A Hierarchical Approach for the Design of Gesture-to-Sound Mappings

Jules Françoise, Baptiste Caramiaux, Frédéric Bevilacqua

A HIERARCHICAL APPROACH FOR THE DESIGN OF GESTURE-TO-SOUND MAPPINGS

Jules Françoise, Baptiste Caramiaux, Frédéric Bevilacqua
STMS Lab, Ircam–CNRS–UPMC
1, Place Igor Strasvinsky
75004 Paris, France
jules.francoise@ircam.fr

ABSTRACT

We propose a hierarchical approach for the design of gesture-to-sound mappings, with the goal to take into account multilevel time structures in both gesture and sound processes. This allows for the integration of temporal mapping strategies, complementing mapping systems based on instantaneous relationships between gesture and sound synthesis parameters. As an example, we propose the implementation of Hierarchical Hidden Markov Models to model gesture input, with a flexible structure that can be authored by the user. Moreover, some parameters can be adjusted through a learning phase. We show some examples of gesture segmentations based on this approach, considering several phases such as preparation, attack, sustain, release. Finally we describe an application, developed in Max/MSP, illustrating the use of accelerometer-based sensors to control phase vocoder synthesis techniques based on this approach.

1. INTRODUCTION

The design of computational models of gesture-to-sound mappings remains a central research question, that is necessary for the development of innovative musical interfaces. In most systems, the modeling of the temporal structures of gestures and their relationships to sound description remains very basic. Generally, only "instantaneous" relationships (i.e. at the fastest framerate of the system) between gesture and sound are considered , and larger time scale structures are not taken into account.

Our main motivation is to address these limitations by modeling gesture-to-sound relationships as the mapping of hierarchically structured temporal processes. By introducing hierarchical segmentation models of gesture and sound, we extend the representation of gesture-sound relationships to a multilevel structure, spanning from a fine control of sound details to a high level control on temporal structures. Moreover, we aim at using machine learning techniques, namely Hierarchical Hidden Markov Models, to train some parameters of the model using examples provided by the user.

The paper is organized as follows. After a review of related work in Section 2, we describe the general hierarchical approach to mapping in Section 3, and then report in Section 4 on a case study using Hierarchical Hidden Markov Models for gesture recognition and segmentation. Finally in Section 5, we present the implementation of a concrete application using this hierarchical framework for gesture-based control of audio processing.

2. MOTIVATIONS AND RELATED WORKS

In interactive music systems, the representation of gesture-to-sound synthesis controls led to the notion of so called *mapping*. This has been extensively studied in the early 2000's (see for instance [1, 2]) leading to taxonomies inspired either by its structure (one-to-one, one-to-many, many-to-one) or its degree of determinism (explicit, implicit). The last taxonomy divides between mappings relying on: 1) an explicit function of input parameters to output parameters [3]; 2) an implicit function, determined by training methods [4], or purposely designed as stochastic [5].

Considering implicit methods, various machine learning techniques have been proposed over the years. Early works include for example the use of neural networks for non-linear multidimensional mappings [6,7] or PCA for dimensionality reduction [8]. More recent works have showed renewed experiments with variety of algorithms and new software tools, for example the *Gesture Follower* [9], the SARC EyesWeb Catalog [10], the Wekinator [11] or the libmapper [12, 13].

While most approaches focus on recognizing gesture units independently of the sound process, some recent works propose to learn more directly the mapping between gesture and sound. For gesture-sound relationship analysis, we recently proposed the use of multimodal dimension reduction [14]. In music performance, Merrill et al. [15] proposed such an approach where participants can personalize mapping functions. Fiebrink et al. [16] proposed a set of algorithms to learn the mapping from a performed gesture "played along" with the music. Similarly, we proposed a procedure to learn the temporal relationship between gesture and sound (we call temporal mapping) where the users gestures are recorded while they listen to specific sounds [17, 18].

This type of approaches are very promising, and our long-term goal is to pursue such directions. However, it ap-

peared to us that a major current limitation resides in the temporal modeling of gestures, and in particular, the lack of hierarchical structures.

Similarly, Jordà [19, 20] argued for the need of considering different control levels allowing for either intuitive or compositional decisions. Recently, Caramiaux [21] highlighted the intricate relationship existing between hierarchical temporal structures in both gesture and sound when the gesture is performed in a listening situation. This is consistent with theoretical works by Godøy et al. [22] translating the notion of "chunking" in perception psychology to music. In music performance, chunking can be understood as a hierarchical cognitive organization of both movement and music (note that the author preferably uses the word *coarticulation*).

As a matter of fact, few works deal explicitly with the introduction of multilevel temporal signal models of gesture and sound in a musical context. Bloit et al. [23] used a segmental approach based on hidden Markov modeling for the description and classification of sound morphologies. The model has then been used for clarinetist's ancillary gesture analysis by Caramiaux et al. [24], highlighting consistent patterns in the gesture performance linked to the score structure. Nevertheless, this approach remains difficult to implement in real-time.

In summary, gesture-to-sound mappings taking into account a hierarchical structure seems to have rarely been proposed, or at least not been fully implemented using machine learning techniques. In this paper we argue that such hierarchical temporal mapping should be relevant for the expressive control of sound based on gestures and we propose an implementation using Hierarchical Hidden Markov Models.

3. HIERARCHICAL APPROACH FOR MAPPING

3.1 Overview

In this section, we introduce a hierarchical approach for gesture-to-sound mapping. This approach is based on hierarchical segmentations of both gesture and sound parameters, in order to structure the relationships between the two modalities into a multilevel time architecture.

A graphical representation of the proposed framework is depicted in Figure 1. A multilevel segmentation is applied to the gesture data (represented as a continuous time profile at the top of the figure). This hierarchical structure is put in relationship with different sound processes, that are also naturally structured hierarchically (e.g. in attack/sustain parts, notes, phrases etc). The sound processes can be seen as a series of overlapping units that are combined to form larger sound structures.

This approach leads to gesture-to-sound mappings that include multilevel time structures, spanning from *micro* signal levels to *macro* level units (that might be identified to symbolic levels). Hence, we propose an extension of standard notions of mapping, that generally takes into account only "instantaneous" links between gesture and sound parameters, to include high level control of audio processes at different time scales.

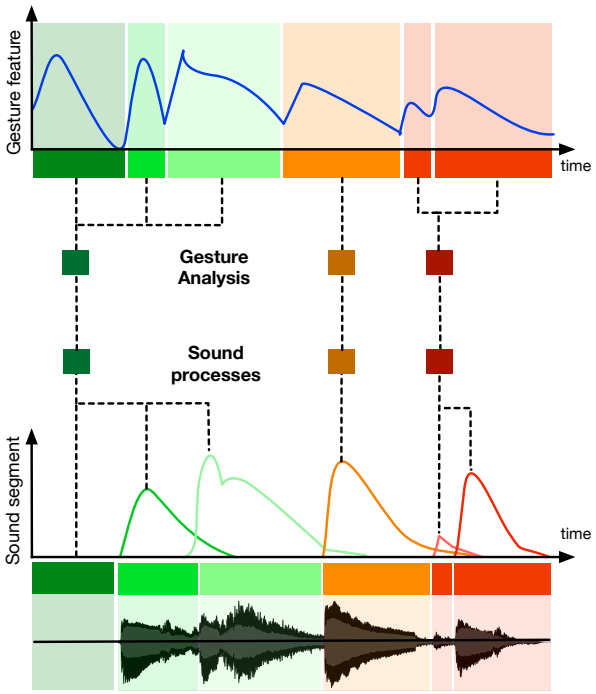

Figure 1. Overview of the proposed approach. Hierarchical structures in both gesture and sound parameters are used to build multilevel mappings from gesture to sound.

We detail in the following sections the three main elements of our framework: gesture modeling, sound processes and multilevel mapping.

3.2 Hierarchical gesture modeling

We consider the gesture signal (e.g. obtained by sensors) as a continuous flow of parameters. The first level of segmentation defines the unitary components we call gesture **segments**, modeled by multidimensional time profiles.

The segments sequencing is governed by a multilevel probabilistic structure encoding their temporal dependencies. A **gesture** can be therefore considered as an ordered **sequence** of segments.

Machine learning techniques can be used to identify the gesture segments and their organization in time from a continuous input. In particular, two main tasks can be computed:

- **segmentation**: identify the temporal boundaries of gesture segments;
- **recognition**: recognize and label gestures at each level of the hierarchy.

Both segmentation and recognition can be computed sequentially or can be coupled: the segmentation being performed conjointly with the recognition task.

3.3 Hierarchical sound processes

To draw a parallel with the gestural representation, we consider a multidimensional continuous stream of parameters

at the input of the sound synthesis system. Similar decomposition can be applied, implying the definition of **sound control segments** and **sound control gestures**.

Basically, segments designate primitive time profiles of sound control parameters. This hierarchical sound synthesis can be performed by scheduling sequences of overlapping segments, augmented with instantaneous control. In particular cases, each of these elements can be easily associated with common sound and musical structures, such as musical phrases, notes, themselves containing subelements such as attack, decay, release, sustain.

Our framework is compatible with any type of sound processing or synthesis. We currently experiment with concatenative synthesis, analysis/synthesis methods (e.g. phase vocoder) or physical models.

In the application of section 5, we focus on sound synthesis techniques based on phase vocoding. In such case, we define sound segments directly, using a manual annotation of sound gestures, e.g. entire audio samples. With physical modeling sound synthesis, a sound control gesture could be defined as an entire excitation pattern, subdivided into multiple phases (e.g. attack, sustain, ...), each associated with sound control segments.

3.4 Multilevel mapping

The hierarchical representations of sound and gesture offer a rich structure to design mapping strategies.

In particular, we aim at integrating various temporal levels in the mapping process by defining input-output relationships for each time scale of the hierarchical models. As illustrated in Figure 2, the mapping relationships at a given instant can be seen as a superimposition of "instantaneous", short-term and long-term mappings.

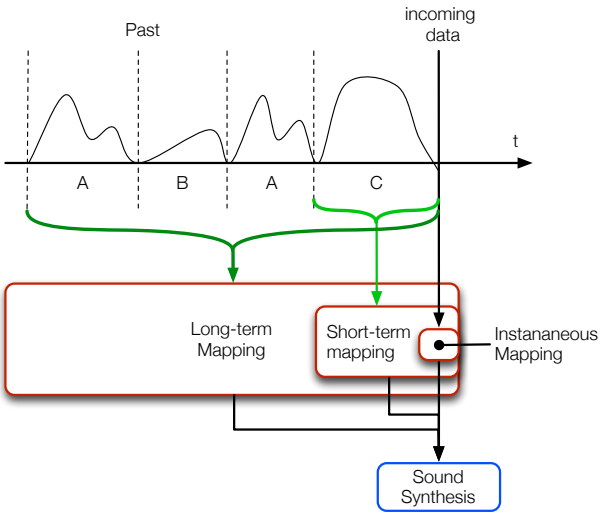

Figure 2. General architecture of the multilevel mapping strategies

For example, one possible strategy consists in the use of the recognition results to trigger various settings of the instantaneous mappings. In such case, the short-term mapping can act as an automatic selector on instantaneous mappings.

Other examples can be built from the concept of temporal mapping introduced in [18], where each gesture segment can be synchronized to sound control segments. The section 5 will further provide with a concrete case based on such concepts.

4. GESTURE SEGMENTATION USING HIERARCHICAL MARKOV MODELS

We explain in this section how Hierarchical Hidden Markov Models can be used in our framework. First, we recall generalities about Markov Models and second, we describe a specific implementation for gesture segmentation and recognition.

4.1 Model for gesture segmentation and recognition

4.1.1 HMM and extensions

Hidden Markov models (HMMs) have been widely used for time series analysis (e.g. gesture, speech, etc) since they are able to capture temporal dependencies between observations of an input time series through an underlying Markov process [25]. In typical HMM-based gesture recognition systems, a model is trained for each gesture using a set of examples. The hidden process is composed by several states encoding the temporal structure of the gesture. When used for segmentation and recognition in real-time, the gesture data is then compared to each model independently, and the model showing the highest likelihood determines the recognized gesture. However, considering the gesture models independently can be a limitation because of the lack of a high level structure modeling the transitions between gestures.

This problem has been addressed by a particular extension of HMMs proposed in the literature: the hierarchical HMM, that integrates multilevel representations of the temporal structures [26]. The hierarchical HMM can be represented as a tree with an arbitrary number of levels, composing a hierarchy of states from a single root to the leaf states which emit observations. In the hierarchy, each state is conditioned on its parents at the superior level. Thus, each hidden state is an autonomous model in the sense that it generates a lower-level Markov process rather than emitting a single observation.

In this paper, we present a two-level hierarchical HMM, that allows for considering fine-grain temporal modeling of gesture segments and long-term temporal relationships.

4.1.2 The two-level Hierarchical HMM

A two-level Hierarchical HMM requires the definition of two types of hidden states we call *segment* states and *signal* states, as illustrated in Figure 3.

In the first level under the root, *segment* states are associated with indexed gesture segments. A *segment* state generates a submodel encoding the fine temporal structure of the segment. Each submodel is itself composed by a time-ordered set of *signal* states that emit observations (associated to the gesture dataflow).

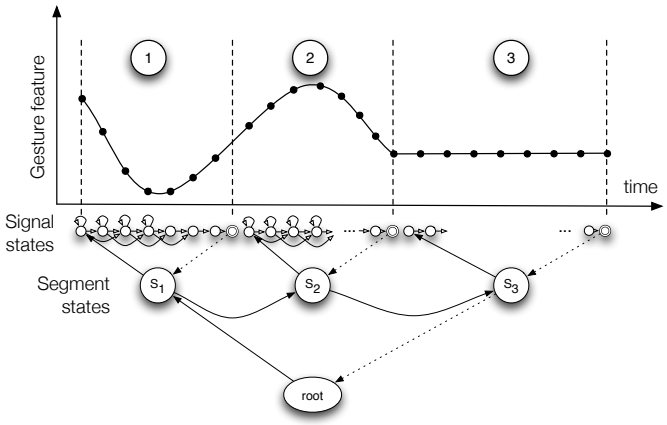

Figure 3. Training procedure of the hierarchical HMM. Given a pre-segmented gesture template, each segment is associated with an *segment* state S_i. The fine temporal structure of each segment is encoded in a submodel.

The model is built using a specific learning procedure. Consider a pre-segmented template gesture signal (e.g. annotated by a user) composed of several segments (Figure 3 shows a case of three segments). Each segment is associated with a *segment* state, denoted by S_i, and the temporal structure of the segment is encoded in a submodel with a left-to-right transition structure. In particular, each sample of the segment is associated to a hidden *signal* state with a gaussian observation probability with fixed variance. Hence, the learning phase only requires one template gesture to build the structure of the model. However, our implementation allows the use of multiple gestures to refine the reference template and learn the variance of the observation probability distributions.

Segmentation and recognition can be performed online thanks to a forward inference procedure [27]. For each incoming sample during the performance, the algorithm evaluates the likeliest *segment* state (i.e. the likeliest gesture segment) and the likeliest *signal* state (i.e. the temporal position within the segment). The algorithm also returns the likelihood of the recognized gesture that is used for recognition tasks.

4.1.3 Evaluation of the model for gesture segmentation

We conducted a comparative study of the performance for three models: *Gesture Follower* [9], the segmental HMM [24] and the two-level hierarchical model presented in this paper. We have evaluated the precision of the segmentation of the models with both offline and online algorithms on a database of simple gestures, performed with a hand-held device embedding accelerometers, and executed at various speeds. A comparison of offline segmentation using the Viterbi algorithm between our model and the segmental HMM highlights a greater ability of the hierarchical HMM to handle important distortions of gesture signals, in particular those introduced by non uniform speed variations during the performance of gestures – for example when the reference and the test gesture are performed at a different speed. For real-time gesture segmentation and recognition using a forward algorithm, the results show that the Hierarchical HMM outperforms *Gesture Follower*, the high-level process preventing errors at the transition between segments. As the evaluation of the model is not the major topic of this paper, we refer the interested reader to [28] for a detailed study covering the representation, implementation and evaluation of the hierarchical model and inference procedures for gesture segmentation.

4.2 Preparation-Attack-Sustain-Release: the PASR representation of gestures

We present in this section an example of decomposition of gestures as ordered sequences of primitives. Inspired from the traditional ADSR representation of sound envelopes – standing for *Attack*, *Decay*, *Sustain*, *Release*, – we introduce, as an example, a decomposition of gestures into 4 typical phases in gestures for sound control, defined as follows:

- **Preparation (P):** anticipation gesture preceding the beginning of the sound.
- **Attack (A):** segment covering the attack transient of the sound.
- **Sustain (S):** segment spanning from the decay to the end of the sound.
- **Release (R):** retraction gesture following the end of the sound.

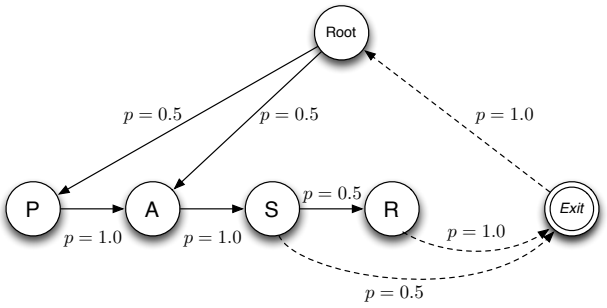

Figure 4. Topology of the PASR gesture models for 1 gesture. The prior probabilities ensure that the gesture can only be entered by the *Preparation* or *Attack* phases. Gesture models are left-to-right and reaching the exit state is only possible from the *Sustain* and *Release* phases.

Importantly, the user can choose among the different transition possibilities, making for example some segments optional (such as the preparation or release) or imposing constraints on the segments ordering or possible repetition.

Figure 4 illustrates the following case. The *segment* states are $S_1 = \mathrm{P}$, $S_2 = \mathrm{A}$, $S_3 = \mathrm{S}$ and $S_4 = \mathrm{R}$, and the parameters of the model are set to allow transitions in the sequential order. For each gesture segment, the prior probabilities, i.e. the probabilities on the starting *segment* states, are equally set to 0.5 on the P and A states, ensuring that the gesture can enter equally through the preparation or the attack phase. Within the gesture, transitions are defined from

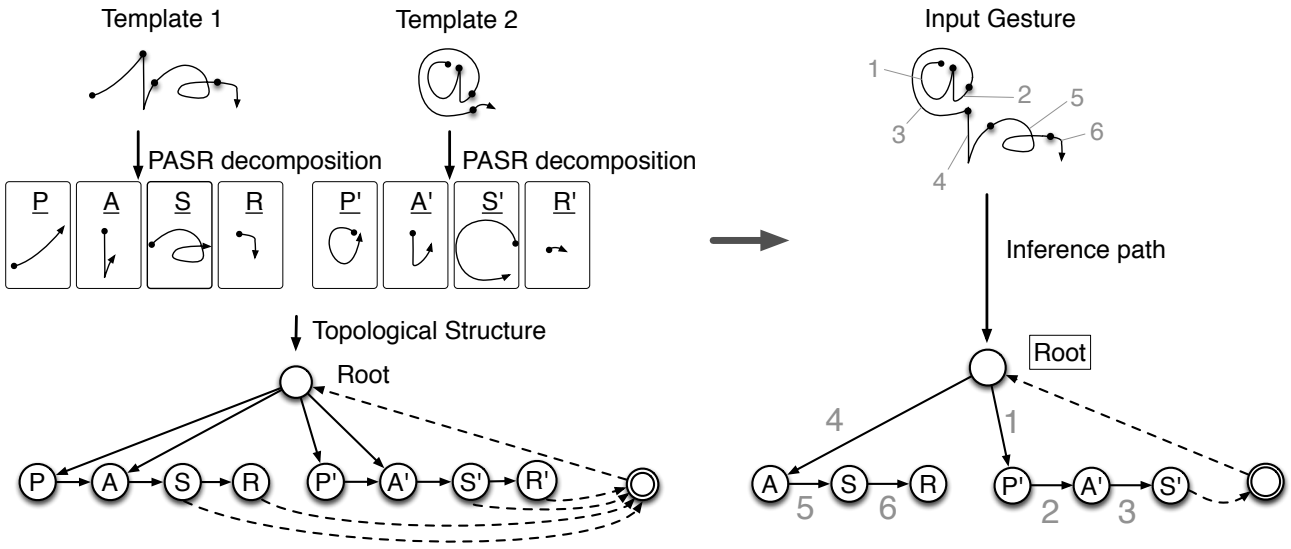

Figure 5. A practical example of the PASR decomposition of gestures. Two template gestures can be learned, represented at the top left of the figure. The decomposition of each gesture defines the structure of the model (bottom left). During performance, a continuous gesture can be performed by sequencing several segments of the original templates (top right). This induces a specific path in the topological graph (bottom right).

left to right to respect the sequential order. Finally, additional probabilities have to be set, which define the possibility of reaching the *exit* state – represented by a double circle on the figure – and go back to the root in order to enter another gesture. These probabilities are equal to 0.5 and 1 for the last two states of the model, restricting the possibility of ending a gesture through the sustain phase or the release phase.

Therefore, two modes are possible when performing sequences of gestures. Each gesture can be performed entirely, from the preparation to the release, or can be sequenced in a shorter form by avoiding the preparation and release segments. Thus, different transitions between gestures are made possible.

In Figure 5, we show an example of the decomposition of a complex gesture based on two gestures templates. On the top left of Figure 5, two different gesture templates are learned. Both are decomposed into the 4 phases P, A, S, and R, which define the topological structure of the two-level Hierarchical HMM, as previously introduced by Figure 4.

On the top right part of the figure, an input gesture is decomposed using the two templates. The inference process segments the input gesture and recognizes the gesture segments. This induces a path in the topological graph, depicted on the bottom right of Figure 5. Note that this type of information can be computed in real-time due to the forward inference.

5. APPLICATION IMPLEMENTATION

A concrete application based on our approach was prototyped and implemented in Max/MSP.

5.1 Application architecture

We focus in this application on the case of accelerometer-based sensors. In particular, we use interfaces called MO which include inertial sensors: 3D accelerometers and 3 axis gyroscopes [29]. Note that other type of gesture signal could also be used with our system.

The scheme presented in Figure 6 details the workflow diagram of the application and figure 7 shows a screenshot of the Max patch. The patch provides visualization and editing tools for both sounds and gesture signal, coupled with a control panel (using the MuBu environnement [30]). The control panel can be used to add or remove buffers, save and load presets, and play the sound (top of Figure 7).

We describe below first the learning mode, necessary to build the hierarchical gesture models from templates recorded by the user, and second, the performance mode, where the gesture segmentation and recognition process drives phase vocoder sound processes.

5.2 Learning phase

In the proposed application, the gesture segmentation is computed using the two-level Hierarchical HMM introduced in Section 4. The model has been implemented as an external object for Max/MSP called *hhmm* allowing to perform the multilevel gesture segmentation in real-time. The learning process requires a minimal set of templates recorded by the user.

A sound is represented by its waveform at the top of Figure 7. First, the user must add markers and label the segmentation on the audio buffer, to define the audio segments that will be linked to the gesture segments: Preparation, Attack, Sustain and Release (PASR) (phase (1) in Figure 6).

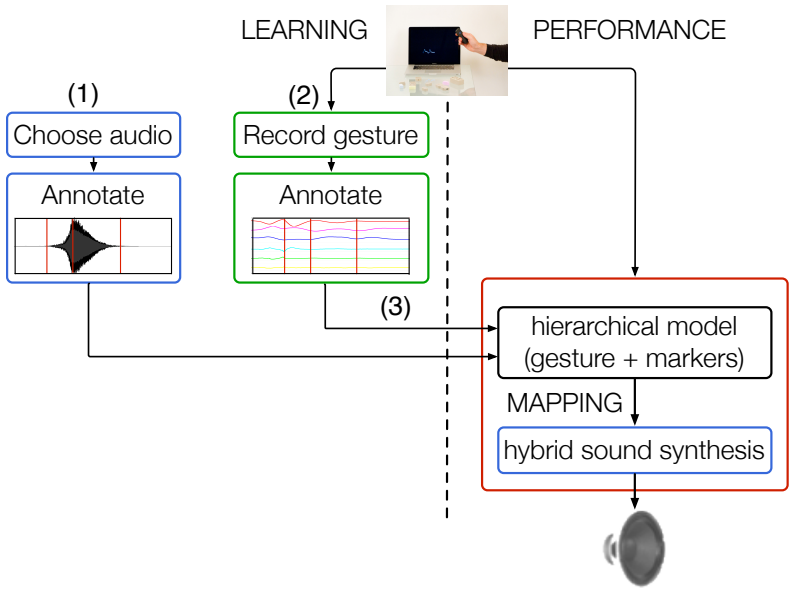

Figure 6. Workflow diagram of the application

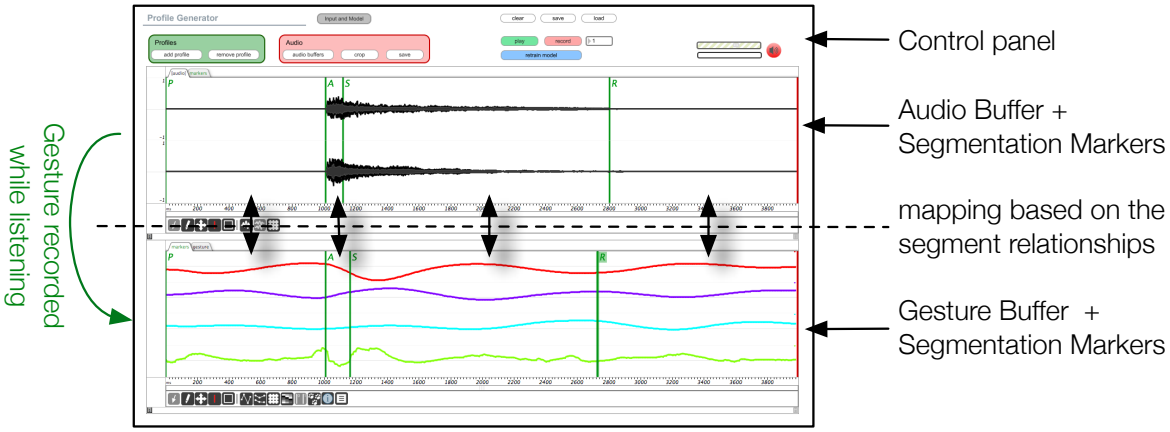

Figure 7. Screenshot of the Max Patch of the application.

Second, the user must perform a gesture, where the PASR decomposition can be operated. One possible strategy is to perform the gesture while listening to the sound, in order to induce structural similarities with the audio sample. This gesture is recorded in a gesture buffer, as shown at the bottom of Figure 7. As with the sound buffer, the gesture data must be annotated with a set of markers defining the P, A, S and R phases of the gesture (phase (2) in Figure 6). If the gesture was performed synchronously with the sound, the markers can be transferred from the audio buffer and re-edited to closely fit the timing of the gesture performance. Finally, the segmented gesture can be used to build to the hierarchical model (phase (3) in Figure 6), and specific messages are used to set the high level parameters (e.g. prior, transition, and exit probabilities) as specified in section 4.2, with respect to the PASR decomposition.

Finally, the user can switch to the performance mode and evaluate the quality of the control. At any moment, he can switch back to the learning mode to re-learn the model.

5.3 Performance phase

In performance mode, the gesture dataflow is segmented and labelled automatically (using the object *hhmm*) and this information is used to control sound synthesis in Max/MSP.

Precisely, the object outputs a set of parameters at each new observation: the likelihood of each gesture segments, the time progression and the estimated speed of the input gesture compared with the templates. Therefore, the object continuously updates the following information: the index of segment currently performed and the temporal position within the segment.

This information is used to control temporal dynamics of recorded sounds, mixing sampling and phase vocoder techniques. Technically, we use *superVP* in conjunction with the Mubu objects [30], to build this modular real-time synthesis engine of annotated audio samples. At each time step, gesture recognition is used to interactively select and time-stretch the audio segments according to the esti-

mation of the temporal alignment on the reference. The segmental annotation of audio samples is used to design specific settings adapted to each type of gesture segments. Typically, the *Preparation* is linked to silence, *Attack* to a non-stretchable sound segment, *Sustain* to a stretchable sound segment, and *Release* to fading effect. Sustain segments are thus stretched or shortened whereas attack phases are played at the initial speed. In specific case when the attack phase of the gesture is longer than that of the sound, the end of the gesture segment is time stretched to smooth the transition between audio processes.

Globally, we found that the high level transition structure between segments enables to plan and execute gestures sequences with a high level of control. Notably, the gesture recognition is improved compared with previous systems [28], the addition of the high level transition structure being efficient at disambiguating between gesture segments.

6. CONCLUSION AND FUTURE DIRECTIONS

We described a new approach for gesture-to-sound mapping which aims at overcoming the limitation of instantaneous strategies by integrating different time-scales in the mapping process. Precisely, we have proposed a multilevel mapping method based on hierarchical representations of both gesture and sound. As an example, a gesture segmentation method based on the Hierarchical HMM has been detailed and implemented, supported by a sequential decomposition of gestures called PASR. Finally we have presented an application for the control of audio processing.

If we described a special case of mapping strategy based on the alignment on gestures over audio samples, we believe that the hierarchical approach can generalize to other types of mappings and sound synthesis. We are currently experimenting with physical modeling sound synthesis, superimposing instantaneous mapping strategies at various levels of the hierarchy. For example, the proposed decomposition can be used to modulate the types of excitation of a physical model given particular gestures, with possibilities of characterizing the sustain phase of continuous excitation or the preparation preceding a strike, bringing contextual information to the mapping process.

Finally, we aim at generalizing such machine learning algorithms, in order to learn both temporal and spatial aspects of the mapping. Precisely, we are working on a generalization of the learning approach by developing multimodal methods able to learn jointly the gesture and sound models, and to characterize their relationships. Another major prospect is to conduct evaluations, both computational and from a user's viewpoint, in order to quantify the benefits of the proposed systems in terms of expressivity and personalization.

7. ACKNOWLEDGMENTS

We acknowledge support from the project ANR - LEGOS (11 BS02 012). We thank all members of the Real Time Musical Interactions team and members of the Legos project for fruitful discussions.

8. REFERENCES

[1] A. Hunt, M. M. Wanderley, and M. Paradis, "The importance of parameter mapping in electronic instrument design," in *Proceedings of the 2002 conference on New interfaces for musical expression*. National University of Singapore, 2002, pp. 1–6.

[2] J. Rovan, M. M. Wanderley, S. Dubnov, and P. Depalle, "Instrumental Gestural Mapping Strategies as Expressivity Determinants in Computer Music Performance," in *Kansei, The Technology of Emotion. Proceedings of the AIMI International Workshop.*, 1997, pp. 68–73.

[3] D. Van Nort, M. M. Wanderley, and P. Depalle, "On the Choice of Mappings Based on Geometric Properties," in *Proceedings of the 2004 Conference on New Interfaces for Musical Expression*, Hamamatsu, Japan, 2004.

[4] M. A. Lee and D. Wessel, "Connectionist models for real-time control of synthesis and compositional algorithms," in *Proceedings of the International Computer Music Conference (ICMC)*, San Jose, CA, 1992, pp. 277–280.

[5] J. Chadabe, "The Limitations of Mapping as a Structural Descriptive in Electronic Instruments," in *Proceedings of the 2002 conference on New interfaces for musical expression*. National University of Singapore, 2002, pp. 1–5.

[6] S. Fels and G. Hinton, "Glove-talk: A neural network interface between a data-glove and a speech synthesizer," *Neural Networks, IEEE Transactions on*, vol. 4, no. 1, pp. 2–8, 1993.

[7] A. Cont, T. Coduys, and C. Henry, "Real-time Gesture Mapping in Pd Environment using Neural Networks," in *Proceedings of the 2004 conference on New interfaces for musical expression*. National University of Singapore, 2004, pp. 39–42.

[8] F. Bevilacqua, R. Müller, and N. Schnell, "MnM: a Max/MSP mapping toolbox," in *Proceedings of the 2005 conference on New interfaces for musical expression*, 2005.

[9] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana, "Continuous realtime gesture following and recognition," in *Embodied Communication and Human-Computer Interaction, volume 5934 of Lecture Notes in Computer Science*. Springer Verlag, 2010, pp. 73–84.

[10] N. Gillian and R. Knapp, "A Machine Learning Toolbox For Musician Computer Interaction," in *Proceedings of the 2011 Conference on New Interfaces for Musical Expression*, 2011.

[11] R. Fiebrink, "Real-time human interaction with supervised learning algorithms for music composition and performance," Ph.D. dissertation, Faculty of Princeton University, 2011.

[12] J. Malloch, S. Sinclair, and M. M. Wanderley, "A Network-Based Framework for Collaborative Development and Performance of Digital Musical Instruments," in *Computer Music Modeling and Retrieval. Sense of Sounds*. Springer, 2008, pp. 401–425.

[13] J. Malloch, S. Sinclair, A. Hollinger, and M. Wanderley, *Input Devices and Music Interaction*, ser. Springer Tracts in Advanced Robotics. Springer Berlin / Heidelberg, 2011, vol. 74, pp. 67–83.

[14] B. Caramiaux, F. Bevilacqua, and N. Schnell, "Towards a gesture-sound cross-modal analysis," *Gesture in Embodied Communication and Human-Computer Interaction*, vol. 5934 of Lecture Notes in Computer Science, pp. 158–170, 2010.

[15] D. J. Merrill and J. A. Paradiso, "Personalization, expressivity, and learnability of an implicit mapping strategy for physical interfaces," in *Proc of CHI 2005 Conference on Human Factors in Computing Systems*. Citeseer, 2005, p. 2152.

[16] R. Fiebrink, P. R. Cook, and D. Trueman, "Play-along mapping of musical controllers," in *In Proceedings of the International Computer Music Conference (ICMC)*, 2009.

[17] F. Bevilacqua and N. Schnell, "Wireless sensor interface and gesture-follower for music pedagogy," *Proceedings of the 7th international conference on New interfaces for musical expression*, pp. 124–129, 2007.

[18] F. Bevilacqua, N. Schnell, and N. Rasamimanana, "Online Gesture Analysis and Control of Audio Processing," in *Musical Robots and Interactive Multimodal Systems*. Springer, 2011, pp. 127–142.

[19] S. Jordà, "Digital Lutherie: Crafting musical computers for new musics performance and improvisation," Ph.D. dissertation, UPF, 2005.

[20] ——, "On Stage: the Reactable and other Musical Tangibles go Real," *International Journal of Arts and Technology*, vol. 1, pp. 268–287, 2008.

[21] B. Caramiaux, "Etudes sur la relation geste-son en performance musicale," PhD Dissertation, Ircam - Université Pierre et Marie Curie, 2012.

[22] R. I. Godoy, A. R. Jensenius, and K. Nymoen, "Chunking in music by coarticulation," *Acta Acustica united with Acustica*, vol. 96, no. 4, pp. 690–700, 2010.

[23] J. Bloit, N. Rasamimanana, and F. Bevilacqua, "Modeling and segmentation of audio descriptor profiles with segmental models," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1507–1513, Sep. 2010.

[24] B. Caramiaux, M. M. Wanderley, and F. Bevilacqua, "Segmenting and Parsing Instrumentalist's Gestures," *Journal of New Music Research*, vol. 41, no. 1, pp. 13–29, 2012.

[25] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[26] S. Fine, Y. Singer, and N. Tishby, "The hierarchical hidden Markov model: Analysis and applications," *Machine learning*, vol. 32, no. 1, pp. 41–62, 1998.

[27] K. P. Murphy and M. A. Paskin, "Linear Time Inference in Hierarchical HMMs," in *Advances in Neural Information Processing Systems*, 2001.

[28] J. Françoise, "Realtime Segmentation and Recognition of Gestures using Hierarchical Markov Models," Master's Thesis, Université Pierre et Marie Curie, Ircam, 2011.

[29] N. Rasamimanana, F. Bevilacqua, N. Schnell, E. Fléty, and B. Zamborlin, "Modular Musical Objects Towards Embodied Control Of Digital Music Real Time Musical Interactions," *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction*, pp. 9—-12, 2011.

[30] N. Schnell, A. Röbel, D. Schwarz, G. Peeters, and R. Borghesi, "Mubu & friends - assembling tools for content based real-time interactive audio processing in max/msp," in *Proceedings of International Computer Music Conference*, Montreal, 2009.