Mimicking Sound with Gesture as Interaction Paradigm

Baptiste Caramiaux, Frédéric Bevilacqua, Bruno Zamborlin and Norbert Schnell

IRCAM, CNRS-UMR STMS, 1 Place Igor Stravinsky, 75004 PARIS, France
{baptiste.caramiaux,frederic.bevilacqua,bruno.zamborlin,norbert.schnell}@ircam.fr

Abstract. We presented a novel approach based on actions mimicking sound for gestural control in interactive systems. In particular, we define two types of mimicking: *actions producing sound* and *actions describing sound*. A general architecture was implemented, using modules for action recognition and similarity measures between gesture and sound features. Case studies were performed in the form of a game using either a Wiiremote game controller or a mobile phone.

1 Introduction

The dominant paradigm in the design of gestural sound control is based on mapping sensor values to sound synthesis parameters. Different mapping strategies have been experimented and discussed for example in [12]. To overcome shortcomings of direct mapping between low-level sensors data to sound synthesis parameters, several authors proposed mapping strategies based on the computation of high-level descriptors, for both gesture and sound ([17], [4]). Recently, some authors proposed to describe gesture and sound from an emotion perspective ([3], [6]).

Nevertheless, most of the methods described in the literature assume that gestures and sounds are set independently from each other. Actually, such a possibility is generally seen as one of the advantage of working in the realm of digital media. We propose here a different standpoint: the gestures are set in relationship to the sound to be produced. Importantly, we wish to design interaction processes that rely on gesture units that could be directly perceived as relating to sound properties, instead of using a mapping between sensors data and sound parameters. Therefore, we are interested in exploring how a user can embody a sound and use such an information to build the sonic interaction. Particularly, we consider in this article physical gestures that can be considered as "mimicking" either the sound production or the sound characteristics.

We report here such an approach. First, we explicit two different interaction paradigms where gestures are defined in relation to a sound. This implies establishing methods to analyse gesture and sound, and their relationships in realtime. Second, we describe a case study corresponding to an implementation of our approach. Finally, we discuss our findings and their implications for sonic interaction.

2 Interaction paradigms

The proposed sonic interaction paradigm could be considered as an extension of sound sample playback allowing for the embodiment of the recorded sounds. Similarly to sampling techniques, we first define targeted sounds, with given time duration, and possibly

complex timbre characteristics. Nevertheless, instead of restraining the interaction to simple triggering, we consider complex gestures or actions, and complex sound transformations possibly combining different synthesis techniques such as the phase vocoder and concatenative synthesis. We describe in this section two different approaches to define and analyse gestures, in relationship to the sound characteristics.

2.1 Mimicking sound

Sound mimicry has recently be the objects of several studies ([9], [10], [13], [14]), generally considering actions imitating the sound produced by an acoustic instrument (for example "air-playing performance"). In our study, sound mimicry is extended to non-instrumental sounds, such as sounds that can be heard in everyday experiences.

Two main strategies could be drawn and could capture gestures intention in sound mimicry (see [7], [13]). Either the actions are related to the gesture producing the sound (*actions producing sound*) or they are directly linked sound properties (*actions describing sound*). Note that in this paper we refer both cases to sound "mimicking" (thus using this terminology in a possibly broader meaning than some authors).

Actions producing sound

We refer to *actions producing sound* in the case of sounds that can be clearly associated to an human action. For example, the sound of "reving a motorcycle engine" can be naturally linked to the action of rotating the grip. In such a case, the action does not describe the sound characteristics but the action leading to the sound production. Obviously some sounds can be associated to culturally well defined gestures, while others could remain ambiguous.

This brief description allows us to precise requirements for an interactive system working within this paradigm. The general scheme is illustrated by figure 1, describing a gesture recognition system. Generic gestures can be associated to each sound, and are first learned by the system. Then, the system can estimate the likelihoods of a given performed movement of being associated to the previously recorded gestures, thus to each sound. For example, Hidden Markov Model (HMM, see [15]) can provide a robust method to estimate likelihoods between two multidimensional signals.

Actions describing sound

We refer to *actions describing sound* in the case of sounds that exhibit particular morphology that could be "analogously" described by body movements. For example, the motion could follow the pitch contour of a sustained sound. Generally, the motion tends to describe spectro-temporal sound characteristics, quantified by various standard audio descriptors.

As shown in figure 2, the system associates the time profile of audio descriptors with the gesture profiles and quantifies similarities between both of them. For example, general multidimensional linear regression called Canonical Correlation Analysis (CCA, see [8]) can provide linear relationships between two sets of variables (considered as stationary stochastic processes) and constitutes an analysis of inherent many-to-many mappings.

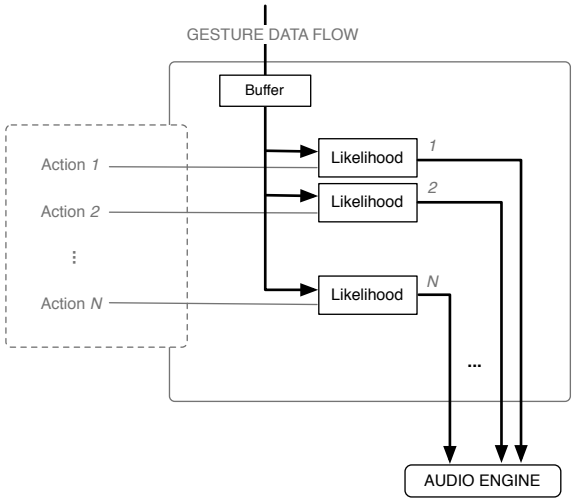

Fig. 1. Action recognition. Generic gestures (actions) are associated to each sound and are learned first by the system. Then likelihoods are estimated of an incoming gesture of being associated to an action.

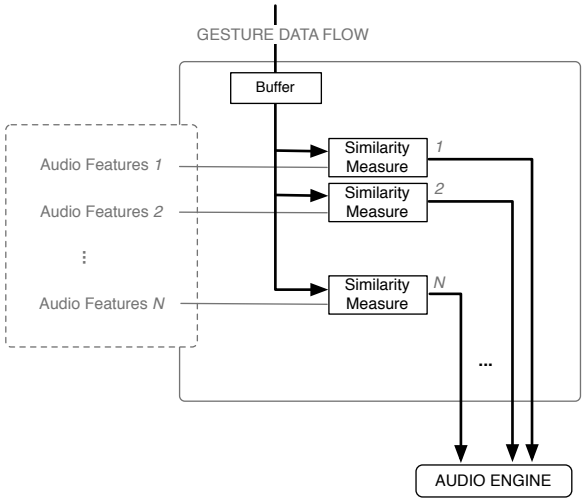

Fig. 2. Similarity measures. The system estimates the similarities between the time profile of audio descriptors and the incoming gesture profiles.

Mixed cases

The two main categories of actions mimicking sound we previously described can overlap. In several cases, actions mimicking sound production might reveal characteristics that could also be directly associated to sound descriptors. In the example of the "motorcycle engine", the sound intensity and timbre would naturally scale with the motion amplitude.

2.2 Sonic interaction

As illustrated in figure 3, the results of the gesture analysis feed an audio engine. Precisely, the different results, likelihood and similarity measures, obtained from the two types of analysis for "*actions producing sound*" and "*actions describing sound*" allow for selecting the corresponding audio files. For example, the similarity measures can be mapped to the volume of each sound, which allow the users to hear sound that matches closely their gestures. This creates an auditory feedback loop: the user is encouraged to pursue a gesture by listening to the sonic results.

Further sound transformations can also be processed, using parameters that are output from the gesture analysis. We report here three possible cases (since such processes can make the whole interaction more difficult to apprehend for first time users, this was not used in the case of the experimentation discussed in 3.3).

- *Time stretching/compression*. Using phase vocoder techniques, the sound can be time stretched or compressed according to the pace of performing the gesture (without altering the pitch). This parameter can be obtained from the gesture analysis jointly with the likelihood and/or similarity measures. Such a transformation can add an adaptive process to the feedback loop.
- *Filtering*. Additional filtering can be processed. For example, the global intensity of a gesture can be set to affect dynamically the spectral shape of the played sound.
- *Synthesis*. Using a high-level sound synthesis engine (e.g. [11]), new sounds can be synthesized with spectro-temporal characteristics approaching those from original sound.

3 Case study

In order to evaluate this approach, we develop a game-like application. The goal of the game is to play different sounds by mimicking them gesturally. We called this the "fishing game" since the control of a given sound ("catching the sound") is achieved if the appropriate gesture is performed. The scenario described here is general, and different versions have been already tested, for example with or without visual feedbacks.

3.1 Scenario

The scenario is composed of different levels the user must gradually succeed. At each level, typically two sounds or actions are presented to the users (either sonically or

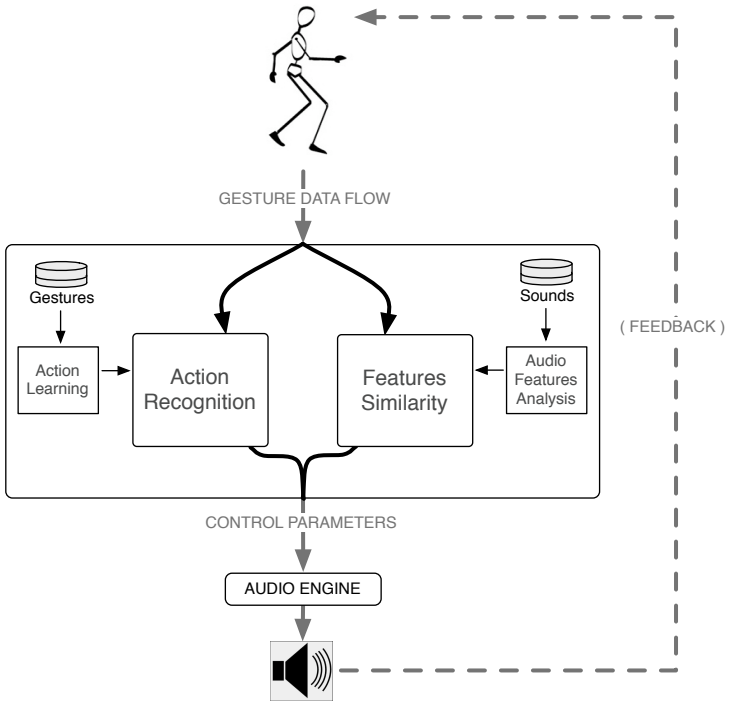

Fig. 3. The figure illustrates the general architecture of the interaction between user body movements and sound. A first part corresponds to the gesture analysis. A second part corresponds to the sound rendering using the analysis results.

with a written text). The user must mimic gesturally these sounds, one after the other. Each time the system evaluates which gesture is performed and the user will hear the corresponding sound (with various levels of transformation). If the user succeeds in a given time, the level increases.

Two specific versions of this scenario have been implemented and tested (discussed in section 3.3). In the first case, only actions producing sound were used.

In the second case, the sound corpus was chosen to contain sounds that can be easily associated to a particular hand gesture (category 1) and sounds that are not related to an action possibly achieved by a body motion (category 2). Each level includes one sound of each category (see table 1).

Category 1	**Category 2**
Reving a motorcycle	Repetitive low frequency sound
Sword fighting	Crow
Egg beating	Repetitive blast
Pouring a beer	Coin turning on a table
Page turning	Unvoiced vocal sounds

Table 1. Sound categorization. The sound corpus is reported here following the two categories. In the second version of the "fishing game" each level involves two distinct sounds: one from category 1; one from category 2.

3.2 Technical Implementation

The developed prototype application corresponds to an implementation of the architecture described in the figure 3. Precisely, the prototype was developed in the Max environnement taking advantage of the library FTM&Co ([16], [2]). Specifically, the action recognition is handled by a module called *gesture follower* [1] and the CCA algorithm was developed as a separate Max external ([5]).

Different types of sensors could be used as input. As a first step, we used accelerometer based interfaces. Both, the Wiimote game controller and a mobile phone were used.

3.3 Results and discussion

The first version of the prototype application was presented at the IRCAM -Agora festival 2009. Only *actions producing sound* were used with the action recognition module (*gesture follower*). During three days, more than fifty people played successively with this version, using a mobile phone as interface. A video feedback gave in real-time the score of their performance. Interestingly, it revealed that users trying the game several times could effectively adapt to the imposed set of gestures, and thus easily learn to

select sounds with their hand movements. Technically, the method was validated and led to the development of a revised version of the application.

The second version of the prototype included the *features similarity* block illustrated in figure 3. The experimental evaluation was performed on a sample composed by 13 subjects (experts and non-experts). During the experiment, the general instruction was to spontaneously mimic particular sounds. In this version the users had only one try to achieve the end of the game. Information was collected during the experiment and interviews were conducted to collect the user feedbacks.

Only one subject has perceived that each level was composed of two sounds from distinct categories, showing that in general the subjects did not have established consciously a mimicking strategy. Although, the interviews revealed that a minority of subjects (23%) have followed a mixed strategy and were able to finish the game. The remaining subjects (77%) adopted a single strategy related either to mimicking *actions producing sound* or *actions describing sound*, and performed worse than the others (only 40% succeeded).

For this version the Wiimote controller was used. Many subjects were not at ease with this interface (only 61,5% have appreciated it). While tangible interfaces can be expected to well suite the case of *actions producing sound*, the shape and handling of these interfaces can be nevertheless problematic. In this regard the specific handling of the Wiimote controller were not always well adapted to the chosen gestures. In the case of mimicking *actions describing sound*, the type of the interface becomes even more critical. As an illustration, subject 11 commented "How can I mimic a crow with a thing in my hand ?".

4 Conclusion

We presented a novel approach based on actions mimicking sound for gestural control in interactive systems. Two types of mimicking, *actions producing sound* and *actions describing sound* were identified. Case studies were conducted in the form of a game and allowed us to test the validity of this approach.

Globally, we found that users can quickly adapt to such system. Learning might be required for specific cases, but generally users can easily assimilate gestural strategies to achieve immediate and intuitive control.

The approach particularly favors auditory feedback occurring between sound and gesture. A sound corresponding to the mimicking gesture is produced, further encouraging the user to pursue and refine the performed gesture. The role of this feedback loop will be studied in detail.

Our case studies allowed us to define three main axes for further investigations. First, more complex sound transformations will be added enabling finer sound control. Second, special attention will be drawn to the matching between the tangible interfaces and the chosen sound. Finally, the concept of sound mimicry will be extended to more complex relationships between action and sound while pertaining our foundational focus on the correspondence between the perceived features of action and sound.

5 Acknowledgments

This work has been partially supported by the European Commission 7th Framework Programme SAME project (no. 215749) and by the ANR (Agence Nationale de la Recherche) project EarToy. We would like to thank Olivier Warusfel and Markus Noisterning who contributed through several discussions to the elaboration of the "Fishing Game".

References

1. F. Bevilacqua, F. Guedy, E. Fléty, N. Leroy, and N. Schnell. Wireless sensor interface and gesture-follower for music pedagogy. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, 2007.
2. Frédéric Bevilacqua, Rémy Muller, and Norbert Schnell. Mnm: a max/msp mapping toolbox. In *New Interfaces for Musical Expression*, pages 85–88, Vancouver, Mai 2005.
3. Antonio Camurri, Gualtiero Volpe, Giovanni De Poli, and Marc Leman. Communicating expressiveness and affect in multimodal interactive systems. *IEEE MultiMedia*, 12(1):43–53, 2005.
4. Camurri, A., Coletta, P., Mazzarino, B., Trocca, R., and Volpe, G. Improving the man-machine interface through the analysis of expressiveness in human movement. *Proc IEEE Intl.Workshop on Robot and Human Interactive Communication (ROMAN2002)*, pages 417–422, 2002.
5. Caramiaux, Baptiste and Schnell, Norbert. Towards an analysis tool for gesture-sound mapping. In *Gesture Workshop*, Bielefeld, Germany, 2009.
6. Ginevra Castellano, Roberto Bresin, Antonio Camurri, and Gualtiero Volpe. Expressive control of music and visual media by full-body movement. In *NIME '07: Proceedings of the 7th international conference on New interfaces for musical expression*, pages 390–391, New York, NY, USA, 2007. ACM.
7. William W. Gaver. How do we hear in the world? explorations in ecological acoustics. *Ecological Psychology*, 5(4):285–313, 1993.
8. R. Gittins. *Canonical analysis; a review with applications in ecology*. Springer-Verlag, Berlin, Germany, 1985.
9. Godøy, R.I., Haga, E., and Jensenius, A. R. Playing "air instruments": Mimicry of sound-producing gestures by novices and experts. In Sylvie Gibet, Nicolas Courty, and Jean-François Kamp, editors, *Gesture Workshop*, volume 3881 of *Lecture Notes in Computer Science*, pages 256–267. Springer, 2005.
10. Godøy, R.I., Haga, E., and Jensenius, A. R. Exploring music-related gestures by sound-tracing - a preliminary study. In *2nd ConGAS International Symposium on Gesture Interfaces for Multimedia Systems*, Leeds, UK, 2006.
11. Hoffman, M. and Cook, P.R. Feature-based synthesis: Mapping from acoustic and perceptual features to synthesis parameters. In *Proceedings of International Computer Music Conference (ICMC)*, New Orleans, LA, USA, November 2006.
12. Andy Hunt and Marcelo M. Wanderley. Mapping performer parameters to synthesis engines. *Organised Sound*, 7(2):97–108, 2002.
13. Alexander Refsum Jensenius. *Action-sound: Developing methods and tools to study music-related body movement*. PhD thesis, University of Oslo, Department of Musicology, Oslo, Norway, 2007.
14. Leman, Marc. *Embodied Music Cognition and Mediation Technology*. Massachusetts Institute of Technology Press, Cambridge, USA, 2008.

15. Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.
16. Norbert Schnell, Riccardo Borghesi, Diemo Schwarz, Frédéric Bevilacqua, and Remy Müller. Ftm — complex data structures for max. In *International Computer Music Conference (ICMC)*, Barcelona, Spain, Septembre 2005.
17. Marcelo Wanderley, guest editor. Mapping strategies in real-time computer music. *Organised Sound*, 7(02), 2002.