A Multimodal Probabilistic Model for Gesture–based Control of Sound Synthesis

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

Norbert Schnell
STMS Lab
IRCAM–CNRS–UPMC
1, Place Igor Stravinsky 75004
Paris, France
norbert.schnell@ircam.fr

Frédéric Bevilacqua
STMS Lab
IRCAM–CNRS–UPMC
1, Place Igor Stravinsky 75004
Paris, France
frederic.bevilacqua@ircam.fr

ABSTRACT

In this paper, we propose a multimodal approach to create the *mapping* between gesture and sound in interactive music systems. Specifically, we propose to use a multimodal HMM to conjointly model the gesture and sound parameters. Our approach is compatible with a learning method that allows users to define the gesture–sound relationships interactively. We describe an implementation of this method for the control of physical modeling sound synthesis. Our model is promising to capture expressive gesture variations while guaranteeing a consistent relationship between gesture and sound.

Categories and Subject Descriptors

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

Keywords

music, gesture, sound synthesis, music performance, HMM, multimodal

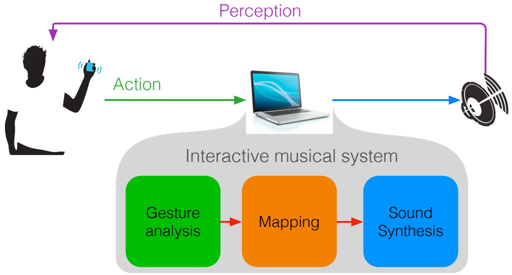


Figure 1: Overview of an interactive musical system. The *mapping* defines the coupling between actions and sound processing.

1. INTRODUCTION

Gestural interaction with audio and/or visual media has become ubiquitous. The development of gestural systems for sound control concerns not only musical applications, but also novel gaming applications [10], sonic interaction design to facilitate objects manipulation [11] or rehabilitation where interactive sound feedback could inform users on their movement performance.

In this paper, we focus on gesture–based control of sound synthesis targeting applications to music performance and movement sonification for rehabilitation. Specifically, we address current issues in designing the *relationship* (often called *mapping*) between gesture features and sound control parameters in such interactive systems (cf. figure 1).

Our research follows recent approaches based on machine learning which aim at defining the mapping between gesture and sound from examples rather than by an analytical formulation. In this paper, we support a multimodal perspective on mapping by introducing a probabilistic model of the gesture–sound relationships. The system is based on a single multimodal Hidden Markov Model (HMM) representing both gesture and sound parameter morphologies. The model is trained by one or multiple gesture performances associated to sound templates. It captures the temporal structure of gesture and sound as well as the variations which occur between multiple performances.

Section 2 gives a short overview of existing research and issues related to gesture–sound mapping in musical instruments. Related works in other fields of multimedia are described in section 3. The multimodal HMM is described in section 4, with specific implementation details. Finally, section 5 presents an application of the model to gesture–based control of physical modeling sound synthesis.

2. GESTURE-SOUND RELATIONSHIPS IN DIGITAL MUSICAL INSTRUMENTS

Creating relationships between gesture and sound parameters has been recognized one of the crucial aspects of the design of digital musical instruments. Specific choices in this design influence the interaction possibilities, varying in terms of ease–of–use, expressivity, and learning possibilities. Most approaches require an analytical definition of the relationships between gesture and sound, for example using *explicit* parameter wiring or *implicit* models, e.g. dynamic systems.

.

Recent works support interaction–driven approaches that take advantage of machine learning to define the mapping by *demonstration*; i.e. from training examples provided interactively by the user. Notably, number of strategies exploit gesture recognition methods to link the identification of user-specific gestures to sound control [7, 5]. Fiebrink [4] recently proposed a *play-along* paradigm enabling users to edit the training examples by performing gestures while listening to sound examples.

We propose a similar approach, based on machine learning, allowing users to define the mapping by direct demonstration of their intended control strategy: the mapping is learned from examples of the relationships between gesture and sound. This implies several requirements. First, for the method to remain practical, the model should be trained with a limited number of examples. Second, gesture and sound parameters, as well as their relationship, evolve over time. A computational model must take into account these temporal variations. Finally, the model should be able to characterize the expressive variations between multiple performances of the same gesture.

Previous works already partially addressed these issues. Systems based on regression methods such as neural networks [4, 5], for example, generalize from multiple interpretations but are often limited to static relationships. Gesture recognition methods often model the temporal dynamics of gestures but are usually limited to classification, restricting mapping strategies to discrete interactions such as triggering [7, 5]. The *temporal mapping* method we recently proposed can learn the temporal relationship between a gesture and a sound using continuous gesture recognition and alignment [1]. Nevertheless, the method is limited to modeling the gesture only. Moreover, the model does not allow for characterizing expressive inter–performance variations. Our goal is to address such limitations by adopting a multimodal perspective on gesture–sound mapping through the introduction of a fully probabilistic multimodal model.

3. RELATED WORKS 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 [2, 13, 3]. Similar issues are found in speaker conversion [14] or acoustic–articulatory inversion [8, 14], 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, such as HMM remapping [2], the input-output HMM [6] and the multimodal HMM (sometimes called audio-visual HMM or HMM inversion) [8, 13, 3].

In spite of their similarity, our applications in gestural sound control contrast in critical points the applications we just mentioned. While the movements and sounds related to speech are generated by the same action and thus intrinsically related, the gesture–sound relationship can be arbitrarily defined by the user. Hereby, the choice might depend on the application, which limits the practicality of creating extensive databases. In many cases, the training examples should be provided by the user within a single workflow integrating training and performance. These aspects require adapting models and algorithms to an interactive learning paradigm, as argued by Fiebrink [5].

4. GESTURE–SOUND MAPPING WITH MULTIMODAL HMM

This section details the multimodal HMM and the specificities of our implementation, designed to fit the context of real-time music performance. In the remainder of this article, *gesture features* or *gesture observations* refer to the parameters extracted from a gesture capture system. Similarly, *sound features* or *sound observations* refer to the vector of control parameters of a sound synthesis model.

4.1 Multimodal HMM

The multimodal HMM is an extension of traditional HMM for handling multimodal data, allowing for the prediction of missing features. We assume that gesture and sound are generated by the same underlying Markov process by representing jointly their observation sequences. The Dynamic Bayesian Network representation of the model is presented in figure 2, where $\mathbf{q}_t$ is the hidden states at time t, $\mathbf{o}_t^g$ (resp. $\mathbf{o}_t^s$) is the gesture (resp. sound) observation vector.

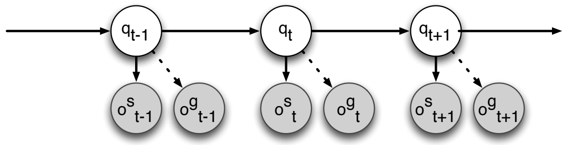


Figure 2: Graphical model of the Multimodal HMM.

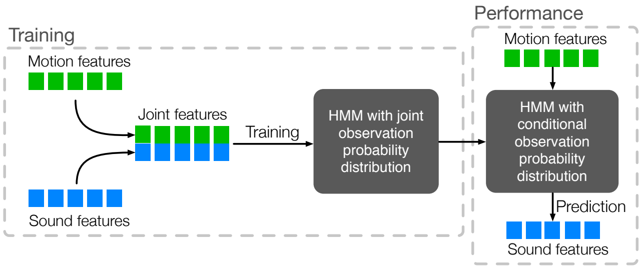


Figure 3: Gesture–sound mapping with the multimodal HMM.

For training, gesture and sound feature vectors are concatenated to form a multimodal sequence (cf. figure 3, left). The observation probability distribution at time t of a state j of the HMM with model parameters λ is defined as a joint multimodal Gaussian distribution:

$$P(\mathbf{o}_t^g, \mathbf{o}_t^s | q_t = j, \lambda) = \mathcal{N}\left([\mathbf{o}_t^g, \mathbf{o}_t^s]\,; \mu_j, \mathbf{\Sigma}_j\right)$$

where the mean of the distribution μ_j is the concatenation of the mean vectors for each modality, and $\mathbf{\Sigma}_j$ is a covariance matrix which can be decomposed in four sub-matrices representing unimodal and cross-modal covariances between gesture and sound parameters:

$$\mu_j = \left[\mu_j^g, \mu_j^s\right]$$

$$\mathbf{\Sigma}_j = \left[\begin{array}{cc} \mathbf{\Sigma}_j^{gg} & \mathbf{\Sigma}_j^{gs} \\ \mathbf{\Sigma}_j^{sg} & \mathbf{\Sigma}_j^{ss} \end{array}\right]$$

The HMM is trained using an Expectation-Maximization algorithm on the multimodal observations sequences [9].

For prediction, the joint observation distributions are converted to conditional distributions by marginalizing over the gesture observations at time t:

$$p(\mathbf{o}_t^s | \mathbf{o}_t^g, q_t = j, \lambda) = \mathcal{N}\left(\mathbf{o}_t^s; \hat{\mu}_j^s(\mathbf{o}_t^g), \hat{\mathbf{\Sigma}}_j^{ss}\right)$$

where the mean $\hat{\mu}_j^s(\mathbf{o}_t^g)$ and covariance $\hat{\mathbf{\Sigma}}_j^{ss}$ of the sound are re-estimated by combining the learned mean of the sound with a linear regression over the gesture features:

$$\begin{aligned}\hat{\mu}_j^s(\mathbf{o}_t^g) &= \mu_j^s + \mathbf{\Sigma}_j^{sg}\left(\mathbf{\Sigma}_j^{gg}\right)^{-1}\left(\mathbf{o}_t^g - \mu_j^g\right) \\ \hat{\mathbf{\Sigma}}_j^{ss} &= \mathbf{\Sigma}_j^{ss} - \mathbf{\Sigma}_j^{sg}\left(\mathbf{\Sigma}_j^{gg}\right)^{-1}\mathbf{\Sigma}_j^{gs}\end{aligned}$$

Given a new gesture, the associated sequence of sound features is estimated by a real-time prediction algorithm presented in section 4.2.2 (figure 3, right).

4.2 Interactive learning and implementation specificities

4.2.1 Interactive learning workflow

An interactive learning situation places the user in constant interaction with the system. The workflow of a typical setup is an interaction loop divided in two steps:

1. *Training*: The user can edit the training set by demonstrating examples of gesture and sound *segments* (i.e. time profiles of parameters). The user can then adjust the parameters and train the model.
2. *Performance*: The user evaluates the results by interacting with the trained model, which synthesizes sonic feedback to his gestures in real-time.

This interaction paradigm enables the user to iteratively evaluate and adjust the model until it converges to his purposes. In this paper, we only consider "direct" evaluation of the results, i.e. user evaluation on subjective criteria during performance, as defined by Fiebrink [5].

4.2.2 Performance: real-time prediction of sound control parameters

Reactivity and low latency are strong requirements of a gesture–based music performance system: sound control parameters must be synthesized in real-time to provide instantaneous feedback to the user's gestures. This constraint prevents the use of either Viterbi estimation of the state sequence or an iterative estimation of the sound sequence based on the EM algorithm [13, 3]. Our implementation features a real-time prediction algorithm based on a causal estimation of state probabilities. At each incoming gesture observation $\mathbf{o}_t^g$, the sound feature vector $\mathbf{o}_t^s$ is estimated by maximum likelihood based on the state probabilities computed by a forward algorithm:

$$\mathbf{o}_t^s = \sum_{i=1}^{N} \alpha_t(i) \cdot \underset{\mathbf{o}_t^s}{\operatorname{argmax}}\left[p(\mathbf{o}_t^s | \mathbf{o}_t^g, q_t = i)\right]$$

where $\alpha_t(i) = P(\mathbf{o}_{1:t}^g, q_t = i | \lambda)$ is the probability of the partial gesture observation sequence $\mathbf{o}_{1:t}^g$ and state i at time t [9]. This causal estimation ensures a good reactivity of the system while guaranteeing the smoothness of the estimated sound feature sequence. Typically, 10ms latency is often considered sufficient to guarantee the consistency of the relationships between gesture and sound [12]. In our experiments with inertial sensors (see section 5), we reach a latency of approximately 5ms.

5. APPLICATION: GESTURE CONTROL OF PHYSICAL MODELING SOUND SYNTHESIS

5.1 Motivation

Physical modeling sound synthesis aims at simulating the acoustic behavior of physical objects. The gestural control of such sound physical models remains difficult since the captured gestural parameters are generally different from the physical input parameters of the sound synthesis algorithm. For example, sensing gestures using accelerometers might pose difficulties for controlling the physical model of a bowed string where force, velocity, and pressure are the expected input.

Our approach of defining gesture-sound relationship by an intermediate probabilistic model seems appropriate to tackle such an issue. We therefore propose a system for gesture–based control of physical models which relies on our interactive learning system. It allows users to craft gestural control strategies by demonstrating gestures associated with particular sounds designed with virtual physical instruments. Our goal is to provide a solution for quick prototyping and evaluation these control strategies using the multimodal HMM.

5.2 Application description

5.2.1 Modal synthesis

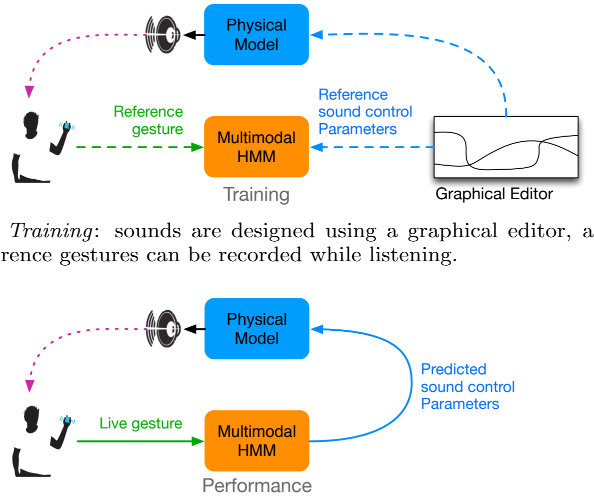
We use *Modalys*, a software dedicated to modal synthesis, i.e. that simulates the acoustic response of vibrating structures under an external excitation. Virtual instruments are built by combining *modal elements* – e.g. plates, strings, membranes – with various types of connections and excitators – e.g. bows, hammers, etc. Each model is governed by a set of physical parameters – e.g. speed, position and pressure of a bow. Specific sounds and playing modes can be created by designing time profiles combining these parameters.

5.2.2 Application Workflow

The workflow of the application fits in the interactive learning workflow described in section 4.2.1, and can be divided in a *training* phase and a *performance phase*, as illustrated in figure 4.

In the *training* phase, the user can draw time profiles of control parameters of the physical models to design particular sounds. Each of these *segments* can be visualized, modified, and played using a graphical editor. Then, the user can perform one or several demonstrations of the gesture he intents to associate with the sound example (figure 4a). In our experiments, gestures were captured using inertial sensors (3D accelerometers), and various synthesis models were used, such as a modified bowed string or a clarinet. Gesture and sound are recorded to a multimodal data container for storage, visualization and editing. Optionally, segments can be manually altered using the user interface.

During the *performance* phase, the user can gesturally control the sound synthesis. The system allows the explo-



(a) *Training*: sounds are designed using a graphical editor, and reference gestures can be recorded while listening.

(b) *Performance*: the model is used to predict the sound parameters associated with a live gesture.

Figure 4: Application workflow.

ration of all the parameter variations that are defined by the training examples. Sound parameters are predicted in real-time to provide the user with instantaneous audio feedback (figure 4b). If needed, the user can switch back to *training* and adjust the training set or model parameters.

6. DISCUSSION AND CONCLUSION

A preliminary evaluation of the system, based on both qualitative observations and computational evaluation on synthetic signals, has been performed and shows promising results. In comparison with the template–based model introduced in [1], the multimodal HMM tolerates larger variation occurring between the performance and the training data. Precisely, an evaluation of the algorithm on synthetic data indicates a better ability of the multimodal HMM to capture gesture variations in relationship with sound variations. This is related to the fact that the gesture–sound relationships are encoded globally by the transition structure of the HMM, and locally by the covariance matrices at each state. The model seems therefore able to capture both the temporal dynamics of the mapping and its expressive variations on between several interpretations.

We plan to perform more detailed evaluations of the model. First, the model will be evaluated on a database of sound-related gestures in order to assess its ability to capture expressive variations between gesture performances, and to evaluate the benefits of the temporal modeling in comparison with static regression methods. Second, the model will be evaluated in the context of movement sonification for rehabilitation. The model will be used to generate informative audio feedback to physical gestures in order to enhance motor learning, for example by continuously sonifying the range of variations between the gesture of a patient with motor disabilities and a target movement.

7. ACKNOWLEDGMENTS

We acknowledge support from the French National Research Agency (ANR project Legos 11 BS02 012). We thank Thierry Artières for fruitful discussions.

8. 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] M. Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28. 1999.

[3] 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.

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

[5] R. Fiebrink, P. R. Cook, and D. Trueman. Human model evaluation in interactive supervised learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 147–156. 2011.

[6] 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.

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

[8] G. Hofer. *Speech-driven animation using multi-modal hidden Markov models*. Phd dissertation, University of Edimburgh, 2009.

[9] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[10] 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. 2012.

[11] 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.

[12] D. Wessel and M. Wright. Problems and Prospects for Intimate Musical Control of Computers. *Computer Music Journal*, 26(3):11–22, Sept. 2002.

[13] E. Yamamoto, S. Nakamura, and K. Shikano. Speech-to-lip movement synthesis based on the EM algorithm using audio-visual HMMs. In *Multimedia Signal Processing, 1998 IEEE Second Workshop on*, pages 2–5, 1998.

[14] 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.