Gesture–Sound Mapping by Demonstration in Interactive Music Systems

Jules Françoise
STMS Lab — IRCAM–CNRS–UPMC
1, Place Igor Stravinsky 75004 Paris, France
jules.francoise@ircam.fr
Advised by Frédéric Bevilacqua

ABSTRACT

In this paper we address the issue of mapping between gesture and sound in interactive music systems. Our approach, we call *mapping by demonstration*, aims at learning the mapping from examples provided by users while interacting with the system. We propose a general framework for modeling gesture–sound sequences based on a probabilistic, multimodal and hierarchical model. Two orthogonal modeling aspects are detailed and we describe planned research directions to improve and evaluate the proposed models.

Categories and Subject Descriptors

H.5.5 [**Information Interfaces And Presentation**]: Sound and Music Computing; J.5 [**Arts and Humanities**]: Music

Keywords

mapping, gesture, sound synthesis, music performance, HMM, multimodal, hierarchical modeling

1. INTRODUCTION

Gesture interaction with audio and/or visual media has become ubiquitous, requiring the development of intuitive software solutions for interaction design. In this paper, we focus on gesture–based interactive systems for sound and music performance. Today, the range of applications of such systems is widening, extending beyond the musical context to novel gaming applications [11], sonic interaction design to facilitate objects manipulation [12], or rehabilitation where interactive sound feedback could inform users on their movement performance.

In this paper we address current issues in the design of the relationship (often called *mapping*) between gesture features and sound control parameters in such interactive systems (cf. figure 1). Mapping has been recognized a crucial aspect of interactive music systems as it determines the degree of control accessible to a user interacting gesturally with sound processes. Indeed, specific choices in the design of such mapping impact on the interaction possibilities, for example in terms of ease-of-use, expressivity, or metaphor.

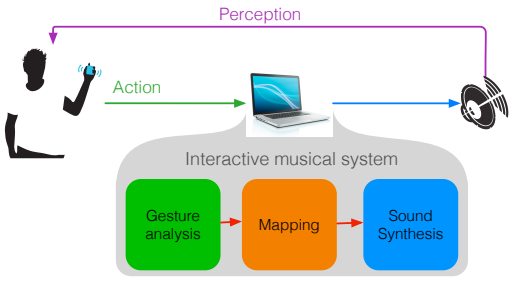


Figure 1: Overview of a gesture–based interactive music system. The *Mapping* relates gesture features to sound control parameters.

This work investigates an approach we call *mapping by demonstration*[1], aiming to learn the mapping from examples provided interactively by the user, e.g. gestures performed while listening to sound examples. A typical situation involves a two-step interaction loop integrating *training* and *performance* (see Figure 2). During the *training* phase, a multimodal model of the mapping is trained on a set of user demonstrations. Once learned, the model can be used for *performance*: sound control sequences are predicted in real-time based on new gestures. This situation involves challenging research questions in terms of modeling, considering the complexity and diversity of possible mappings, and in terms of machine learning, regarding the need for efficient and interactive learning algorithms.

Section 2 details related work about mapping in music and multimedia. The research questions and specificities of our approach are detailed in section 3. Current research is summarized in section 4, and prospective work is described in section 5.

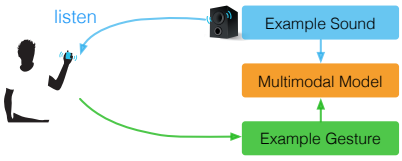
2. RELATED WORK

2.1 Gesture–Sound Mapping by Demonstration

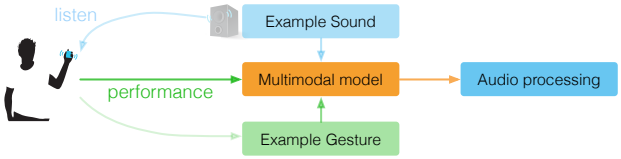
This section details current approaches to mapping by demonstration. We identify three classes of approaches according to two properties of the learned mapping: the tem-

[1]This terminology is inspired by the field of robotics called *Programming by demonstration* [3].

MM'13, October 21-25, 2013, Barcelona, Spain



(a) *Training* phase.



(b) *Performance* phase.

Figure 2: Workflow of a Mapping by Demonstration system.

poral aspect, and the discrete/continuous nature of interactions.

Various methods have been proposed to learn non-linear continuous mappings between gesture parameters and sound control parameters. Regression methods, in particular neural networks, have been extensively used for mapping by demonstration [10]. In particular, Fiebrink et al. proposed a *play-along* paradigm that allows to define a training set interactively by performing gestures associated with sound parameter sequences [5]. This class of methods, we call *Regression Mapping*, are often limited to static relationships between gesture and sound parameters.

Other approaches have investigated gesture recognition methods to consider the temporal aspect of the relationships between gesture and sound. We call this class of approaches *temporal mapping* and distinguish between discrete and continuous strategies. In *Discrete Temporal Mapping*, discrete gesture recognition is used to identify gestures carrying a particular meaning or metaphor, in order to trigger sound controls. Several gesture recognition techniques have been studied such as Hidden Markov Models (HMMs), Dynamic Time Warping (DTW) and Support Vector Machines (SVMs) [9]. *Continuous Temporal Mapping* refers to strategies using continuous–time models of gestures in relationship with sounds, therefore allowing to align gesture time profiles to sound morphologies [1]. The method is based on continuous gesture recognition — also called *gesture following* — using HMMs [2]. The method allows accurate control on the temporal dimension, but is limited by a one-shot learning procedure preventing the integration of expressive variations occurring in several interpretations of the same gesture.

If regression methods are powerful to learn non-linear parameter mappings from multiple demonstrations, they do not or poorly model the temporal relationships between gesture and sound. Discrete temporal mapping partially addresses this issue, but restricting gesture analysis to classification limits control strategies to simple triggering or selection. Finally, if continuous temporal mapping methods model accurately the temporal relationship between modalities, they poorly handle the variability occurring between several performances of the same gesture. Moreover, in this case gestures are modeled as continuous time profiles, limiting the integration of higher-level temporal representations related to musical sequences. Our goal is to address the limitations of current approaches by adopting a hierarchical and multimodal perspective on mapping inspired by recent work in other fields of multimedia.

2.2 Mapping in Other Fields of Multimedia

Many recent systems for virtual character animation use speech–driven approaches, involving mapping from acoustic features to lip, face or body motion [4]. Similar issues are found in acoustic–articulatory inversion [13], aiming at the retrieval of speech–producing movements from acoustic signals. Most of the current approaches are based on sequence models representing the output process in conjunction with the input modality. In particular, various extensions of HMMs were proved efficient for feature mapping [8].

In spite of their similarity, our applications in gestural sound control contrasts in critical points the applications we just mentioned. While speech-related movements and sounds are generated by the same action and thus intrinsically related, the gesture–sound relationships can be arbitrarily defined by the user. In many cases, the training examples should be provided by the user in a *mapping by demonstration* workflow integrating training and performance. These aspects require adapting models and algorithms to an interactive learning paradigm.

3. RESEARCH QUESTIONS AND PROPOSED APPROACH

We aim at overcoming the limitations of current approaches by the development of a general modeling framework for mapping integrating the following properties:

1. **temporal**: each example is a particular instance, intrinsically dynamic, of the underlying mapping. Therefore, learning from user demonstrations require modeling the temporal aspect of the mapping.

2. **expressive variations**: the model must be able to abstract not only the average time structure of gesture–sound segments but also the expressive variations arising between several interpretations of a same mapping.

3. **hierarchical**: in order to integrate high level representation of musical structures, the model must consider multiple time scales and build the mapping accordingly using hierarchical representations. Learned mappings must provide both accurate and high–level degrees of control.

Moreover, we stress the importance of the constraints related to the interactive learning situation required by a mapping by demonstration approach. The training examples are interactively demonstrated by the user. Therefore, the model must be trained from few examples. Moreover, during the *performance* phase, the prediction algorithm must work in real-time to ensure instantaneous feedback to the user's movements.

Our goal is to propose a fully probabilistic, multimodal and hierarchical sequence model of gesture–sound mappings addressing the above constraints. An overview of the general approach is represented on figure 3. Our approach takes advantage of multilevel segmentation of both gesture and

sound to implement mapping strategies spanning over different time scales, from fine control to higher level strategies. In the next section we detail two contributions addressing orthogonal aspects of this framework. The former details a hierarchical approach to mapping, while the latter focuses on multimodal modeling.

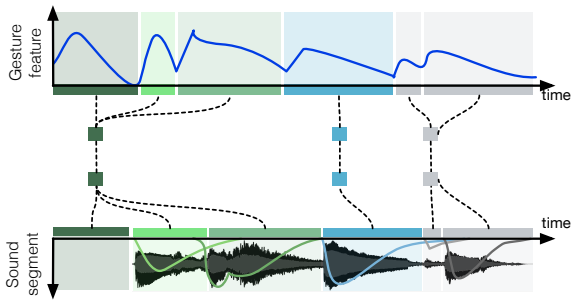


Figure 3: Overview of the general approach. The modeling framework is based on multilevel segmentations of gesture and sound to implement mapping strategies with different time granularity.

4. CONTRIBUTIONS

4.1 Hierarchical Approach to Mapping

This work follows recent developments in the Real-Time Musical Interactions team at Ircam regarding continuous temporal mapping. Our goal is to extend current systems for continuous gesture recognition [2] in order to integrate higher level representations of musical sequences. This work is described in [6] and addresses the issue (3) highlighted in section 3.

We propose the use of a Hierarchical HMM with two levels for real-time gesture segmentation and recognition. We adopt a template–based learning procedure allowing to build the model from a single example, as illustrated on figure 4. From a segmented example gesture, we build the model by associating each segment to a high level state, which generates a submodel encoding the time structure of the segment. The submodel is built by associating to each frame of the example a state in a left–right transition structure. The high–level transition structure can be authored by the user. Segmentation and recognition are then performed in real-time using a forward algorithm, estimating for each incoming observation the likeliest segment and the time alignment of the gesture to the reference.

The approach offers interesting possibilities for gesture–based control of sound synthesis. First, the recognition is improved because informed by the high level structure. Moreover, the possibilities of authoring the high level transition probabilities enable user to define sequence models or particular representations of gestures (see for example the PASR[2] representation we proposed in [6]). Finally, the segmental representation of sound–related gestures provides a means to define interactively some constraints on the sound synthesis for particular segments — e.g. imposing transient conservation on the attack phases of the sound.

[2]PASR is a representation of gestures as a sequence of 4 segments: Preparation-Attack-Sustain-Release

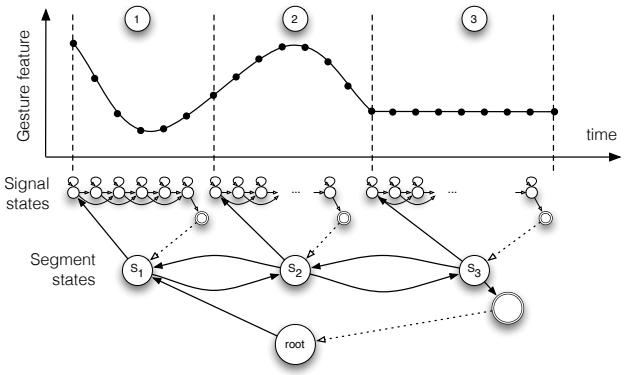


Figure 4: The two-level hierarchical HMM. The model is built from a single segmented example gesture.

4.2 Multimodal Modeling of Gesture–Sound Mapping

This section details another aspect of modeling addressing the constraints (1) and (2) highlighted in section 3. We support a multimodal perspective on mapping by introducing a probabilistic model able to learn gesture–sound mapping from several examples. The system is based on a multimodal HMM modeling the dependencies between gesture and sound. A more detailed description of the model with applications to gesture–based control of physical modeling sound synthesis can be found in [7].

The model assumes that gesture and sound are generated by the same underlying markov process. During training, gesture and sound feature vectors are concatenated to form multimodal sequences (figure 5, left). The observation probability distributions are therefore defined as joint multimodal Gaussian distributions with full covariance. The model is trained using an EM algorithm with pre-estimated parameters to ensure fast and efficient convergence. For prediction, the joint observation distributions are converted to conditional distributions. The sound observation sequence associated to a new gesture sequence can be predicted in real-time using a forward algorithm (figure 5, right).

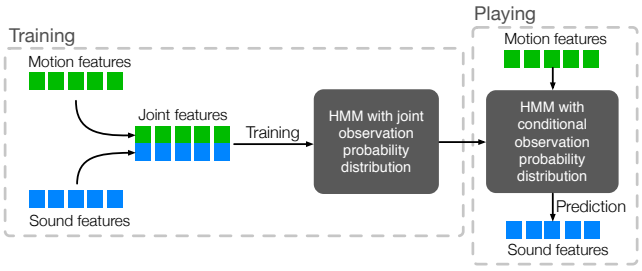


Figure 5: Mapping with the multimodal HMM.

A preliminary assessment of the system, based on informal observation and evaluation on synthetic signals, has been performed and shows promising results. In comparison with the template–based model introduced in [1], the multimodal HMM show a better ability to model the expressive gesture variations between different interpretations, guaran-

teeing a more coherent translation of gesture variations to sound variations.

5. PLANNED WORK

5.1 Modeling

We plan to address several issues of the proposed models. First, the convergence of the training procedure of the multimodal HMM might not be guaranteed in cases where the number of examples is critically limited. We are currently improving this limitation by examining various approaches for adding constraints in the training procedure. Moreover, the smoothness and consistency of the prediction could be improved by the integration of a trajectory model, as proposed by Zen [13]. Second, the two contributions presented in section 4 address orthogonal dimensions of the general approach we describe in section 3. The multimodal HMM precisely models the temporal and spatial relationships between gesture and sound at the segment level, while the hierarchical HMM provides higher–level control of musical sequences. We are currently working on a combination the two models to form a hierarchical multimodal model. Finally, in the current implementation, the high level structure of the hierarchical HMM can only be edited by the user. We plan to investigate online learning methods to adapt incrementally both the segment models and the transition parameters during the performance phase.

5.2 Evaluation

For now, the proposed models have not been extensively evaluated. One difficulty resides in the specificities of the context of music performance, for which there is no unique relationship between modalities but the design itself of this mapping is part of the creative process. We are planning two directions for the evaluation of the systems:

1. **Computational**: the models will be evaluated on a multimodal database implementing several specific mapping strategies. The models will be compared to state of the art methods regarding their ability to learn a mapping incrementally and to model temporal dynamics and expressive gesture variations.

2. **User Evaluation**: we plan to conduct two types of experiments. First, using a database similar to the one for computational evaluation, a qualitative assessment of the prediction of the algorithms will be performed by a panel of experts on subjective criteria. The second type of experiment will investigate user interaction with the machine learning models. Specifically, we plan to study how the user and the system can co-adapt and converge during the process of designing mapping strategies.

5.3 Applications

We plan to develop several concrete applications of the models in the contexts of music performance, sonic interaction design, and rehabilitation. In the framework of the ANR project *LEGOS*[3], we are investigating the benefits of auditory feedback for sensori-motor learning. In this case, the model will be used to learn mapping strategies for movement sonification, where the sound feedback would inform users on their movement performance.

[3] `http://legos.ircam.fr`

6. ACKNOWLEDGMENTS

We acknowledge support from the French National Research Agency (ANR project LEGOS 11 BS02 012).

7. REFERENCES

[1] F. Bevilacqua, N. Schnell, N. Rasamimanana, B. Zamborlin, and F. Guédy. Online Gesture Analysis and Control of Audio Processing. In *Musical Robots and Interactive Multimodal Systems*, pages 127–142. Springer, 2011.

[2] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana. Continuous realtime gesture following and recognition. *Gesture in Embodied Communication and Human-Computer Interaction*, pages 73–84, 2010.

[3] A. Billard, S. Calinon, R. Dillmann, and S. Schaal. Robot programming by demonstration. *Handbook of robotics*, 1, 2008.

[4] K. Choi, Y. Luo, and J.-n. Hwang. Hidden Markov model inversion for audio-to-visual conversion in an MPEG-4 facial animation system. *The Journal of VLSI Signal Processing*, 29(1):51–61, 2001.

[5] R. Fiebrink, P. R. Cook, and D. Trueman. Play-along mapping of musical controllers. In *In Proceedings of the International Computer Music Conference*, 2009.

[6] J. Françoise, B. Caramiaux, and F. Bevilacqua. A Hierarchical Approach for the Design of Gesture-to-Sound Mappings. In *Proceedings of the 9th Sound and Music Computing Conference*, pages 233–240, Copenhagen, Denmark, 2012.

[7] J. Françoise, N. Schnell, and F. Bevilacqua. A Multimodal Probabilistic Model for Gesture-based Control of Sound Synthesis. In *Proceedings of the 21st ACM international conference on Multimedia (MM'13)*, Barcelona, Spain, 2013.

[8] S. Fu, R. Gutierrez-Osuna, A. Esposito, P. Kakumanu, and O. Garcia. Audio/visual mapping with cross-modal hidden Markov models. *Multimedia, IEEE Transactions on*, 7(2):243–252, Apr. 2005.

[9] N. Gillian and R. Knapp. A Machine Learning Toolbox For Musician Computer Interaction. In *proceedings of the International Conference on New Interfaces for Musical Expression*, Oslo, Norway, 2011.

[10] P. Modler. Neural Networks for Mapping Hand Gestures to Sound Synthesis parameters. *Trends in Gestural Control of Music*, pages 301–314, 2000.

[11] N. Rasamimanana, F. Bevilacqua, J. Bloit, N. Schnell, E. Fléty, A. Cera, U. Petrevski, and J.-L. Frechin. The urban musical game: using sport balls as musical interfaces. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, pages 1027–1030, Austin, Texas, USA, 2012.

[12] D. Rocchesso, S. Serafin, F. Behrendt, N. Bernardini, R. Bresin, G. Eckel, K. Franinovic, T. Hermann, S. Pauletto, P. Susini, and Y. Visell. Sonic interaction design: sound, information and experience. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, pages 3969–3972, 2008.

[13] H. Zen, Y. Nankaku, and K. Tokuda. Continuous Stochastic Feature Mapping Based on Trajectory HMMs. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(2):417–430, 2011.