Segmenting and Parsing Instrumentalists' Gestures

Baptiste Caramiaux, Marcelo Mortensen Wanderley, Frédéric Bevilacqua

draft version - Segmenting and Parsing Instrumentalists' Gestures
B Caramiaux, MM Wanderley, F Bevilacqua
Journal of New Music Research 41 (1), 13-29, 2012

final version available at the JNMR website
http://www.tandfonline.com/doi/abs/10.1080/09298215.2011.643314

Segmenting and Parsing Instrumentalist's Gestures

B. Caramiaux[1], M.M. Wanderley[2], and F. Bevilacqua[1]

[1]UMR IRCAM-CNRS, Paris, France
[2]IDMIL, CIRMMT, McGill University, Montreal, Canada

Abstract

This article presents a segmentation model applied to musician movements, taking into account different time structures. In particular we report on ancillary gestures that are not directly linked to sound production, whilst still being entirely part of the global instrumental gesture. Precisely, we study movements of the clarinet captured with an optical 3D motion capture system, analyzing ancillary movements assuming that they can be considered as a sequence of primitive actions regarded as base shapes. A stochastic model called segmental hidden Markov model is used. It allows for the representation of a continuous trajectory as a sequence of primitive temporal profiles taken from a given dictionary. We evaluate the model using two criteria: the Euclidean norm and the lg-likelihood. We show that the size of the dictionary is not a predominant influence in the fitting accuracy and we propose a method for building a dictionary based on the log-likelihood criterion. Finally, we show that the sequence of primitive shapes can also be considered as a sequence of symbols enabling us to interpret the data as symbolic patterns and motifs. Based on this representation, we show that circular patterns occur in all players' performances. This symbolic step produces a different layer of interpretation, linked to a larger time scale, which might not be obvious from a direct signal representation.

1 Introduction

Physical gestures of a music performance are commonly called *instrumental gestures* [7, 29, 16]. In this context, instrumentalists' gestures that are not directly involved in sound production (or music production) are usually called *accompanying gestures* [7] or *ancillary gestures* [29]. In this article, we propose a methodology for clarinetist's ancillary gestures segmentation highlighting their inherent multi-level *information* content. This work is related to other recent studies on musical gestures, in particular instrumental gesture modeling ([11, 19, 9, 20, 25]) and computational models for investigating expressive music performance [32, 33]. This can give important perceptive insights for the design of virtual instruments, sound installations and sound design applications.

Ancillary gestures

Ancillary gestures are musical gestures [16] that are not related to sound production but convey relevant information about the player's expressivity during a performance. In [10], the author shows that the expressive intentions of musical performers are carried most accurately by their movements. This was later shown for the particular case of

clarinet performance [28]. Vines et al. explore how expressive gestures of a professional clarinetist contribute to the perception of structural and emotional information in musical performance. A main result is that the visual component carries much of the same structural information as the audio. Previously, Wanderley in [30] has investigated clarinetists' ancillary gestures providing findings that can be summarized as follows: the same player performing a musical piece several times tends to use the same gestures; some gestures related to structural characteristics of the piece tend to be similar across the performances of different musicians whilst others remain subjective. These results are specified in [31]. The authors show that clarinetists' subjective interpretation can be clustered according to which parts of the body are the most involved in ancillary gestures: some expressed in their knee and others used waist-bending gestures. Moreover, from the clarinetists' point of view, they show that the players feel uncomfortable when they try to consciously restrain their gestures whereas most of them seem to be aware of their expressive movements but not conscious of the gesture details. A recent study by Teixeira et al. [26] has highlighted movement patterns in the clarinetists' head by an empirical analysis and a qualitative segmentation. Results from these works highlight importance of ancillary gestures in communicating intention to the audience as well as understanding their expressive and spontaneous nature.

However, most of these works remain qualitative and do not propose quantitative methods for characterizing subjective gesture expressivity. One of the reasons is the problem of retrieving which part of the body (or which movement feature) is relevant for the analysis from high dimensional captured data often provided by a 3D full body motion capture system. Two main approaches can be used.

1. The *top-down approach* considers all the data and tries to find a subset of relevant features explaining the gesture. Usual techniques are PCA [13], Neural Networks [21], or CCA for cross-modal dimension reduction [8].

2. The *bottom-up approach* considers a restricted part of the movement selected by prior knowledge and shows that it can be used to make suitable assessments on gesture expressivity ([9, 20, 25]) .

In the scope of this paper, we follow the second approach in selecting a specific part of the captured elements, namely the instrument, that has been shown to be pertinent to characterize instrumentalists' ancillary gestures [30, 31].

Gesture as a sequence of primitive actions

Our basic hypothesis is that musical gestures can be considered as a sequence of primitive actions understood as primitive shapes. Previous works in cognitive sciences stated that people "make sense of continuous streams of observed behavior [like actions, music, ...] in part by segmenting them into events" [18]. Events can occur simultaneously at different time scales and according to a hierarchical structure.

In [15], the authors propose to adapt the linguistic framework to model human activity. The proposed human activity language consists of a three-level architecture: kinetology, morphology and syntax. Interestingly, kinetology "provides a symbolic representation for human movement that [...] has applications for compressing, decompressing and indexing motion data". Similarly, in activity recognition literature, a usual technique is to recognize actions defined as human motion units and activities defined as sequences of actions (see [27] for a survey).

In a recent paper, Godøy et al. [14] explore the theory that "perceived [music related] actions and sounds are broken down into a series of chunks in peoples' minds when they perceive or imagine music". In other words we holistically perceive series of both gesture and sound units: gesture and sound are cut into smaller units and the fusion and transformation of respective units lead to larger and more solid units.

We consider that music-related gestures, like ancillary gestures, can be described according to different timescales, meaning different segmentation levels (or *chunking* levels) like for the *human activity language* defined in [15]. Our aim is to provide a robust quantitative analysis technique, first, to represent the gesture signal into sequences of symbolic units (segmentation and indexing) and, second, to allow for further analysis of ancillary gestures taken as sequences of symbols (parsing). In this study, we will show that a trade-off has to be made between these two goals. This work is complementary to the previous work by Widmer et al. [32], showing that quantitative methods from machine learning, data mining or pattern recognition are suitable for the analysis of music expression and allow for retrieving the various structural scales in music. An important difference resides in the data used. Widmer et al. used MIDI like data while we use continuous data from a full-body motion capture system.

This paper is organized as follows. We first report previous work on human motion segmentation in the next section. Then we propose an overview of the general architecture of our methodology in section 3. The system is based on two main parts: first, the definition of a suitable dictionary for expressive gesture representation (in section 4); second, the stochastic modeling by SHMM that is formally presented in section 5. Section 6 details our experiments on a chosen database. First, we evaluate the pertinency of the model for representing the data using a geometric and a probabilistic criterion. Then, we show that it is suitable for motion pattern analysis of clarinetists' interpretation of a music piece. In section 7, we conclude and propose short-term prospective works.

2 Related work

Motion segmentation methods can be roughly categorized into either unsupervised or supervised techniques. Unsupervised segmentation algorithms do not need any prior knowledge of incoming signals. Barbič et al. [3] have shown that human behaviors can be segmented using simple methods like Principal Component Analysis (PCA), probabilistic PCA (PPCA) and Gaussian Mixture Model (GMM) that are only based on information available in the motion sequence. Changes in intrinsic data dimension (PCA, PPCA methods) or changes in distribution (GMM method) define segments' limits. Other methods use velocity properties of the joint angles [12] or perform implicit segmentation as a learning process [5]. In this last paper, Brand et al. use an unsupervised learning process on human motion to build *stylistic HMM* defined as a generic HMM (for instance describing bipedal human motion dynamics) changing according to a style parameter (for instance describing walking or strutting). This method allows complex human motions to be segmented and re-synthesized. However the internal states defining motion units are difficult to interpret and the method gives access to solely one timescale for movement description. More sophisticated methods like the nonparametric Bayesian process are used to model learning of action segmentation and its causal nature but are specific to goal-oriented actions [6].

The second set of algorithms are supervised techniques, where *primitives* (in a wide sense) attributed to the signal segments are known. Arikan et al. in [1] have proposed

a motion synthesis process based on a sequence of primitive actions (e.g. `walking–jumping–walking`) given by the user. It allows for higher-level control on motion synthesis but requires an annotated gesture database. Annotations are usually provided by the user and make sense for either action-oriented movement or activity synthesis. Our solution defines motion primitives as temporal profiles or shapes rather than words. An interesting model previously used for shape modeling and segmentation is the segmental hidden Markov model (SHMM). This model has been studied in different research fields: speech recognition [23], handwriting recognition [2] and time profile recognition of pitch and intensity evolution in [4]. SHMM allows continuous signals to be segmented and indexed at the same time. The system first needs a base shape dictionary used to describe input gestures as a sequence of basic shapes. Then a model defined as a sequence of shapes can be learned (or fixed) for recognition process. Hence, a character is a sequence of strokes [2] or a violin sketch like *tremolo* is a sequence of pitch shapes [4]. Our contribution is to show that the SHMM-based approach can efficiently represent clarinetists' ancillary gestures as a sequence of primitive shapes useful for the analysis of gesture patterns characterizing idiosyncratic player interpretations.

3 System overview

Figure 1 illustrates the general architecture of the methodology. It is specified for clarinetists' ancillary gestures but the methodology can be used for other kinds of gesture inputs like action-oriented human motion. Here, we focus on data captured by a 3D motion capture system that gives access to marker positions along the Cartesian axis, allowing the skeleton and the instrument movement to be reconstructed.

We assume a *prior knowledge* on the dataset that consists of a selection of relevant features for gesture representation and the definition of a set of primitive shapes (namely the base shape dictionary). This prior knowledge is based on previous work in the literature. It corresponds to the first two blocks in figure 1 and will be further detailed in the next section. Even if using a supervised segmentation technique, the methodology is modular and these blocks could be replaced by a learning process either supervised or unsupervised. For instance, from a suitable gesture parameterization, we could learn primitive shapes of the dictionary by specifying the number of shapes we require or by using previously annotated clarinetists' gestures.

The stochastic model is based on a segmental hidden Markov model that represents the input gesture signal from the database by the likeliest sequence of continuous and time-warped shapes taken in the dictionary. SHMM requires that both dictionary elements and input signal have the same representation. Interpretation allows the initial hypothesis to be validated, i.e. that clarinetist's expressive gestures can be represented as a sequence of meaningful primitive shapes. Interpretation consists of verification with recorded video and observation.

4 Gesture parameterization and dictionary

In this section we present the chosen gesture parameterization and the set of shapes composing a dictionary. This prior knowledge is based on previous work [30, 31, 24] on clarinetists' ancillary gesture analysis. We first select the gesture features; then we propose four base shape dictionaries that will be used in this paper. These dictionaries

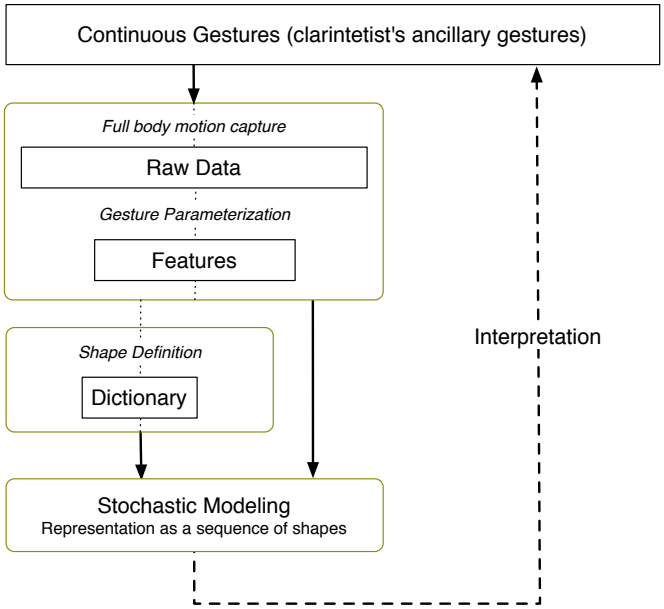

Figure 1: Architecture of our system. Continuous gestures are captured by a specified motion capture system. From the raw data, we define the gesture features (in this paper we consider 2 features) and build a base shape dictionary. A segmental HMM is applied on captured gesture based on the defined dictionary. Resulting gesture representation can be interpreted in terms of clarinetist's expressive movements.

are defined to be flexible enough to handle expressive gesture representation.

4.1 Features for clarinetist's gestures

Full body motion capture

As mentioned in the introduction, we suppose that we have access to marker positions from a 3D motion capture system. An absolute cartesian frame (x, y, z) is defined by the motion capture system calibration. From previous work (see [30], [31] and [24]), the bell's movements have been shown to convey relevant information about clarinetists' gestural expressivity. A local frame is defined in which we describe the movements of the bell. The origin of the local coordinate system is set to the reed marker, and we consider the vector drawn by the clarinet $(bell - reed)$ in the Cartesian frame.

Gesture parameterization

Because both the origin and the clarinet's dimensions are fixed, the clarinet's movements can be entirely defined by its angles in the relative spherical coordinates. Let $\mathbf{C}$ be a vector representing the clarinet in the cartesian coordinates system: $\mathbf{C}(x, y, z) = (bell - reed)(x, y, z)$. Transformation to a spherical system as depicted in figure 2 is denoted by: $\mathbf{C}(\rho, \theta, \phi)$. In the spherical coordinates, ρ is the radius, θ the azimuth angle and ϕ the inclination angle. Here ϕ is preferably called elevation angle. Since ρ is constant, we consider only θ, ϕ.

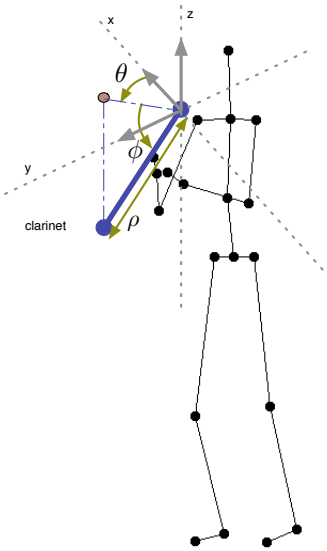

Figure 2: The clarinet's movement from cartesian to spherical coordinates system. The radius ρ is constant so we choose azimuth θ and elevation ϕ to describe the clarinet's bell movements.

4.2 Dictionary

A dictionary is a set of primitive shapes defining the basis on which an input gesture is decomposed. As mentioned above, each shape is parameterized by the azimuth and elevation angles θ, ϕ.

Defining segments

In this paper we consider four dictionaries of different sizes, as depicted in figure 3. The figure details all the shapes used to build each dictionary. Each shape in the figure describes the evolution of the bell's movements described by the evolution of the spherical coordinates (θ, ϕ). The first dictionary contains 8 primitive shapes that describe four equal parts of two circles: clockwise and counterclockwise directions. The second dictionary contains 44 primitive shapes. It generalizes the previous one and contains it. It takes into account the diagonals and intermediate trajectories between quarter-circles and diagonals. The third dictionary is an intermediate between dictionary 1 and 2. It contains 12 primitive shapes that include with quarter-circles and diagonals. The fourth dictionary also contains 12 primitive shapes: 8 from dictionary 1 plus vertical and horizontal trajectories.

Indexing primitive shapes

Each shape from the dictionaries is associated with an (integer) index. Figure 3 illustrates the four considered dictionaries and the shape indices. Globally, all the shapes shared by several dictionaries have the same index within these dictionaries. This simplifies the comparison of sequences of indices, obtained from distinct dictionaries. In the same way, we associate intermediate integer indices to intermediate shapes (e.g. between quarter-circle and diagonal in dictionary 2). Finally, dictionary 4 has four

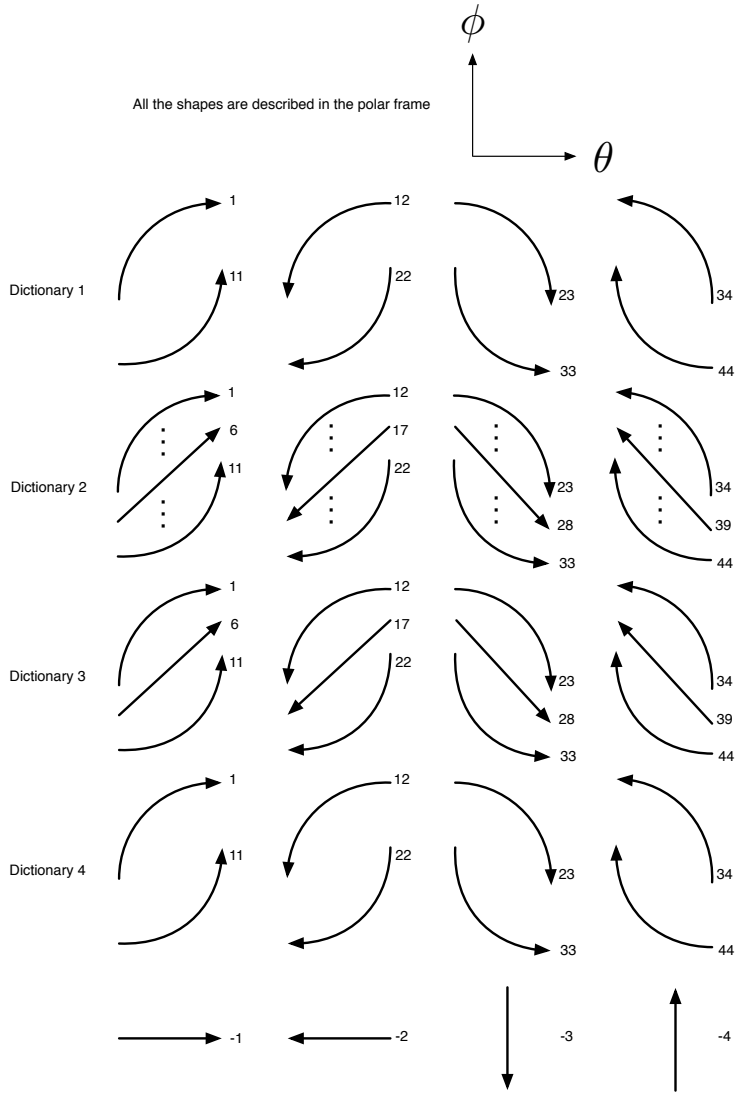

Figure 3: The four dictionaries defined for our study. From top to bottom: the first dictionary contains 8 elements corresponding to quarter-circles; the second dictionary also has the diagonal and shapes in-between quarter-circle and diagonal leading to 44 elements; the third one contains 12 elements which are quarter-circles and diagonals; finally the last one has 12 elements that are 8 elements from dictionary 1 plus horizontals and verticals.

shapes that are not included in the other dictionaries and not intermediate of previously defined shapes, so we chose to index them by negative values.

5 Stochastic modeling

We first present the theory of Segment Hidden Markov Models and then further discuss learning, preprocessing and inference.

5.1 Segment hidden Markov model (SHMM)

SHMM is a generative model used for shape-modeling. It generalizes classical HMM in the sense that emission probability density functions are defined at a segment level instead of at the sample level. Therefore, a segment generated by SHMM is allowed to follow a chosen regression form [17]. Figure 4 illustrates SHMM technique applied to an input signal curve. A uniformly sampled signal is taken as input of the segment model. An inference process returns the likeliest sequence of states (which correspond to elements of the dictionary) that fits the input signal and their respective durations.

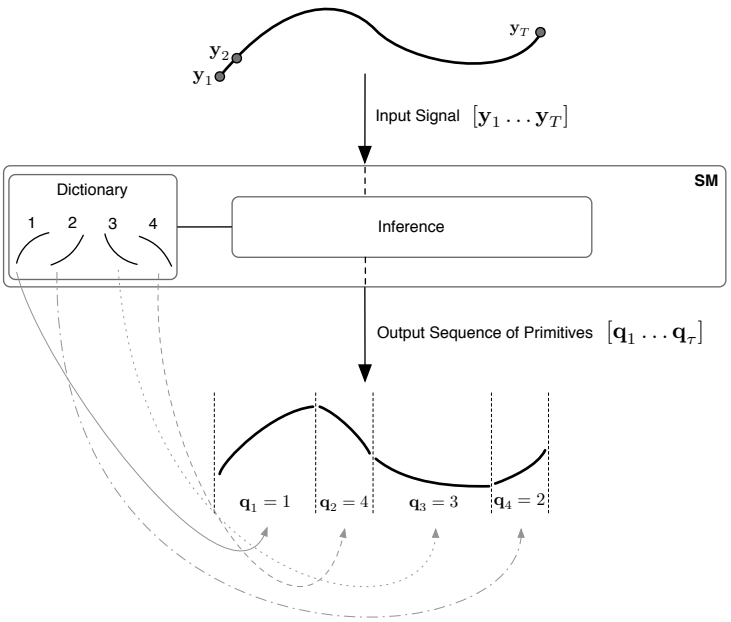

Figure 4: Illustration of the application of the segment model on a uniformly sampled input continuous curve. The inference process finds the likeliest sequence of states and their durations that generates the input signal. States are elements of the dictionary that can be time stretched to fit the input curve.

Formally, we denote $\mathbf{y} = [\mathbf{y}_1 \dots \mathbf{y}_T]$ the whole incoming feature vector sequence. A sub-sequence of $\mathbf{y}$ from t_1 to t_2 (with $t_1 < t_2$) is denoted $\mathbf{y}_{t_1}^{t_2} = [\mathbf{y}_{t_1} \dots \mathbf{y}_{t_2}]$ and its length is written $l = t_2 - t_1 + 1$. SHMM allows for representing $\mathbf{y}$ as a sequence of segments:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1^{l_1} & \mathbf{y}_{l_1+1}^{l_1+l_2} & \dots & \mathbf{y}_{\sum_i l_i}^{T} \end{bmatrix}$$

Each segment is of length l_i and we have $\sum_{i=1}^{\tau} l_i = T$ where τ is the number of segments inferred by SHMM to represent $\mathbf{y}$. Hence each SHMM state q emits a sequence $\mathbf{y}_{t_1}^{t_2}$ and a length $l = t_2 - t_1 + 1$ according to a density function $p(\mathbf{y}_{t_1}^{t_2}, l|q) = p(\mathbf{y}_{t_1}^{t_2}|l, q) \times p(l|q)$. The distribution of segment durations is $p(l|q)$, and the likelihood of the sequence is $p(\mathbf{y}_{t_1}^{t_2}|l, q)$. If we write $q_1^{\tau} = [q_1 \dots q_\tau]$ the sequence of states, taking values in a finite set $\mathcal{S}$, associated to the input sequence $\mathbf{y}$, and $l_1^{\tau} = [l_1 \dots l_\tau]$ the sequence of lengths, taking values in a finite set $\mathcal{L}$, the probability that the model

generates the input sequence $\mathbf{y}_1^T$ is:

$$p(\mathbf{y}_1^T|q_1^\tau) = \sum_{l_1^\tau} p(\mathbf{y}_1^T|l_1^\tau, q_1^\tau)p(l_1^\tau|q_1^\tau) \tag{1}$$

Where the sum is over all possible duration sequences. Considering that the segments are conditionally independent given the state and the duration and considering that the pairs $(state, duration)$ are themselves independent, we can rewrite the probabilities as:

$$\begin{aligned} p(\mathbf{y}_1^T|l_1^\tau, q_1^\tau) &= \prod_{i=1}^{\tau} p\left(\mathbf{y}_{l_{i-1}+1}^{l_i}|l_i, q_i\right) \\ p(l_1^\tau|q_1^\tau) &= \prod_{i=1}^{\tau} p\left(l_i|q_i\right) \end{aligned} \tag{2}$$

Figure 5 represents an unrolled SHMM as a graphical model where each arrow represents a conditional dependency. At the bottom is the generated sequence $\mathbf{y}_1^T$. Hidden layer states are: q_t (a segment index); l_t (the segment's duration); and X_t (a state from a segment emitting the observation $\mathbf{y}_t$).

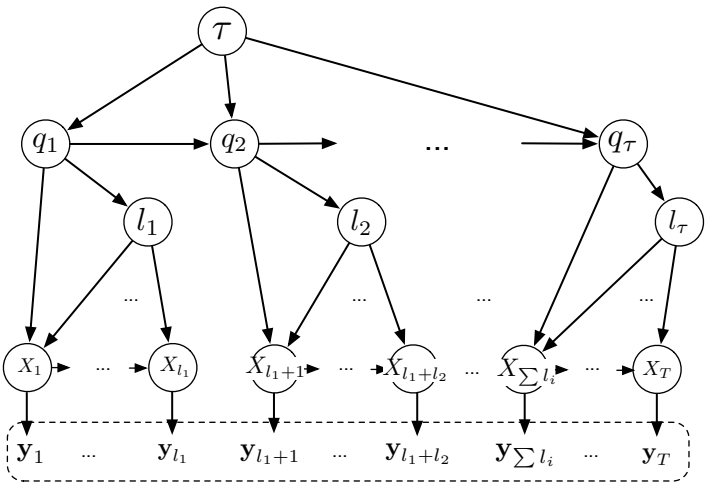

Figure 5: Graphical model of a Segmental Hidden Markov Model [22]. Arrows are conditional dependencies between variables. Inference is finding the likeliest q_1^τ and l_1^τ that generate $\mathbf{y}_1^T$. Hence duration l_i is dependent on state q_i and generated segment is dependent on the duration and the current state.

5.2 Learning and inference of SHMM

Learning SHMM

From the previous description, the hidden layer dynamics of SHMM can be modeled by three probability distribution functions:

1. State prior distribution: $\pi(i) = p(q_1 = s_i), \forall s_i \in \mathcal{S}$ or how the first shape of the sequence is chosen.

2. State duration distribution: $p(l_n|q_n = s_i)$, $l_n \in \mathcal{L}$ or how segment durations are weighted during inference.

3. State transition distribution: $p(q_{n+1} = s_j | q_n = s_i)$ denoted a_{ij}, $\forall (i,j) \in [1 \dots \tau]^2$ or how shapes in a dictionary are weighted during the inference.

Typical tools used in HMM framework for training can be used to learn SHMM parameters (e.g. Expectation–Maximization algorithm [23]). As mentioned in section 2, no well-defined ancillary gesture vocabulary can be used for training. Thus, the methodology adopted is to define prior dictionaries (section 4.2), then show that SHMM is relevant for ancillary gestures representation (section 6.2) and discuss the construction of *inter-* and *intra-* players gestures classes (section 6.4). For that purpose we use a generic configuration of SHMM based on uniform distributions:

1. $\pi(i)$ *uniform* means that any shape in the considered dictionary can be used as the first shape in the sequence.

2. $p(l|q_n = s_i)$ *uniform* means that shapes placed in the sequence can have any duration, each possible duration having the same weight. This choice is discussed later.

3. $p(q_{n+1} = s_j | q_n = s_i)$ *uniform* means that for each shape, any shape placed afterwards has the same weight (the so-called *ergodic* model).

Inference

Inference is the estimation of the likeliest state sequence that has emitted the input gesture data. It means estimating the number of segments $\hat{\tau}$, the segment sequence $\hat{q_1}^{\hat{\tau}}$ and the corresponding length sequence $\hat{l_1}^{\hat{\tau}}$. This can be done by finding the arguments maximizing the likelihood function defined by equation (1), that is:

$$(\hat{\tau}, \hat{q_1}^{\hat{\tau}}, \hat{l_1}^{\hat{\tau}}) = arg \max_{\tau, q_1^\tau, l_1^\tau} \sum_{l_1^\tau} p(\mathbf{y}_1^T | l_1^\tau, q_1^\tau) p(l_1^\tau | q_1^\tau) \tag{3}$$

As previously mentioned, transition probability distribution and duration probability distribution are uniform. Here, we define the observation (or emission) probability distribution. An input signal is represented as a bi-dimensional time series, uniformly sampled, representing the evolution of azimuth and elevation angles θ, ϕ. The incoming sequence of observations $\mathbf{y}_1^T = [\mathbf{y}_1 \dots \mathbf{y}_T]$ is defined such that:

$$\mathbf{y}_t = \begin{bmatrix} \theta_t \\ \phi_t \end{bmatrix}$$

Emission probability is defined as:

$$p\left(\begin{bmatrix} \theta_{l_{i-1}+1}^{l_i} \\ \phi_{l_{i-1}+1}^{l_i} \end{bmatrix} | l_i, q = s \right) \propto \exp\left(- \sum_{j=l_{i-1}+1}^{l_i} \frac{\left\{ [\theta_j - \theta(s)_j]^2 + [\phi_j - \phi(s)_j]^2 \right\}}{2\sigma^2} \right) \tag{4}$$

Where $\theta(s)_j$ (respectively $\phi(s)_j$) is the value of the first coordinate (respectively the second) of shape s at time j; and σ is the gaussian standard deviation. Exact inference is made using the forward-backward Viterbi algorithm. For an observation sequence of length T, it takes $O(MDT^2)$ where M is the number of primitives in dictionary, D is the maximum length of possible durations. Hence, doubling the number of elements in a dictionary implies doubling the computation time. It can be a criterion for selecting a dictionary among the others.

Preprocessing and resynthesis

While processing the inference, each segment to be compared to shapes in a dictionary is normalized to $[0, 1]$ meaning that both azimuth and elevation signals are translated by their minimum and scaled by their maximum. Let us consider the i-th segment; we define an offset coefficient

$$\min_{l_{i-1}+1 \leq j \leq l_i} (\begin{bmatrix} \theta_j \\ \phi_j \end{bmatrix})$$

as well as a scaling coefficient

$$\max_{l_{i-1}+1 \leq j \leq l_i} (\begin{bmatrix} \theta_j \\ \phi_j \end{bmatrix})$$

These coefficients are stored for the resynthesis process. During resynthesis, for each shape taken sequentially in the inferred sequence by SHMM, we scale the whole shape by the scaling coefficient and translate the scaled shape by the offset coefficient.

6 Results

In this section we first present the material used for the analysis. Then, we inspect the accuracy of the model with respect to the considered dictionary. Finally we show how the resulting sequence of index can be parsed and what kind of information it gives us for characterizing and analyzing ancillary gestures.

6.1 Database

The database used for experiments was recorded at the Input Devices and Music Interaction Laboratory (IDMIL) at McGill University, Montreal, Canada. From the whole database we have selected four clarinetists playing the first movement of the Brahms *First Clarinet Sonata Opus 120, number 1*. The task was to interpret the piece four times in a neutral way. The set-up was as follows: The clarinet sound was recorded using an external microphone, and a video camera was used to record all the players' performances. Clarinet movements were recorded using a 3D motion capture system. Two of the markers were placed on the clarinet, one the reed and one the bell.

In order to cross-analyze players' performances, we need to align each performance to a single reference. The common reference is the score of the first movement of the Brahms *sonata*. First we build a synthetic interpretation: the score is translated into a pitch time series with a fixed sample rate (see figure 6). Synthetic pitch evolution corresponds to the piece played regularly following a tempo of 100. The pitch evolutions of the subjects' performances are pre-processed to be discrete and aligned to the synthesized signal using the Dynamic Time Warping (DTW) technique. Since audio and gesture streams are recorded synchronously, the resulting alignment is also applied to gesture data.

An example of one performance by each clarinetist is depicted in figure 7. Solid black lines represent azimuth signals, and dashed black lines represent elevation signals. Each signal has been centered to have a zero mean and aligned on the reference interpretation.

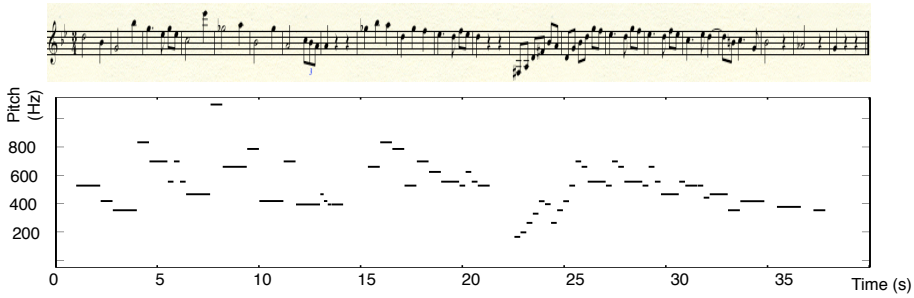

Figure 6: Synthetic interpretation. The figure is the pitch signal (piano roll like) of an interpretation of Brahms' *First Clarinet Sonata Opus 120, number 1* played at tempo 100.

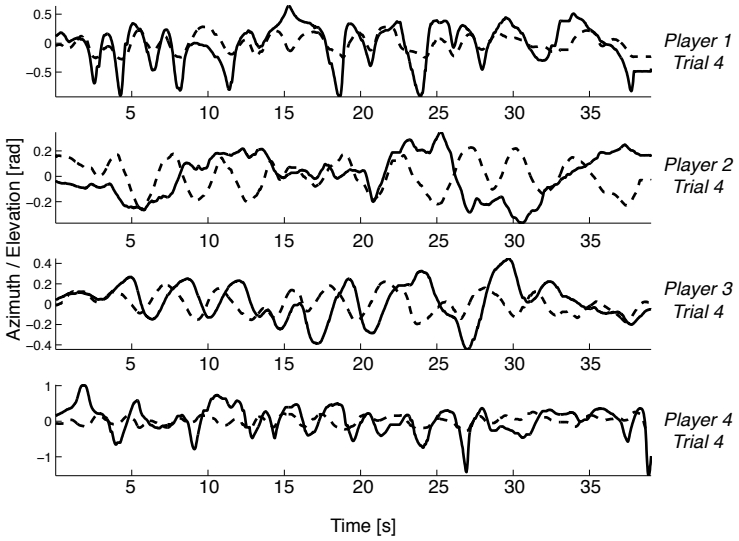

Figure 7: Examples of signals for one interpretation by players 1, 2, 3, 4. Solid lines are azimuth signals θ_t, and dashed lines represent elevation signals ϕ_t

6.2 Ancillary gesture as a sequence of base shapes

To clarify the reading, let us take the example of a part of player 4's fourth performance from 17 seconds to 22 seconds (the whole signal is plotted at the bottom of figure 7). We denote $\mathbf{y}_{t_1}^{t_2}$ where $t_1 = 17$ and $t_2 = 22$, the sequence of observations taken as input for SHMM. Considering the dictionaries 1 and 2, the model inferred a sequence of shapes per dictionary. Figure 8 shows the results: on the upper part are the results obtained with dictionary 1; on the lower part are those obtained with dictionary 2. For each, two plots are depicted: at the top is the azimuth angle ($\theta_{t_1}^{t_2}$); and at the bottom, the elevation angle ($\phi_{t_1}^{t_2}$). Dashed lines are the original angle time series and gray solid lines are the sequences of shapes inferred by SHMM for both dictionaries. Bracketed integers are the shape indices from the considered dictionary.

Intuitively, a more exhaustive dictionary should better represent a given continuous

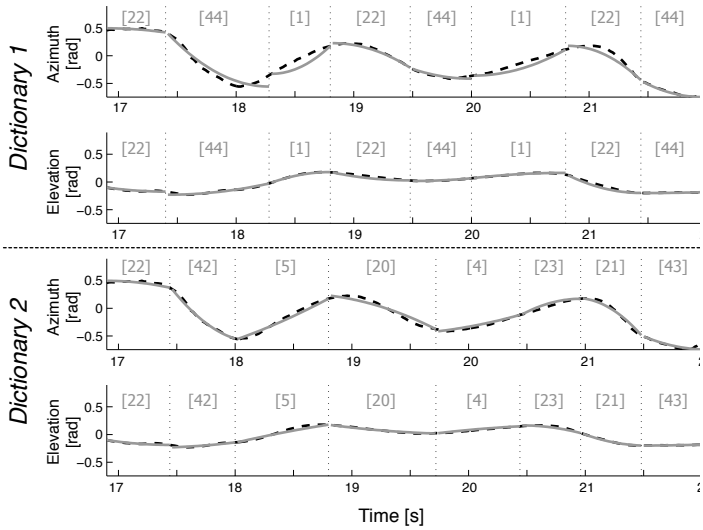

Figure 8: Examples of resynthesized curves. Upper part presents the input signal and the resynthesized signal using dictionary 1. Dashed lines are original curves (azimuth at the top, elevation at the bottom), and piecewise gray lines are the shapes inferred by SHMM. Similarly, the lower part reports the result using the second dictionary

multidimensional curve. This can be observed in figure 8 around 18.5sec. Let us consider dictionary 1, the likeliest shape for the input signal around 18.5sec is the shape indexed by 1, and it better matches the elevation signal than the azimuth signal. In dictionary 2, the result reveals that the likeliest shape to model this part of the signal is shape 5 that does not belong to dictionary 1. This shape is an intermediate shape between 1 and the diagonal 6. On the other hand, this simple example shows a more regular sequence of indices (bracketed integers) on the upper part of figure 8 than on the lower part. Hence, a more exhaustive dictionary seems to provide a more varying sequence of symbols.

This example illustrates how a gesture input is represented by a sequence of shapes taken from a dictionary and shows that differences appear according the choice of the dictionary. Firstly, we want to generalize the approach by systematically evaluating the accuracy of such a representation through other input gestures from the database and the available dictionaries. Secondly, we want to qualitatively analyze the sequences of symbols provided by the model according to the different dictionaries.

6.3 Evaluation of the model

SHMM infers the likeliest sequence of shapes together with the likeliest sequence of the shapes' durations. The concatenation of the inferred time warped shapes offers a new representation of the input signal (as presented in the schematic view figure 4). Here, we inspect how accurate is the re-synthesis in representing ancillary gestures from real performances of clarinetists. To do so, we propose two evaluation criteria: the Euclidean norm and the log-likelihood. As mentioned before, distributions involved in the model are uniform. The possible durations are from 0.1sec to 1.3sec (step of 0.1sec) and the standard deviation used is $\sigma = 0.1$ radian.

Evaluation using the Euclidean norm $\|.\|_2$

As mentioned in section 6.1, the database contains four natural interpretations of the Brahms *sonata* by four players. We evaluate how these gestures are represented according to each of the four dictionaries defined earlier. An Euclidean distance is computed between the resynthesized signal and the original one. Therefore, one distance value per interpretation is returned. We average over the interpretations so that one value remained per player. Results are reported in table 2 showing means and standard deviations.

	Dictionaries (#elements)			
$\times 10^{-3}$	1 (8)	2 (44)	3 (12)	4 (12)
Player 1	1.41 ± 0.18	0.84 ± 0.12	0.84 ± 0.12	1.40 ± 0.19
Player 2	0.37 ± 0.07	0.19 ± 0.02	0.21 ± 0.01	0.37 ± 0.07
Player 3	0.39 ± 0.07	0.18 ± 0.03	0.24 ± 0.01	0.39 ± 0.07
Player 4	1.28 ± 0.37	1.07 ± 0.51	1.04 ± 0.38	1.33 ± 0.31

Table 1: Averaged Euclidean distance with standard deviation. Values are reported in 10^{-3} radians. Lowest values correspond to better fitting between the resynthesized signal and the original one.

In this table, lowest values mean better fitting. Globally, the results show that the model efficiently fits the incoming data: the maximum mean value is 1.41×10^{-3} radians (player 1, dictionary 1) corresponding to the mean quantity of variation between the two curves. Moreover standard deviations across interpretations are very low, meaning that there are not important variations intra-participant [30]. For players 1, 2 and 3, lowest scores are obtained for dictionary 2 although they are very close to the scores obtained with dictionary 3. Dictionary 1 and 4 return the same scores and a close look at inferred sequences reveals that SHMM returns the same sequence of shapes for both dictionaries, meaning that the vertical and horizontal shapes in dictionary 4 are not involved in the inferred sequence. The case of player 4 is singular because standard deviations are high and the dictionaries can not be statistically discriminated: the Student's t-test shows that the mean obtained for dictionaries 1 to 4 are not significantly different at the 5% significance level. To conclude, according to the Euclidean norm, dictionary 2 and 3 are equivalent as well as dictionaries 1 and 4. Hence the number of elements in a dictionary is not necessarily linked to a better fitting.

Evaluation of the temporal prediction score

We compute the log-likelihood $\log p(\mathbf{y}_1^T | l_1^\tau, s_1^\tau)$ for the observation time series $\mathbf{y}_1^T$ (or test data). This score refers to the likelihood assigned to the incoming observation by the model [17]. Higher scores mean that the model is likely to generate the test data. In other words, the model has a better predictive power. In order to be length independent, we normalize by the length of the observation sequence. The average log-likelihood values are computed for each subject and the results are given in table 2.

The results show that the highest log-likelihood scores are obtained with dictionary 2, meaning that the probability to generate data from the Brahms database is higher with dictionary 2 than either with dictionary 1, 3 or 4. Dictionary 2 is more exhaustive than the other three, and table 2 shows that the scores obtained with dictionary 3 are significantly better than with dictionary 1. An interpretation is that the more exhaustive a dictionary is, the better it is to generate (and consequently to predict) the observed

	Dictionaries (#elements)			
	1 (8)	2 (44)	3 (12)	4 (12)
Player 1	-1.335 ± 0.023	-0.643 ± 0.058	-0.743 ± 0.063	-1.353 ± 0.023
Player 2	-1.294 ± 0.041	-0.641 ± 0.028	-0.712 ± 0.047	-1.309 ± 0.042
Player 3	-0.907 ± 0.148	-0.481 ± 0.033	-0.584 ± 0.037	-0.920 ± 0.148
Player 4	-1.163 ± 0.087	-0.662 ± 0.056	-0.748 ± 0.056	-1.179 ± 0.085

Table 2: Cumulative Log-likelihood ($\sum_{t=1}^{T} \log(p(\mathbf{y}_1, ..., \mathbf{y}_t | q_1^{\tau})))$ averaged over the interpretations and the standard deviations. The criterion returns the likelihood that the model generates the observation data. Highest values correspond to better perdiction.

data. This is in contrast with the results based on Euclidean distance: we add *information*[1] from dictionary 3 to dictionary 2 by adding new primitive shapes whereas it does not affect how the reconstructed curve fits the input curve. This will be refined in section 6.4. However, considering the scores obtained with dictionary 4, a t-test (using a 5% significance level) shows that they are not significantly different from the scores with dictionary 1 (similar to the evaluation using the Euclidean norm) even if dictionary 4 contains more elements than dictionary 1.

This can be explained as follows. As mentioned in section 5, the log-likelihood is based on the observation likelihood and the transition probability. Therefore, adding elements in a dictionary: increases the observation likelihood as long as they better fit the observations (dictionary 2, 3); decreases their transition probability (uniform over the elements of the dictionary) (dictionary 2, 3 and 4)). Hence, a dictionary can be tested using this criterion as follows: starting from a simple dictionary (e.g. dictionary 1), one can add element by element and inspect the resulting log-likelihood. If the score is decreasing, it means that the the fitting criterion is stronger than the decreasing weight. Otherwise, the added element is not relevant for the input observations.

6.4 Parsing the sequences of indices

The sequence of shapes can be analyzed as a sequence of symbols (integer indices). The motivation is to observe the real-world data set of ancillary gesture signals at a higher level than the *shape level* presented in the previous section. A symbolic representation of continuous signal allows the retrieval of patterns and motifs based on *parsing* processes: the continuous signal can be considered as a string. There exists a large set of methods for *string* analysis (from machine learning, pattern recognition, theoretic computer science and so forth). Here we propose a qualitative interpretation by considering ancillary gestures as strings that could be useful for future research in this field. We discuss how the design of a dictionary can determine the pattern analysis.We start by considering the first dictionary as an example, and we compare to the results obtained with dictionaries 2 and 3 (in this section dictionary 4 is discarded since it does not add relevant information).

From shape level to pattern level

In this section, we inspect how the shapes and their durations are temporally distributed in the signal. Figure 9 presents the data for player 3's second interpretation. At the top,

[1] *Information* is to be understood as the accuracy of the model to generate the input data

the temporal evolution of the pitch is represented as a continuous signal and zero values mean silences. Below, we report both the azimuth signal (solid line) and elevation signal (dashed line). Finally, at the bottom the sequence of shapes' index inferred by the stochastic model is plotted as a piecewise constant signal with circles indicating the starting and ending points of the shape. It appears that short-duration shapes are concentrated around 2 seconds (62.5% of shapes have durations lower than 500ms), 13 seconds (80% lower than 500ms), 24 seconds (75% lower than 500ms) and 33 seconds (66.7% lower than 500ms). In-between these instants, sequences of longer duration define a specific recurrent pattern (so-called *motif*) 1–23–22–44 that is highlighted by gray lines. A more detailed study of motifs will be given afterwards.

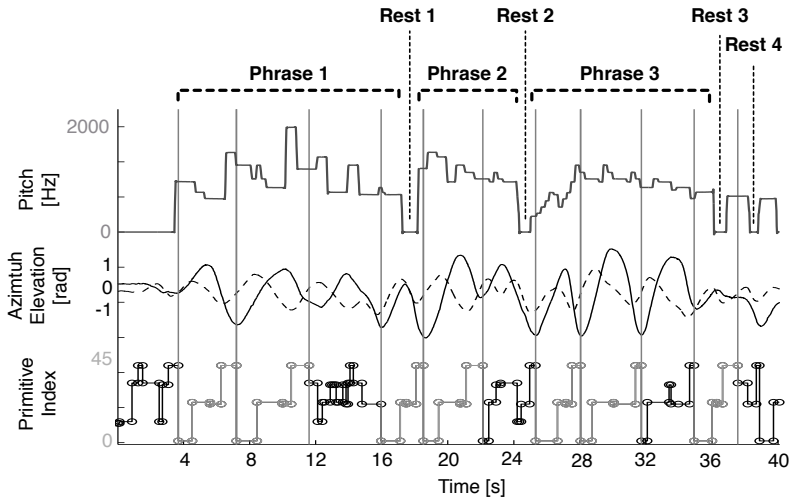

Figure 9: Player 3's second interpretation. At the top we report the pitch evolution corresponding to this specific interpretation. In the middle, the solid line represents the bell's azimuth evolution while the dashed line draws the bell's elevation. At the bottom, we report the segmentation obtained by SHMM using the first dictionary. We added indications about the Brahms sonata: phrases 1, 2 and 3 as well as the various rests.

Three parts are of particular interest: $[12, 16]$ seconds (part 1), $[22, 25]$ seconds (part 2) and $[32, 35]$ seconds (part 3). Part 1 is in-between two identical patterns defined by the sequence 1–23–22–44. Our hypothesis is that it defines an articulation between two phrases separated by a rest. This is reinforced by the second part ($[22, 25]$ seconds): it corresponds to the articulation between the second phrase and the third one which starts by the same pattern 1–23–22–44. Finally, part 3 seems to be of different nature. Since the global pattern is very similar to the recurrent pattern 1–23–22–44, the short-duration shapes detected in the middle of the pattern seem to be due to "noise" (i.e. unwanted perturbation that is not specific to the interpretation). Let us analyze further this assumption by discussing figure 10. The figure depicts the same performance as in figure 9. Pitch signal has been omitted for readability. On the upper part, we report the beginning and the end of the gesture as well as the three parts described above. On the lower part, we depict each occurrence of the pattern 1–23–22–44.

The seven occurrences of the pattern 1–23–22–44 correspond to a circle performed in counterclockwise direction and are mostly occurring during musical phrases. The second and the last circles clearly include the beginning of the next shape: it gives a trace of the articulation between shapes at a fine temporal scale. On the other hand, if

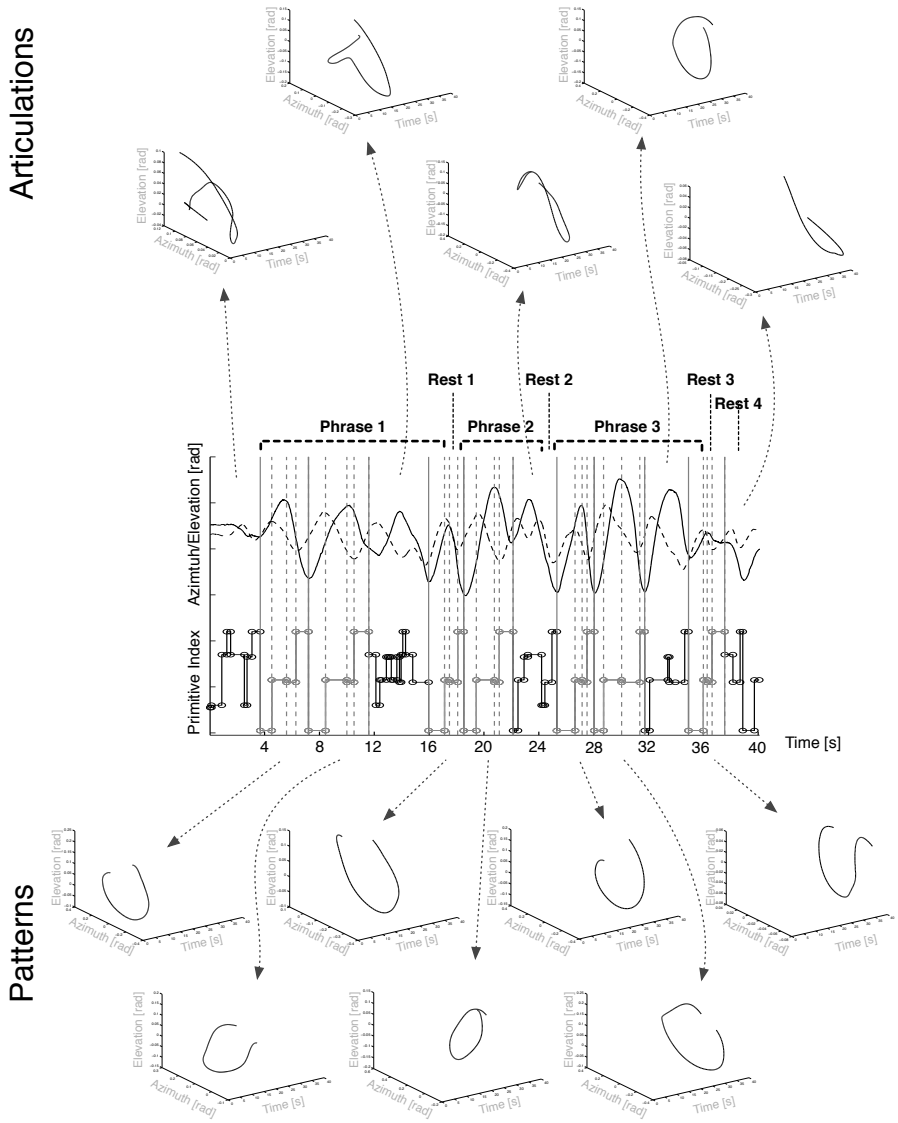

Figure 10: Player 3's second interpretation. This figure shows the azimuth and elevation signals (respectively solid and dashed lines) and the segmentation inferred by SHMM using the first dictionary. Patterns are in gray, and articulations are in black. We explicitly plot the bell's original movement for these particular parts. Patterns are clockwise circles, with idiosyncratic movements in between them. Patterns can be retrieved using the symbolic representation though it is not clear from only the original signals.

considering the three parts of interest described above, part 1 (fig. 10, $[12, 16]$sec) and part 2 (fig. 10, $[22, 25]$sec) reflect a different behavior: an abrupt change of direction

and the articulation between this change and the beginning of the next circle. Part 3 (fig. 10, $[32, 35]$sec) corresponds to a circle performed counterclockwise supporting our previous guess that short-duration shapes in this specific part are due to noise and do not reflect articulations between different behaviors. This study reveals that occurring patterns can be identified using the symbolic representation and highlights higher temporal structure in ancillary gestures (patterns – articulations). Interestingly, this could be handle in the SHMM by defining higher level models constituted by the sequences defining the patterns. Consequently, a non-ergodic model would optimize the recognition of such patterns.

Identifiable motifs

In the previous example, a clockwise circular pattern is repeated seven times and is described as the sub-sequence of indexes 1–23–22–44 in the first dictionary. Here we show that a symbolic representation gives a powerful tool for exact motif retrieval in the continuous gesture signal. A quick enumeration of recurrent patterns in the data from players' interpretations shows that the previously introduced sequence 1–23–22–44 occurs 38 times, mostly in interpretations by player 3 (25 occurrences) and 4 (12 occurrences) than player 1 (1 occurrence) and player 2 (0 occurrences). The pattern begins by shape 1 drawing the circle from the leftmost point. A circular permutation of the pattern index means starting the circle at a different position. A quick study shows that starting from the left point creates the pattern 22–44–1–23 occurring 42 times: 32 occurrences for player 3 and 9 occurrences for player 4.

Figure 11 reports the two symbolic sequences corresponding to the inferred shapes for the fourth interpretations by both players 3 and 4. The co-occurring pattern 1–23–22–44 is emphasized by gray frames: 8 occurrences for player 3 and 5 occurrences for player 4.

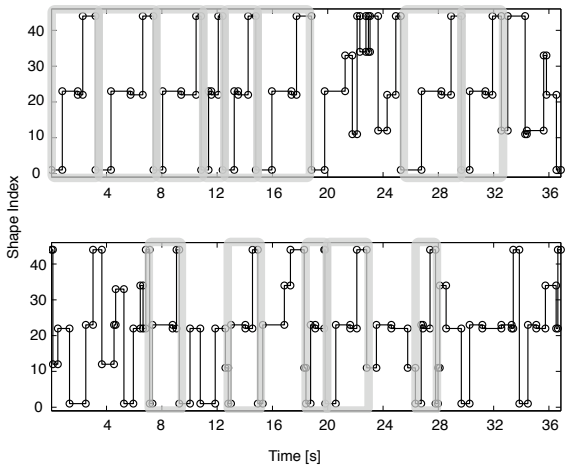

Figure 11: At the top we can see the segmentation of player 3's fourth interpretation. Patterns 1–23–22–44 are highlighted by gray frames. At the bottom, we can see the segmentation of player 4's fourth interpretation. Same patterns are similarly highlighted. 1–23–22–44 pattern is particularly found in the interpretations by these two players.

A similar analysis shows that the pattern 11–34–12 occurs 19 times across all the

interpretations and specifically in player 1's interpretations (12 occurrences) and players 2's interpretations (7 occurrences). This pattern consists in three quarter-circles in counterclockwise direction, that is, the opposite strategy from players 3 and 4. Hence, every clarinetists make use of circles in their interpretations.

Motif-based dictionary information

From previous sections, a dictionary can hold different information contents: number of elements and/or accuracy of these elements for the specific input. In this section, we analyze the influence of a chosen dictionary on ancillary gesture motif recognition. Let us focus on pattern 1–23–22–44 in player 3's interpretations. Figure 12 reports resynthesized patterns in (θ, ϕ) coordinates for dictionaries 1, 2 and 3 together with the original input trajectories. The first line illustrates the patterns inferred by SHMM with dictionary 1. Above each reconstructed pattern, we report the corresponding sequence of indices. The second line (resp. the third) illustrates patterns inferred with dictionary 2 (resp. 3). Circle patterns corresponding to the index sequence 1–23–22–44 are marked by dashed rectangular boxes.

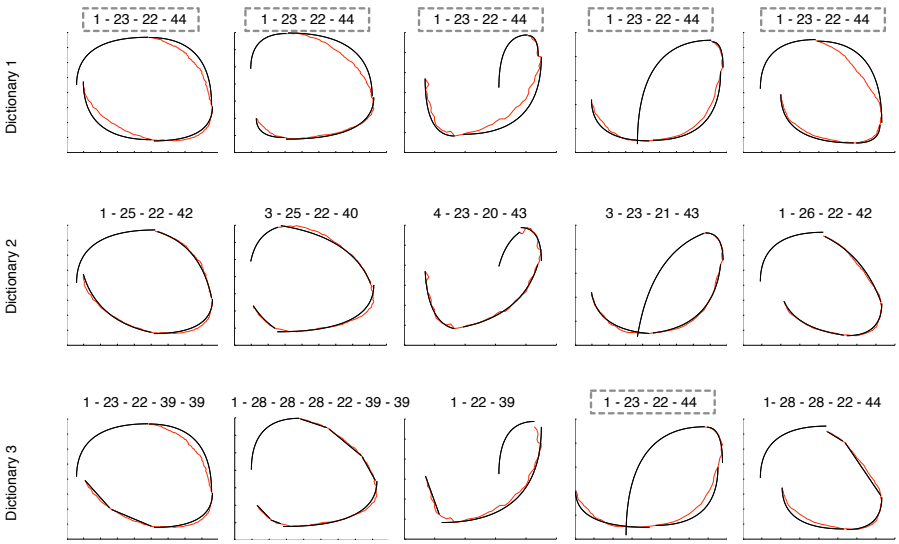

Figure 12: Player 3's fourth interpretation. We can see 1–23–22–44 pattern occurrences in the interpretation inferred using three dictionaries: dictionary 1 at the top, dictionary 2 in the middle and dictionary 3 at the bottom. Shapes approach original signal while the dictionary gets more exhaustive but symbolic representations become highly sensitive to the variations across shapes.

At first sight, the recurring pattern 1–23–22–44 retrieved using dictionary 1 has various shapes: closed circle, open circle, continuous and noncontinuous circles. Even so, all these shapes are labeled 1–23–22–44, that is a circle performed in clockwise direction. Using a more exhaustive dictionary allows for a better fitting of the resynthesized signal from the original signal (see section 6.2 and lines 2 and 3 in figure 12):

SHMM inferred more accurate shapes when using dictionary 2 and 3. For instance, let us consider the first pattern at each line in figure 12. The sequence of symbols given by dictionary 1 (first line) is 1–23–22–44. If we use dictionary 2 (second line), inference induces a novel sequence of indexes that corresponds to the following shape index transformations:

$$\begin{array}{rcl} 1 & \rightarrow & 1; \\ 23 & \rightarrow & 25; \\ 22 & \rightarrow & 22; \\ 44 & \rightarrow & 42; \end{array}$$

Similarly, inferred sequence using dictionary 3 (third line) shows other novel sequence of index that corresponds to the following transformations:

$$\begin{array}{rcl} 1 & \rightarrow & 1; \\ 23 & \rightarrow & 23; \\ 22 & \rightarrow & 22; \\ 44 & \rightarrow & 39-39; \end{array}$$

Hence, proposing a large set of shapes (e.g. dictionary 2) for gesture shape modeling implies a larger set of symbolic representations. In figure 12, all plotted circle patterns have a distinct symbolic representation (distinct sequence of index). It means that SHMM inferred sequences of shapes that tend to adapt to variations in the observations and, consequently, the model is more discriminative. On the contrary, the use of dictionary 1 allows similar continuous patterns (e.g., circles) to be retrieved even if they are not exactly the same at each occurrence (the so-called *motifs* as defined in [34]). The inferred symbolic sequence constitutes of a general *stereotype* in the shape of a clockwise circle. Finally, sequences of indexes inferred by SHMM with dictionary 3 vary in length: some indexes are either repeated, omitted or changed. Even if this dictionary provided with good trade-off between a sample-by-sample fitting criterion (Euclidean norm criterion) and a temporal predictive criterion (log-likelihood criterion), it seems to be less adapted for symbolic representation of considered patterns.

7 Conclusion and perspectives

In this paper, we have presented a segmentation algorithm applied to a real-world data set of clarinetists' ancillary gestures. The hypothesis is that ancillary gestures can be suitably analyzed by multi-level modeling techniques. The general system is based on the definition of a dictionary of primitive shapes and a stochastic temporal modeling of the sequence of shapes that best represents the input gesture signal. The model is tested on a database containing four interpretations of the first movement of the Brahms *First Clarinet Sonata Opus 120, number 1* by four clarinetists.

Firstly, we have shown that the proposed model infers sequences of shapes that accurately fit the input signal. A detailed study of the Euclidean distance between both the re-synthesized and the input signals has shown that the fitting does not depend on the number of elements in a dictionary. We have also shown that the log-likelihood of the temporal sequence increases (i.e better predictive power) if we add relevant elements in the dictionary (e.g dictionary 1 to dictionaries 2 and 3) but might decreases while the number of elements increases if the added elements are not pertinent (e.g, dictionary 4). A trade-off has to be made between the number of elements and how representative of the input observations these elements are. We proposed an incremental process to build a dictionary based on these criteria.

Secondly, we have shown that a greater concentration of short duration shapes occurs between recurrent regular sequences of longer shapes that we called patterns. Short duration shapes model articulations between phrases in the piece while patterns occur within phrases. Such high level structure seems not to be trivially retrievable if we consider only the initial signal.

Finally, we have shown that some patterns occur in every clarinetists' interpretations. The choice of the dictionary makes a difference in retrieving the patterns. A large set of shapes leads to variations in the sequence of symbols that represents the pattern. On the contrary, a smaller set of shapes allows for pattern stereotype definition. Hence, symbolic representation using semantically relevant primitive shapes highlights higher time structures in gesture signals that can be otherwise hidden.

A future improvement of our system will take into account a hierarchical structure in the dictionary. A first dictionary with stereotypical shapes will be defined, accompanied by refinements of each one these elements. We will also pursue a real-time implementation of the model.

8 Acknowledgments

We would like to thank Norbert Schnell for his fruitful advice as well as Delphine Bernardin and Erwin Schoonderwalt for having creating the database. Finally, we would like to thank the COST IC0601 Action on Sonic Interaction Design (SID) for their support in the short-term scientific mission at IDMIL, Montreal.

References

[1] ARIKAN, O., FORSYTH, D., AND O'BRIEN, J. Motion synthesis from annotations. In *ACM SIGGRAPH 2003 Papers* (2003), ACM, pp. 402–408.

[2] ARTIERES, T., MARUKATAT, S., AND GALLINARI, P. Online handwritten shape recognition using segmental hidden markov models. *IEEE Trans. Pattern Anal. Mach. Intell. 29*, 2 (2007), 205–217.

[3] BARBIČ, J., ALLA, S., JIA-YU, P., FALOUTSOS, C., HODGINS, J., AND POLLARD, N. Segmenting motion capture data into distinct behaviors. In *Proceedings of the 2004 Conference on Graphics Interface. London, Ontario, Canada* (Human Motion, Segmentation, Motion Capture, PCA 2004), pp. 185–194.

[4] BLOIT, J., RASAMIMANANA, N., AND BEVILACQUA, F. Modeling and segmentation of audio descriptor profiles with segmental models. *Pattern Recognition Letters* (2010).

[5] BRAND, M., AND HERTZMANN, A. Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques* (2000), ACM Press/Addison-Wesley Publishing Co., pp. 183–192.

[6] BUCHSBAUM, D., GRIFFITHS, T. L., GOPNIK, A., AND BALDWIN, D. Learning from actions and their consequences: Inferring causal variables from continuous sequences of human action. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (2009).

[7] CADOZ, C., AND WANDERELEY, M. M. Gesture-music. *Trends in Gestural Control of Music* (2000).

[8] CARAMIAUX, B., BEVILACQUA, F., AND SCHNELL, N. Towards a gesture-sound cross-modal analysis. In *In Embodied Communication and Human-Computer Interaction, volume 5934 of Lecture Notes in Computer Science*. Springer Verlag, 2010, pp. 158–170.

[9] DAHL, S. The playing of an accent–preliminary observations from temporal and kinematic analysis of percussionists*. *Journal of New Music Research 29*, 3 (2000), 225–233.

[10] DAVIDSON, J. Visual perception of performance manner in the movements of solo musicians. *Psychology of Music 21*, 2 (1993), 103.

[11] ENGEL, K., FLANDERS, M., AND SOECHTING, J. Anticipatory and sequential motor control in piano playing. *Experimental brain research 113*, 2 (1997), 189–199.

[12] FOD, A., MATARIĆ, M., AND JENKINS, O. Automated derivation of primitives for movement classification. *Autonomous robots 12*, 1 (2002), 39–54.

[13] GLARDON, P., BOULIC, R., AND THALMANN, D. Pca-based walking engine using motion capture data. In *In: Proceedings. Computer Graphics International* (2004), IEEE Computer Society.

[14] GODOY, R., JENSENIUS, A., AND NYMOEN, K. Chunking in music by coarticulation. *Acta Acustica united with Acustica 96*, 4 (2010), 690–700.

[15] GUERRA-FILHO, G., AND ALOIMONOS, Y. A language for human action. *Computer 40*, 5 (May 2007), 42–51.

[16] JENSENIUS, A. R., WANDERLEY, M., GODØY, R. I., AND LEMAN, M. Musical gestures: concepts and methods in research. In *Musical gestures: Sound, Movement, and Meaning*. Rolf Inge Godoy and Marc Leman eds., 2009.

[17] KIM, S., AND SMYTH, P. Segmental hidden markov models with random effects for waveform modeling. *The Journal of Machine Learning Research 7* (2006), 969.

[18] KURBY, C., AND ZACKS, J. Segmentation in the perception and memory of events. *Trends in cognitive sciences 12*, 2 (2008), 72–79.

[19] LOEHR, J., AND PALMER, C. Cognitive and biomechanical influences in pianists finger tapping. *Experimental brain research 178*, 4 (2007), 518–528.

[20] MAESTRE, E. *Modeling instrumental gestures: an analysis/synthesis framework for violin bowing*. PhD thesis, Ph. D. dissertation, Univ. Pompeu Fabra, Barcelona, Spain, 2009, 2009.

[21] MOESLUND, T., HILTON, A., AND KRÜGER, V. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding 104*, 2-3 (2006), 90–126.

[22] MURPHY, K. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UC Berkeley, 2002.

[23] OSTENDORF, M., DIGALAKIS, V., AND KIMBALL, O. A. From hmms to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Trans. on Speech and Audio Processing 4* (1996), 360–378.

[24] PALMER, C., KOOPMANS, E., CARTER, C., LOEHR, J., AND WANDERLEY, M. Synchronization of motion and timing in clarinet performance. In *International Symposium on Performance Science* (2009).

[25] RASAMIMANANA, N., AND BEVILACQUA, F. Effort-based analysis of bowing movements: evidence of anticipation effects. *Journal of New Music Research 37*, 4 (2008), 339–351.

[26] TEIXEIRA, E., LOUREIRO, M., AND YEHIA, H. Methodological aspects of the research in musical expressiveness based on corporal movement information.

[27] TURAGA, P., CHELLAPPA, R., SUBRAHMANIAN, V., AND UDREA, O. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on 18*, 11 (2008), 1473–1488.

[28] VINES, B., WANDERLEY, M., KRUMHANSL, C., NUZZO, R., AND LEVITIN, D. Performance gestures of musicians: What structural and emotional information do they convey? *Gesture-based communication in human-computer interaction* (2004), 3887–3887.

[29] WANDERLEY, M. M., AND DEPALLE, P. Gestural control of sound synthesis. *Proceedings of the IEEE 92*, 4 (2005), 632–644.

[30] WANDERLEY, M. M. Quantitative analysis of non-obvious performer gestures. *Gesture and sign language in human-computer interaction. Springer-Verlag* (2002), 241–253.

[31] WANDERLEY, M. M., VINES, B.W., MIDDLETON, N., MCKAY, C., AND HATCH, W. The musical significance of clarinetists' ancillary gestures: An exploration of the field. *Journal of New Music Research 34*, 1 (2005), 97–113.

[32] WIDMER, G., DIXON, S., GOEBL, W., PAMPALK, E., AND TOBUDIC, A. In search of the horowitz factor. *AI Magazine 24*, 3 (2003), 111.

[33] WIDMER, G., AND GOEBL, W. Computational models of expressive music performance: The state of the art. *Journal of New Music Research 33*, 3 (2004), 203–216.

[34] YANKOV, D., KEOGH, E., MEDINA, J., CHIU, B., AND ZORDAN, V. Detecting time series motifs under uniform scaling. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (2007), ACM, pp. 844–853.