}, \ldots, \boldsymbol{\delta}_{k_K})$.

Algorithm 2 KNN regression function $(f_{\mathcal{S}}(\mathbf{v}))$

Input(s):
training set $\mathcal{S} = \{(\mathbf{v}_k, \boldsymbol{\delta}_k)\}_{k=1}^{N_{\mathcal{S}}}$ with size $N_{\mathcal{S}}$
feature vector $\mathbf{v}$
number of neighbors K
distance d in feature space
Output(s): move $\boldsymbol{\delta} \in \mathbb{R}^2$

1: find the K-nearest neighbors of $\mathbf{v}$ in $\mathcal{S}$ as

$$\{k_1, \ldots, k_K\} \leftarrow \underset{\{k_1,\ldots,k_K\}\subset\{1,\ldots,N_{\mathcal{S}}\}}{\arg\min} \sum_{i=1}^{K} d(\mathbf{v}_{k_i}, \mathbf{v})$$

2: average moves of selected neighbors

$$\boldsymbol{\delta} \leftarrow \frac{1}{K} \sum_{i=1}^{K} \boldsymbol{\delta}_{k_i}$$

3. EXPERIMENTS AND EVALUATION

Designing an evaluation framework for gesture continuation is a complex topic for which we first raise a number of issues.

First of all, one should keep in mind the main goal – continuing subjective gestures with no a priori contents –, even as an unreachable objective; in particular, one should find how this can properly make part of an evaluation.

The space of musical gestures with no a priori may have a complexity much larger than what can be represented in some training and testing sets, and gathering a representative and statistically-consistent set of gestures may be a vain wish. This issue will impact the classical use of a training set (*e.g.*, for tuning parameters by cross-validation) as well as the validity of performance assessment on a testing set.

In terms of performance measure, one may not hope for a well-defined global score available beforehand. One may even consider it is a too challenging task to evaluate the quality of a predicted musical gesture by integrating complex aspects like: the intention of the gesture author and his or her subjective idea of what is a good continuation; differences in the audio rendering of various possible gesture continuations belonging to some kind of equivalence class, beyond a singular groundtruth gesture. In such conditions, one may characterize the main evaluation objective by the following two uncommon statements: the evaluation criteria may vary from one gesture to another; the evaluation criteria may be established by the performer at the time the

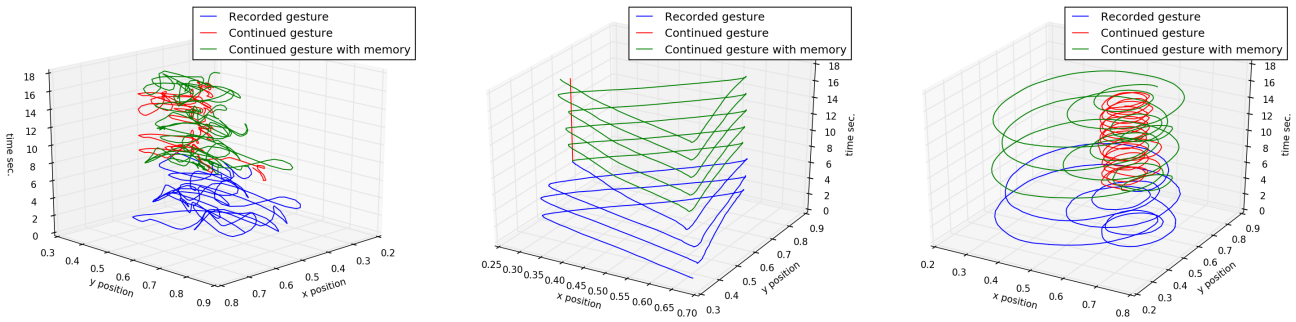

Figure 2. Initial gesture and two possible continuations for : a random-walk-like gesture (left); a triangle-shape gesture (middle) a periodic motion that alternates two small circles and a large one (right). The red and green gestures have been generated with the systems with instantaneous features only ($J = 0$) and with finite-memory ($J = 7$), respectively.

gesture is generated so that it may not be known at the time a continuation system is designed. In such context, one may develop an *a posteriori* multi-criteria evaluation and consider it as a principal evaluation method while usual *a priori* evaluation criteria may play a subordinate role.

Hence, the design of an evaluation framework for musical gesture continuation is an important challenge. We propose a substantial evaluation section which may be thought of as a first step in that direction.

Our experimental protocol is based on a corpus of recorded gesture to be continued. This set has been made with pluricity and maximum variability in mind, combining strong periodic forms to pseudo-random paths. A first set of 12 gestures was composed of basic geometrical shapes like circles, triangles, oscillations. A second set of 18 gestures has been made by a musician who was asked to specifically focus on the sound result. Gestures are sampled from a 2D tactile tablet at a rate of 60 points per seconds. The gesture data is sent to a custom sound synthesis software and stored as textfiles one line per point. Their length varies between 9.3s and 29.6s, the average being 18.8s.

We first analyze isolated gestures to provide a short *a posteriori* analysis in section 3.1. We then define an objective measure for prediction accuracy and apply it to evaluate the effect of finite-memory features in section 3.2. Finally, we propose a multicriteria evaluation framework in order to help the analysis of gesture continuation by pointing out a set of key evaluation criteria.

3.1 Three notable continuation examples

Figure 2 shows, for three gestures, their continuations by the proposed system with instantenous features only ($J = 0$) and with finite-memory features ($J = 7$, *i.e.*, about 2 seconds). The number of neighbors is fixed to $K = 5$. Those example have been chosen to illustrate the ability of the proposed approach to continue any gesture, as well as its limitations.

In the case of a gesture with a strong stochastic component (example on the left), both continuations show some ability to generate a similar stochastic behavior. It seems that finite-memory features help to reproduce a large variability including spatial spread and temporal cues. One may notice that short patterns of the initial gesture are locally reproduced. However, the system does not seem to be trapped in a loop, which would have had a strong negative impact on the perceived musical forms. From this point of view, the evaluation on random-like gestures may be compared to the topic of pseudo-random numbers generation where one tries to avoid any repetitions or period in the sequence of generated numbers.

The continuation of quasi-periodic gestures is illustrated in the other two examples. A recurrent failure case of the system with instantaneous features only is illustrated by the triangle-shaped gesture where speed is constant on edges and is null on corners (the user deliberately stops its movement for a while at each corner). The system is trapped in a corner since the K nearest neighbors have null-speed and the related move is also null. Finite-memory features provide information from the current point history to avoid such problems, as long as that the memory length is greater than the duration of stops. In the obtained (green) continuation, one may observe that the system succeeds in generating triangles, and that they present a variability similar to that of the original gesture.

Another common challenging situation is crosspoints in periodic motions, as illustrated in the third example. The gesture is a repetition of one large circle and two small circles successively. All three circles are tangent at their junction point, which generates an ambiguity since at that point, position, speed and acceleration are similar. Hence, at that position, the system with instantaneous features only is not able to determine whether it should enter a large or a small circle and gets trapped into the small circles here. On the contrary, the system with finite-memory features uses history information and is able to generate an alternation of one large circle and two small circles.

From these examples, one may note how gesture-dependent the evaluation is. Indeed for each example, specific evaluation criteria have been commented on, based on the property of the gesture as well as on the behavior of the system. This shows the importance of *a posteriori* evaluation. In a more synthetic way, one may also conclude from these examples that the proposed system is able to continue gestures of very different natures and that finite-memory features are useful to avoid typical failures. The subsequent sections will provide an extensive evaluation on this topic.

3.2 Prediction accuracy for various memory sizes

The memory size have a dramatic effect on the prediction results: one may wonder how large it should be set and how the system behaves when it varies. We propose to introduce and use an objective measure to assess the qual-

ity of the prediction at various horizons after the last point used to learn the gesture. This measure is subsequently used to analyze the prediction accuracy as a function of the memory size.

Let us consider a recorded gesture of total length $\widetilde{N}$: for clarity, $\widetilde{\mathbf{p}}(t)$ denote the recorded position for $1 \leq t \leq \widetilde{N}$. We denote by $N_0 < \widetilde{N}$ a minimum size considered for training. For a training size N such that $N_0 \leq N < \widetilde{N}$, the system is trained on the first N positions only and for $t > N$, $\widehat{\mathbf{p}}^{(N)}(t)$ denotes the position predicted by this system at time t. In such conditions, for any $n \leq \widetilde{N} - N$, position $\widehat{\mathbf{p}}^{(N)}(N+n)$ is the position predicted at a horizon n after the last position N known by the system. We define the mean prediction error at horizon n, $1 \leq n \leq \widetilde{N} - N_0$, by

$$\epsilon(n) \triangleq \frac{\sum_{N=N_0}^{\widetilde{N}-n} \left\| \widehat{\mathbf{p}}^{(N)}(N+n) - \widetilde{\mathbf{p}}(N+n) \right\|_2}{\widetilde{N} - n - N_0 + 1}. \quad (3)$$

In other words, for a fixed horizon n, $\epsilon(n)$ is the prediction error averaged among the predictions at horizon n obtained by training the system on different sizes N of training set, using the same recorded gesture.

Figure 3 shows the mean error averaged over all the gestures we considered, when fixing the number of neighbors to $K = 5$ and and the minimum training size to $N_0 = \frac{2}{3}\widetilde{N}$ (first two thirds of each gesture). The error increases with the horizon, since it is harder to make an accurate prediction when the horizon is large. Each curve can be split into two parts. During a first phase (memory size below 0.5 second), increasing the memory helps decreasing the error significantly. However, increasing the memory size beyond 0.5 second does not improve the prediction and sometimes drives up the error. These two trends (decreasing and then increasing) are found in most of the examples we considered, with different optimal memory sizes from one gesture to the other, and show that the proposed system has a limited capacity to learn from past points.

One may also note that this evaluation measure is not well suited for some gestures. For instance, if a gesture is made up of randomness, all possible realizations of this randomness are satisfying ways to continue it. As a consequence, a valid extrapolated gesture might be very far from the actual continuation made by the user. In this perspective, it appears useful to introduce other evaluation criteria.

3.3 Multicriteria evaluation

Evaluation may be thought within a multicriteria framework, relying on multiple evidence, by extending the use of objective performance measures. Since the criteria are not combined into a single score, this methodology is not dedicated to learn parameters or to rank concurrent systems. Generic multiple criteria may be used as a set of objective features that are automatically generated to help human interpretation or analysis.

We propose a set a evaluation criteria in order to compare a continued gesture $\widehat{\mathbf{p}}$ and a groundtruth continuation $\widetilde{\mathbf{p}}$. The experimental setting consists in splitting each gesture with a ratio $(2/3, 1/3)$ so that the first part is used for learning and is continued by the system to obtain $\widetilde{\mathbf{p}}$ while the second part is taken as the groundtruth $\widetilde{\mathbf{p}}$ for performance assessment ($\widehat{\mathbf{p}}$ and $\widetilde{\mathbf{p}}$ having the same length). The

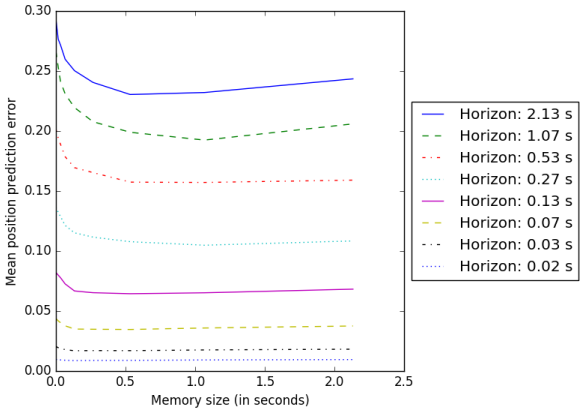

Figure 3. Mean prediction error averaged over all gestures for several prediction horizons n, as a function of the memory size J (n and J have been converted in seconds).

proposed criteria are based on instantaneous features in the continued gesture: position, speed and acceleration vectors, as well as their norms and angles. In each of those 9 possible cases, the feature of interest is extracted from the continued gesture $\widehat{\mathbf{p}}$ and from the groundtruth $\widetilde{\mathbf{p}}$ at each available sample time, resulting in two feature vectors to be compared.

A first family of criteria aims at analyzing the distribution of the instantaneous features. Distributions are considered by building histograms from the coefficients of feature vectors. For each feature, we compare the histogram for the continued gesture and that of the grountruth using a simple histogram difference measure, both histograms being computed on a common support with size $N_b = 25$ bins. Results are represented in the top part of Figure 4. In order to separate the gesture trajectory from its dynamics, a second family of criteria is proposed, based on dynamic time warping (DTW). DTW is used to align the continued gesture and the groundtruth, which cancels the effect of possible time stretching : the obtained distance measure quantifies only the difference in the trajectories, evaluating spatial cues only. Results are denoted by DTW/position in the middle part of Figure 4. As a possible extension, we also represent the DTW computed on the vectors of instantaneous speed, acceleration and speed norm instead of positions. Finally, since many gesture have one or several oscillating components – sometimes in position, speed, acceleration, and so on –, we also computed the Fourier transform of feature vectors. For each feature, spectra from the continued gesture and from the groundtruth are compared using the log-spectral distance and results are presented in the bottom part of Figure 4.

Results shown in Figure 4 confirm the advantage of finite-memory features in the proposed continuation system, since almost all criteria are improved on average. This multicriteria framework may also be used to detect gestures that are not well continued – *e.g.*, by automatically selecting gestures that are in the higher quartile – in order to draw a detailed analysis. As not all criteria are of interest for a given gesture, the performer may select them from this full dashboard, on a gesture-dependent basis, adopting an *a posteriori* evaluation.

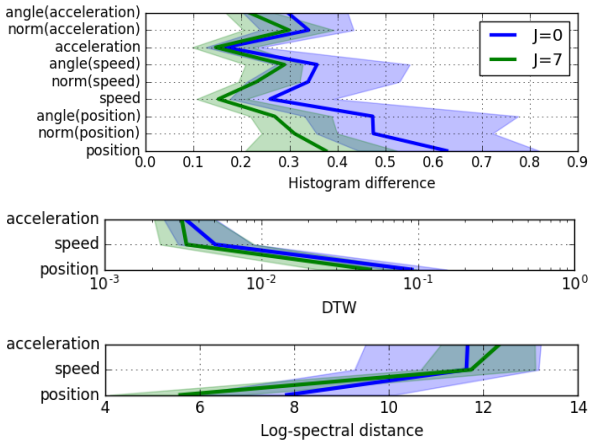

Figure 4. Multicriteria evaluation of the proposed system with instantaneous features only ($J = 0$, $K = 5$) and finite-memory features ($J = 7$, $K = 5$). Plain curves are median values among all gestures, with interquartile ranges as shaded areas. Various families of criteria are represented from top to bottom.

4. CONCLUSION AND PERSPECTIVES

We would like the main conclusion of this paper to be that the problem of musical gesture continuation, despite its vague definition, is not a vain or absurd task. To support this conclusion, we have shown that a system based basic features and KNN regression is able to continue any arbitrary gesture in an automatic way. We have also proposed the guidelines for an evaluation framework, including some particular considerations on specific gestures (null velocity issues, periodic and random components), a prediction accuracy measure and a large set of multicriteria objective measures that may be used in an *a priori* evaluation setting as well as for *a posteriori* evaluation. Those elements form preliminary contributions for works on musical gesture continuation, with several open directions.

The problem setting and the evaluation framework should go beyond the proposed ideas. 2D gestures may include multiple strokes generated simultaneously (*e.g.*, by several fingers) and sequentially (with arbitrary stops between strokes). They may also be extended to 3D gestures. The set of evaluation criteria may be completed by other features and comparison measure computed on the gesture itself, as well as criteria in the audio domain. This may also be the opportunity to analyze the relation between gesture and audio domains. Finally, subjective evaluation should also be considered and would first require the design of dedicated test protocols.

The proposed system for gesture continuation may be extended in some interesting directions. As shown in this paper, a significant improvement results from the extension of instantaneous features to finite-memory features. Adding more features may be even more useful to capture the right information, using feature selection method at training time. As a more fundamental issue, one may design or learn an appropriate distance in the feature domain while features are numerous and of different natures. We think that metric learning approaches would play an important role in order to have continuation systems that adapt to each gesture. One may also explore the wide range of possible non-parametric prediction functions. For instance, hidden Markov models may be successful to model the time dependencies as well as to control variations from the reference gesture as in [8].

Is the sky the limit? In many aspects, the problem of musical gesture continuation raises important questions about how to go beyond the limits we usually set for prediction tasks: how to deal with the dilemma of characterizing musical gestures with no *a priori*? How to address ill-posed problems as such? How to design systems when evaluation criteria are not known? Eventually, would such works be of interest to revisit conclusions from well-established tasks, as they may be questioned in [9]?

5. REFERENCES

[1] M. Wanderley and P. Depalle, "Gestural control of sound synthesis," *Proceedings of the IEEE*, vol. 92, no. 4, pp. 632–644, Apr 2004.

[2] A. Hunt, M. M. Wanderley, and M. Paradis, "The Importance of Parameter Mapping in Electronic Instrument Design," *Journal of New Music Research*, vol. 32, no. 4, pp. 429–440, 2003.

[3] G. Lucchese, M. Field, J. Ho, R. Gutierrez-Osuna, and T. Hammond, "GestureCommander: continuous touch-based gesture prediction," in *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2012.

[4] M. Takahashi, K. Irie, K. Terabayashi, and K. Umeda, "Gesture recognition based on the detection of periodic motion," in *Int. Symp. on Optomechatronic Technologies (ISOT)*. IEEE, 2010, pp. 1–6.

[5] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana, "Continuous realtime gesture following and recognition," in *Gesture in Embodied Communication and Human-Computer Interaction*, ser. LNCS. Springer Verlag, 2010, vol. 5934, pp. 73–84.

[6] F. Pachet, "The Continuator: Musical Interaction with Style," *Journal of New Music Research*, vol. 32, no. 3, pp. 333–341, 2003.

[7] G. Assayag and S. Dubnov, "Using Factor Oracles for machine Improvisation," *Soft Computing*, vol. 8, no. 9, Sep. 2004.

[8] B. Caramiaux, N. Montecchio, A. Tanaka, and F. Bevilacqua, "Adaptive Gesture Recognition with Variation Estimation for Interactive Systems," *ACM Trans. Interact. Intell. Syst.*, vol. 4, no. 4, pp. 18:1–18:34, Dec. 2014.

[9] B. L. Sturm, "A Simple Method to Determine if a Music Information Retrieval System is a Horse," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1636–1644, Oct 2014.