Dynamic Bayesian networks for musical interaction

Baptiste Caramiaux, Jules Françoise, Frédéric Bevilacqua

Dynamic Bayesian Networks for Musical Interaction

Baptiste Caramiaux[1,2], Jules Françoise[1] and Frédéric Bevilacqua[1]

[1]UMR STMS Ircam-CNRS-UPMC, Paris, France
[2]McGill University, Montreal, Canada

Introduction

We describe in this chapter a consistent set of temporal models that we have developed over the years for analyzing movement in real-time musical interaction. These models are probabilistic and can be unified and generalized under the formalism of dynamic Bayesian networks (DBNs). We believe that this unified approach offers new perspectives for embodied music cognition and interaction, both in terms of fundamental studies and technological development, as this will be discussed after the general presentation of the framework and the review of specific models previously implemented.

By musical interaction, we refer to a broad body of work from gesture-based control of sound synthesis (Wanderley and Depalle, 2004) to more general interaction paradigms involving human movements and recorded media (Schnell, 2013), as investigated by the NIME community (New Interfaces for Musical Expression; (Bevilacqua et al., 2013)). The different interaction models behind such systems vary greatly and range from fully deterministic approaches to probabilistic models. In particular, machine learning techniques have been used by researchers and creative practitioners to teach the machines particular behaviors instead of programming them (Fiebrink and Caramiaux, 2016). In this context, several systems use non-temporal methods such as principal component analysis, support vector machines, or Gaussian mixture models in order to accomplish a given task such as gesture classification, motion-to-sound regression, or gesture clustering. However, these methods and models are not suitable to fully capture the characteristics and variations in movement execution because they only consider snapshots of the movement being executed. This motivated researchers to look at other approaches, such as hidden Markov models and dynamical systems, in order to model dynamic behaviors. Our previous work on motion modeling has contributed to this endeavor by proposing a family of methods capable of performing real-time analysis of human motion by taking into account the history of the executed gesture.

In this chapter, our aim is to gather these methods under a general model of movement execution and its possible variations that can originate both from noise in motor control and from conscious variations driven by a particular performance interpretation. We propose to use the general framework of DBNs that allows for modeling at various temporal scales and handles the intrinsic variability of the measured movement features. The framework is introduced in the section “Modeling strategy”. The models afforded by DBNs are generic enough to be configured in order to capture short-term and long-term temporal dependences. We show three implementations that illustrate the analysis at various temporal levels from our previous work in the section “Temporal levels”. In addition,

the probabilistic nature of DBNs allows for modeling of spatiotemporal variations inherent to human motion. We present two examples from our previous work in the section "Spatiotemporal variations". In the section "Discussion", we discuss applications of these models as tools for experimental music research.

Modeling Strategy

Our goal is to determine in real time the characteristics of the physical movement that can be used for two purposes: performance analysis and/or control parameters in interactive music systems. Various challenges arise from analyzing such performance data: (a) the motion characteristics might not be independent of each other; (b) the correspondence between these characteristics and the signal measured with motion capture might be non-trivial or corrupted by noise; and (c) these characteristics may often be time-varying. In this section, we remind the reader that data variability can be handled within a probabilistic framework, and we introduce dynamic Bayesian networks, which handle dependences between variables of interest and time.

Modeling Strategy

Variability is inherent to any observed data. If we observe several times the same movement performed by a musician, the movements, although similar, will differ to some degree. Differences across performances are due to several factors: (1) various sources from the different technological layers used to capture, transmit, and process the musician's movements generate noise; (2) inherent variability arises from the performer's motor system (Zatorre et al., 2007); and (3) performing music also relies on subjective interpretation, understood as deliberate variations of timing and dynamics added by the musician to the performance (Palmer, 1997). Therefore, spatiotemporal variation of the mea- sured parameters can be related to interpretation and style.

Variability can be modeled through a probabilistic framework using random variables. A random variable can take a finite set of values – each of them having a probability assigned to it. The set of values together with their probabilities is usually referred to as the probability distribution. Within a probabilistic framework, the dependence between two variables is defined by the conditional probability distribution over the values of one random variable given the values of another random variable. We can also model time dependences by specifying conditional probability distributions between random variables and time (i.e., between random variables and their previous values).

Probability distributions can rarely be computed directly, and conditional probability distributions are often challenging to compute. Bayes' rule provides a way to compute these probabilities through operations between alternative distributions that are often easier to compute. Bayes' rule infers a posteriori belief (probability) on a random variable from our prior belief on this variable and the likelihood of this belief based on what we observe. We will see in the following section how the Bayesian hypothesis can be naturally linked with the time dependence of the system.

Temporal Modeling Using Dynamic Bayesian Network

Commonly used temporal models for human motion, such as hidden Markov models, Kalman filters, or autoregressive filters, have been shown to belong to the same family of models called dynamic Bayesian networks (Murphy, 2002).

Figure 1 (top) illustrates a generic representation of DBNs applied to human motion modeling. A motion trajectory captured by the system generates a sampled trajectory. A DBN models the motion through a set of dependent variables at each time step. Some variables are observable (typically the sampled motion trajectory from a specific capture system) and others are hidden (typically the internal states that characterize the latent structure of the observed data). Hidden variables include the motion characteristics such as execution speed, accelerations, curvature, and tension. The dependences between these hidden variables capture various temporal scales, as we will see in the next section. Dependences between hidden variables and between hidden and observable variables can be represented graphically as a network, as shown in Figure 1. Note that the network is considered dynamic because it is used to model a dynamical system (Murphy, 2002).

The set of dependent hidden variables is updated at each time step t, given a new observation (observable variable) and the set of hidden variables previously computed at time $t-1$. This process of estimating the hidden variables of the model is called Bayesian inference because it makes use of Bayes' rule: Our belief on the internal state of a system at time t is estimated by combining our prior belief on the system with the new evidence at time t (the observation). Precisely, inference estimates the probability distribution over the hidden variables (i.e., computing probabilities of the possible values of the hidden variables) in a slice at a given time step based on the observed data and the previously estimated distribution. Reporting a comprehensive list of inference methods is beyond the scope of this chapter, but the interested reader can refer to the work of (Murphy, 2002).

For computer-mediated real-time musical interaction, we aim at estimating motion characteristics from a live motion-capture data stream in order to interact with a sound synthesis engine. In this case, we use inference as a way to estimate in real time the internal state (hidden variables) representing the motion characteristics and then use the estimation of the characteristics for analysis or controlling sound synthesis parameters.

Finally, note that the topology of the internal state may influence the inference method. If we want to infer the speed of an observed human motion, we may be able to write an analytical formulation of the relationships between the captured motion positions and the speed. In this case, we can perform exact inference and estimate the probability distributions according to the analytical solution to the problem using, for instance, Kalman filters. While exact inference can be implemented and solved in many cases, complex networks with non-linear and/or non-Gaussian dependences might not have an analytical solution, or the solution might not be tractable. In this case, approximate methods, such as sampling (Arulampalam et al., 2002), can be used to compute an approximation of the probability distribution.

Designing DBNs

As stated earlier, in our previous work we have developed a set of temporal models of human motion that can all be considered as specific implementations of DBNs. In the next sections we will present these models as particular implementations of DBNs. These models differ in the choice of the hidden variables and the way their dependence is chosen. The next section illustrates how DBNs can account for time dependence at various levels through hierarchical dependence of their hidden variables. Then, the following section illustrates how DBNs can take motion variations into account through the choice of the hidden variables.

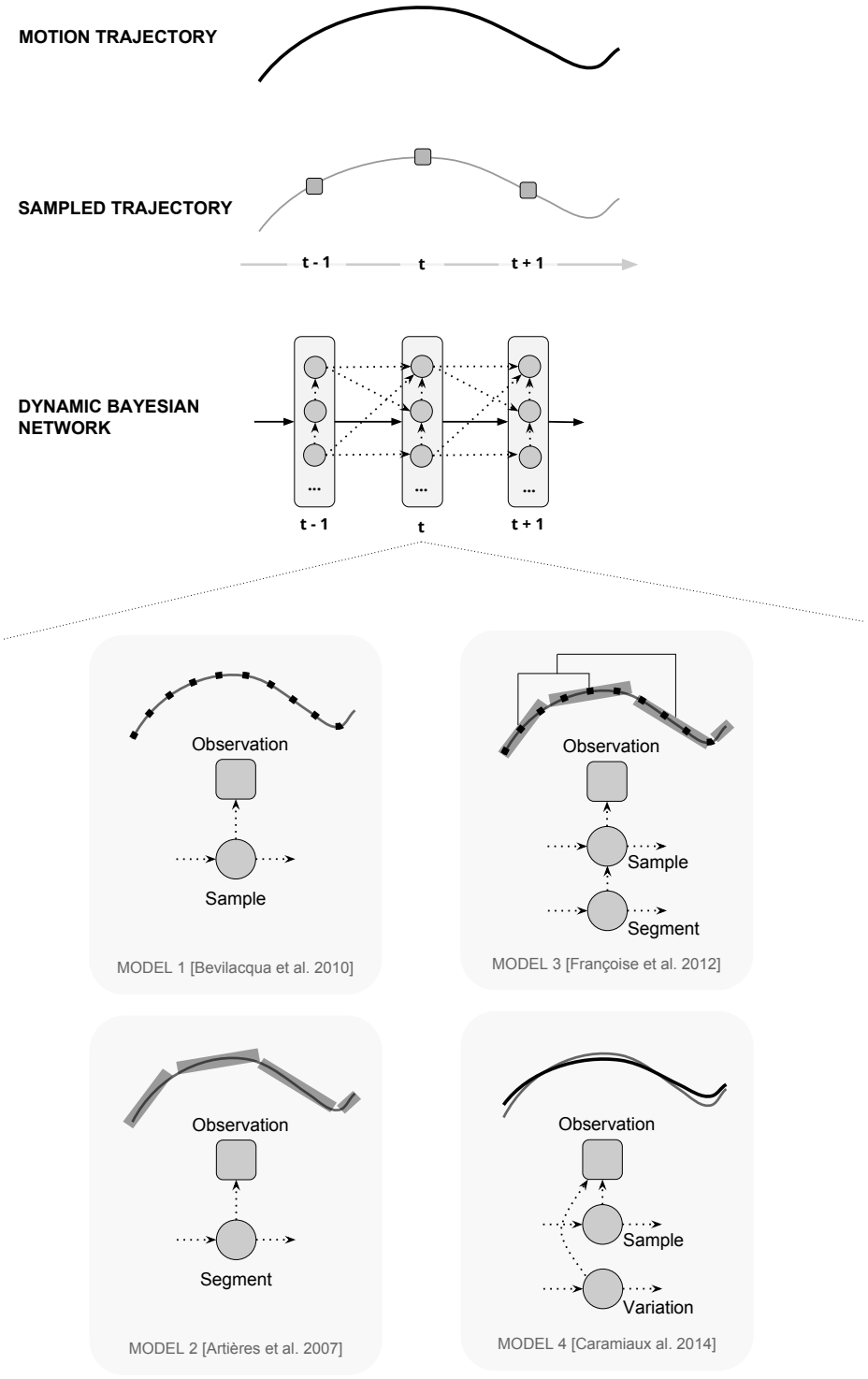

Figure 1: Generic representation of Dynamic Bayesian Networks (DBNs) applied to human motion modeling. Top: Motion trajectory is sampled by a capture system; a DBN models these data through a set of dependent variables at each time step. Bottom: Various implementations of DBNs from previous work, handling multi-level temporal structures (sample-level [model 1], segment-level [model 2], hierarchical [model 3]) and explicit movement variations (model 4).

Temporal Levels

Temporal models of human motion can capture time dependence at various levels, from short- to long-term temporal dependence, and nested time dependences. In this section, we present how our applications in music interaction motivated the design and implementation of several models in order to examine changes in human motion at different temporal scales.

Sample Level

Temporal models, as formalized in DBNs, can handle changes at the level of the sample (considered as the atomic representation of the movement provided by the capture system) by updating the state values at each new observation (see Figure 1). Practically, such level of modeling can afford continuous interactions by enabling the computer to act at each new received motion sample – while the movement is being performed.

In our prior work, we proposed such implementation to perform real-time alignment of a motion trajectory (or gesture) onto pre-recorded motion templates through a model called the gesture follower (Bevilacqua et al., 2009). The model is depicted in Figure 1 (bottom, Model 1). The model has a structure of a hidden Markov model (HMM), which is the simplest case of a DBN: It has one hidden state that generates the observation. In our implementation, the hidden state is a random variable representing the sample index in pre-recorded gesture templates. This simple structure allows for tractable and exact inference that is used to decode the most probable sequence of samples (from a pre-recorded template) given a new sequence of motion samples. The inference is incremental: At each new observation (incoming motion sample), the model infers the likeliest sample index in a template based on the previously estimated index and the observation.

This model has been employed in several use cases in music and dance for both creation and analysis. In particular, it has been shown to be able to recognize and track professional musicians' gestures while performing a complex piece from contemporary repertoire (Bevilacqua et al., 2012). In this context, the model has been used to follow the gestures of instrumentalists in a string quartet in order to synchronize electronic music pieces to the acoustic performance. In addition, the model captures temporal ambiguities in a musician's performance by dynamically adapting the variance of the probability distributions over the index: Large variance values indicate that the estimation of the index value is ambiguous, whereas small values indicate that the index is statistically well defined.

Segment Level

While the previous section presented a "temporal zoom" into musicians' movements, motivated by the need to follow motion trajectories (or morphologies) for continuous interaction, here we inspect temporal modeling of human movements in music performance – either due to the structure of a written score or due to perception and planning (Godøy et al., 2010; Janata and Grafton, 2003). One way to do so is to consider musicians' movements as a sequence of motion primitives, understood as basic units such as patterns of movement kinematics, where each primitive can be represented as a sequence of samples.

In our prior work we used such a model to segment instrumentalists' gestures in sequences of primitive segments taken in a set of pre-defined primitives (called a dictionary) and used the inferred sequences for gesture analysis (Caramiaux et al., 2012).

The model is also based on an HMM, called segmental HMM, and has been previously proposed for handwritten shape recognition (Artieres et al., 2007). In a segmental HMM, the hidden state now represents the current segment in which the observation belongs

(Figure 1 bottom, Model 2). In addition, each segment has an explicit duration in the model (not represented in the figure for the sake of clarity). Probability distributions on the segment lengths allow for imposing short lengths to some specific segments and a wider range of lengths for others.

We used this model to analyze clarinetists' ancillary movements (here the movements of the clarinet bell) executed while performing a piece from the classical repertoire (Caramiaux et al., 2012). First, we defined a set of primitive trajectories on which to decompose the clarinetists' ancillary motion trajectories. Then, we parsed the sequence of primitives in order to inspect the regularities and differences among performers. As a result, the model has shown regularities such as circles made with the clarinet, which temporal boundaries can be linked to the structure of the composition. In addition, we found subjective deformations of these shapes that are characteristic of each player's subjective interpretation of the piece.

Interestingly, varying the probability distribution defined over the segment lengths changes the resolution of the analysis. If small length values have high probabilities, then the resulting sequence will be made of a high number of small segments, which has been shown to allow the analysis of subtle idiosyncratic structures in the clarinetists' performance. On the contrary, if the probability density is centered on higher length values, the resulting sequence will be made of fewer and longer segments, which tend to highlight commonalities across performances of the clarinetists in our study.

Hierarchical Structure

The previous sections introduced models operating at different timescales: a high temporal resolution inspecting continuous changes in movement timing and a low temporal resolution that parses gesture segments on a longer timescale. These complementary views can be unified through hierarchical representations.

In previous work (Françoise et al., 2012), we proposed a hierarchical representation of music-related gestures using the hierarchical HMM (HHMM; (Fine et al., 1998)), which extends standard HMM with a hierarchy of hidden states (see Figure 1, Model 3). In the hierarchical HMM, both short- and long-term temporal dependences can be modeled through a hierarchy of hidden states in each slice. The highest level of hidden states corresponds to the largest timescale in the movement representation. In a model with two levels, each state in the highest-level layer is associated to a motion segment. Each of these "segment" states generates a sub-model at a lower timescale that encodes the continuous trajectory of motion parameters associated with the segment – for example, at the sample level, as proposed in (Françoise et al., 2012), or with a fixed number of states (Françoise et al., 2015).

In the proposed implementation of segmental HMM, each segment is modeled as a geometric shape that can only be stretched uniformly. The power of the hierarchical HMM resides in the unification of high and low temporal resolution within the same structure. Segments are modeled at the sample level using an implicit time model, but they are also embedded in a long-term transition structure. The transition probabilities between segments can be manually authored or learned to integrate long-term dependences and information on how segments can be sequenced. This allows for continuously tracking and aligning a new gesture segment over a reference and, simultaneously, for anticipating a transition to another segment when the current segment approaches its end.

We proposed a particular representation of musical gestures where each sound-related gesture is represented as four segments: preparation (a "pre-gesture" that anticipates the beginning of the sound), attack and sustain (analogous to the sounds' attack and sustain),

and release (their accompaniment or recovery gesture that might continue after the end of the sound). This representation can be encoded in a hierarchical HMM by allocating one high-level state to each segment, and by specifying how one can make the transition from one segment to another. For example, we can authorize the entry into the gesture through preparation and attack only, and allow exiting from gesture via sustain or release segments only.

This structure was shown to be efficient for interactive control of recorded sounds: Gestures can be divided into a sequence of constitutive segments for learning, and these segments can then be played in an arbitrary order (as allowed by the model). Moreover, different mapping strategies can be specified for each segment. For example, consistent synthesis of the transients on attack phases can be guaranteed while allowing for continuous timing variations on the sustain and release phases.

Note that in our previous work, we implemented this model with a two-level topology (segment and sample), but it can be extended to an arbitrary number of layers to model time processes on longer timescales such as motifs, phrases, or parts of a musical score.

Spatiotemporal Variations

Movement performance is fundamentally variable among performers, executions, expertise, tasks, and context. By design, DBNs have the ability to handle such variability. We have proposed strategies to handle spatiotemporal variations in a Bayesian framework either as a means to movement analysis or as an expressive vector in real-time musical interaction.

Learning Movement Variability

In this section, we report an example of the use of dynamic Bayesian networks to investigate the temporal evolution of the movement variability from one performance to another with regards to the performer's expertise (Françoise et al., 2015). In this case, the goal was to develop a method for off-line analysis that uses information encoded in models trained from several performances of the same sequence.

We analyzed several performances of a known sequence of t'ai chi movements by an expert and a novice performer. All performances were used to train a hierarchical HMM representing the full movement sequence of each performer. We postulated that the configuration of the trained network, along with the values of its internal states, incorporate critical information on the variability of the performer over several executions of the same sequence – in particular, through the variances of each state of the DBN.

To examine such variability in an intelligible way, we synthesized the average movement trajectory, along with the associated variances over time. We found that the synthesis highlighted important variations of the variance over time. The variance significantly and consistently decreased on a set of key gestures in the sequence. This decrease of the variability across performances was even stronger as the expertise of the performer increased. Moreover, changing the properties of the DBN can lead to additional insights. For instance, increasing the number of states used to model each gesture clearly highlighted that the expert's movements were performed more accurately at a higher temporal resolution.

One of the critical advantages of DBN for this analysis task is their flexibility for temporal modeling. DBNs are typically trained from several examples of the same gesture or sequence, without requiring prior alignment of the various recordings. As a result, DBNs can be used for analyzing both continuous changes in timing (using HMMs for

sequence alignment) and the variability in dynamics across several performances (using the variances estimated by the training algorithm).

Adaptation to Movement Variations

In the previous approach, the goal was to develop an off-line tool to analyze motion variability, based on the learned variances of the motion characteristics. Here, we look at on-line analysis, which requires adapting to movement characteristics that vary while the movement is performed. A practical application is the design of movement-based musical interaction that supports expressive motion variations as they occur in traditional music performance.

The models we have presented so far are based on an HMM-like structure (Figure 1, bottom, Models 1, 2, and 3) that is characterized by hidden states taking values in a discrete set. For instance, the HMM-based model presented in the section "Sample level" (Figure 1, Model 1) approximates a continuous motion template by its discrete set of samples whose index are the hidden states of the HMM. The model we are presenting in this section proposes to consider motion variations as continuous variables, and it is depicted in Figure 1, Model 4.

A DBN for which the hidden states are continuous is called a dynamical system (a well-known linear dynamical system is the Kalman filter). The proposed model is a non-linear dynamical system that is able to handle a finite set of simple motion variations defined as amplitude, speed, and orientation, called the gesture variation follower (see (Caramiaux et al., 2015), for a detailed description of the model and examples). The relationship between the observations, typically the motion sample values (such as x and y values on a plane), and the variations to be inferred from the observations (typically the rotation angle, scale coefficients, and speed) can be complex (e.g., non-linear). In our model we used a sampling method called particle filtering (see (Caramiaux et al., 2015), for more details), which consider at any time a large number of potential variations and their likelihood.

Our typical use case in musical interaction is the continuous control of sound parameters (or audio effect) in real time. The scenario is as follows: As a performer starts a gesture, the method infers continuous variations of the current gesture according to pre-recorded templates, such as rotation angles, size, and speed. These continuous variations are then mapped onto sound parameters such as the cutoff frequency of a digital filter, audio volume, and playback speed. We have shown that such a system is usable by performers (Caramiaux et al., 2015)). In addition, the model has also been evaluated in terms of user experience, and we showed that it affords an expressive, attractive, and hedonic experience to the user when compared to more traditional interaction techniques such as menus and sliders (Caramiaux et al., 2013).

Discussion

In this chapter we have presented a series of DBN models of human motion. While general-purpose temporal models of human motion have been widely used to perform recognition, tracking, or generation, our models have been designed to allow for real-time inference of motion characteristics that can then be used in musical interaction (e.g., to be mapped to sound synthesis parameters).

Interestingly, each of the four models, as depicted in Figure 1 (bottom), conceives the notion of temporal structure in human motion differently. Model 1 captures the

temporal structure as a temporal sequence of samples following a spatial trajectory. Model 2 assimilates the temporal structure as a temporal sequence of segments. Model 3 brings together both Model 1 and Model 2 by capturing the hierarchical structure of a sequence of segments themselves considered as sampled trajectories. Finally, Model 4 considers the temporal structure of the motion variations as a continuous dynamical system rather than as a sequence of elements.

As an element of an interactive system bridging the input motion and digital media, each model affords time-dependent actions that have rich potential for novel interaction designs. Other models can then be imagined with higher levels of hierarchy and more complex relationships between states across levels. However, besides the potential for interaction, increasing the complexity of the model implies an increasing difficulty in training the model and requires a larger amount of data. In our model we often used heuristics that fit our end goal.

With regards to spatiotemporal variations, we saw that characteristics of variations could be modeled through the variance of the hidden states. High variance in the hidden states means variability in the observed data. While this variability can be due to a lack of expertise in performance, it can also be attributed by voluntary motion variations for real-time music performance.

Developing models of motion execution and variations based on dynamic Bayesian networks has the potential to address important challenges in music performance and perception. In the following paragraphs, we consider three main challenges: modeling co-articulation, modeling coordination, and linking perception with motor control.

Considering DBN as a tool to model dependences between dynamic variables, a first challenge would be to formalize the problem of co-articulation as a DBN, where co-articulation is defined as the fusion of small-scale events into phrase-level segments (Godøy et al., 2010). This first challenge would require the careful definition of the relevant variables and their dependences, as well as the need for a dataset embedding co-articulatory elements (Bevilacqua et al., 2016).

The first challenge can then be conceptually extended to the problem of coordination among various physical limbs producing a movement or various musicians performing in an ensemble. While the problem of co-articulation inspects interdependence within a sequence of elements, the problem of coordination involves the definition of several processes that are mutually dependent. Such models already exist in the literature, such as a simple version, the coupled HMM (Brand et al., 1997).

A second, and more general, challenge is the link between the proposed framework and the literature in music psychology. Such literature has long shown that time plays a fundamental role in music production and perception. On the one hand, fine-grained time-dependent processes have been investigated in behavioral experimental research that shows, for instance, the capacity of musicians to be temporally more accurate than non-musicians using a sensorimotor synchronization paradigm (Aschersleben, 2002), or to be more sensitive to action–reaction delays (van Vugt and Tillmann, 2014). On the other hand, segment-level time-dependent processes govern sequence planning and control. Within these sequences, elements are co-articulated by means of cognitive and biomechanical processes (see, for instance, (Loehr and Palmer, 2007)).

The processes underlying sensorimotor synchronization can be examined by various HMM topologies (such as the coupled or multi-modal HMM). The notion of expertise could also be tackled through a Bayesian framework. In a recent article, Braun and colleagues have formulated the problem of "learning to learn" actions through structure learning in a Bayesian network (Braun et al., 2010). Methods exist within the DBN

framework to learn internal structures and their time dependence. Future research in music psychology inspecting the role of expertise in music performance can leverage on such Bayesian formulations.

We believe that, in general, Bayesian models can provide a computational account for the cognitive processes at play during such tasks. Previous work in cognitive neuroscience already shed light on the Bayesian nature of integration in sensorimotor learning (Körding and Wolpert, 2006). As stated by the authors, "The central nervous system [...] employs probabilistic models during sensorimotor learning". In our view, this calls for a thorough exploration of the potential of dynamic Bayesian networks in order to examine, through a different perspective, various research topics in music psychology.

Acknowledgements

This work is supported by the Marie Skłodowska-Curie Action of the European Union (H2020- MSCA-IF-2014, IF-GF, grant agreement no. 659232) and by the Rapid-Mix EU project (H2020- ICT-2014-1, project ID 644862).

References

Artieres, T., Marukatat, S., and Gallinari, P. (2007). Online handwritten shape recognition using segmental hidden markov models. *IEEE transactions on pattern analysis and machine intelligence*, 29(2).

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on signal processing*, 50(2):174–188.

Aschersleben, G. (2002). Temporal control of movements in sensorimotor synchronization. *Brain and cognition*, 48(1):66–79.

Bevilacqua, F., Baschet, F., and Lemouton, S. (2012). The augmented string quartet: experiments and gesture following. *Journal of New Music Research*, 41(1):103–119.

Bevilacqua, F., Caramiaux, B., and Françoise, J. (2016). Perspectives on real-time computation of movement coarticulation. In *Proceedings of the 3rd International Symposium on Movement and Computing*, page 35. ACM.

Bevilacqua, F., Fels, S., Jensenius, A. R., Lyons, M. J., Schnell, N., and Tanaka, A. (2013). Sig nime: music, technology, and human-computer interaction. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 2529–2532. ACM.

Bevilacqua, F., Zamborlin, B., Sypniewski, A., Schnell, N., Guédy, F., and Rasamimanana, N. (2009). Continuous realtime gesture following and recognition. In *International gesture workshop*, pages 73–84. Springer.

Brand, M., Oliver, N., and Pentland, A. (1997). Coupled hidden markov models for complex action recognition. In *Computer vision and pattern recognition, 1997. proceedings., 1997 ieee computer society conference on*, pages 994–999. IEEE.

Braun, D. A., Mehring, C., and Wolpert, D. M. (2010). Structure learning in action. *Behavioural brain research*, 206(2):157–165.

Caramiaux, B., Bevilacqua, F., and Tanaka, A. (2013). Beyond recognition: using gesture variation for continuous interaction. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 2109–2118. ACM.

Caramiaux, B., Montecchio, N., Tanaka, A., and Bevilacqua, F. (2015). Adaptive gesture recognition with variation estimation for interactive systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 4(4):18.

Caramiaux, B., Wanderley, M. M., and Bevilacqua, F. (2012). Segmenting and parsing instrumentalists' gestures. *Journal of New Music Research*, 41(1):13–29.

Fiebrink, R. and Caramiaux, B. (2016). The machine learning algorithm as creative musical tool. *arXiv preprint arXiv:1611.00379*.

Fine, S., Singer, Y., and Tishby, N. (1998). The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32(1):41–62.

Françoise, J., Caramiaux, B., and Bevilacqua, F. (2012). A hierarchical approach for the design of gesture-to-sound mappings. In *9th Sound and Music Computing Conference*, pages 233–240.

Françoise, J., Roby-Brami, A., Riboud, N., and Bevilacqua, F. (2015). Movement sequence analysis using hidden markov models: a case study in tai chi performance. In *Proceedings of the 2nd International Workshop on Movement and Computing*, pages 29–36. ACM.

Godøy, R. I., Jensenius, A. R., and Nymoen, K. (2010). Chunking in music by coarticulation. *Acta Acustica united with Acustica*, 96(4):690–700.

Janata, P. and Grafton, S. T. (2003). Swinging in the brain: shared neural substrates for behaviors related to sequencing and music. *Nature neuroscience*, 6(7):682.

Körding, K. P. and Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in cognitive sciences*, 10(7):319–326.

Loehr, J. D. and Palmer, C. (2007). Cognitive and biomechanical influences in pianists finger tapping. *Experimental brain research*, 178(4):518–528.

Murphy, K. P. (2002). *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley.

Palmer, C. (1997). Music performance. *Annual review of psychology*, 48(1):115–138.

Schnell, N. (2013). *Playing (with) Sound-Of the Animation of Digitized Sounds and their Reenactment by Playful Scenarios in the Design of Interactive Audio Applications*. na.

van Vugt, F. T. and Tillmann, B. (2014). Thresholds of auditory-motor coupling measured with a simple task in musicians and non-musicians: was the sound simultaneous to the key press? *PLoS One*, 9(2):e87176.

Wanderley, M. M. and Depalle, P. (2004). Gestural control of sound synthesis. *Proceedings of the IEEE*, 92(4):632–644.

Zatorre, R. J., Chen, J. L., and Penhune, V. B. (2007). When the brain plays music: auditory-motor interactions in music perception and production. *Nature reviews. Neuroscience*, 8(7):547.