View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Goldsmiths Research Online

Mapping Through Listening

Baptiste Caramiaux

Department of Computing

Goldsmiths, University of London

United Kingdom

b.caramiaux@gold.ac.uk

Jules Françoise, Norbert Schnell, Frédéric Bevilacqua

UMR STMS IRCAM CNRS UPMC

Paris, France

{ jules.francoise, norbert.schnell, frederic.bevilacqua}@ircam.fr

Abstract

Gesture-*to*-sound mapping is generally defined as the association between gestural and sound parameters. This article describes an approach that brings forward the perception-action loop as a fundamental design principle for gesture–sound mapping in digital music instrument. Our approach considers the processes of listening as the foundation – and the first step – in the design of action-sound relationships. In this design process, the relationship between action and sound is derived from actions that can be perceived in the sound. Building on previous works on listening modes and gestural descriptions we proposed to distinguish between three mapping strategies: *instantaneous*, *temporal*, and *metaphoric*. Our approach makes use of machine learning techniques for building prototypes, from digital music instruments to interactive installations. Four different examples of scenarios and prototypes are described and discussed.

Introduction

In digital musical instrument, gestural inputs obtained from motion sensing systems, image analysis or sound analysis, are commonly used to control or to interact with sound processes or sound synthesis (Miranda and Wanderley (2006)). This has led artists, technologists, scientists to investigate so-called *mapping* strategies between gestural inputs and output sound processes.

Considered as an important vector of expression in computer music performance (Rovan et al. (1997)), the exploration of mapping approaches have led to flourishing research works dealing with taxonomy (Wanderley (2002)), the study of various strategies for example based on perceptual spaces (Arfib et al. (2002)), on mathematical formalization (Van Nort et al. (2004)), on dynamical systems (Momeni and Henry (2006)), and evaluation procedures based on user studies and other tools "borrowed from Human-Computer Interaction". (Hunt and Kirk (2000); Wanderley and Orio (2002)).

It has been often discussed that for digital music instruments, unlike most of acoustic instruments (Cadoz (1988); Wanderley and Depalle (2004)), there is no direct coupling between the gesture energy and the acoustic energy. Precisely, as the mapping is programmed in the digital realm, the relationship between the input and output digital data streams can be set arbitrarily. This offers thus unprecedented opportunities to create various types of mapping, which can be seen as part of the creative endeavour in building novel digital instruments.

After several years of experimentation we have developed an approach that brings back the perception-action loop as a fundamental design principle. Complementary to approaches that focus on building active haptic feedback to enhance the action-perception loop (Castagne et al. (2004)), we propose a methodology that is rooted in the concept of embodied music cognition. Precisely, the methodology considers listening as a process from which emerge gestures and interactions defining key elements for the design of mappings.

Our approach is anchored in advances in cognitive sciences and precisely rooted in Embodied Cognition (Varela et al. (1991)). The *enactive* point of view on perception and the idea of embodied cognition understand aspects of cognition as shaped by the body which comprises the perceptual and motor system (Varela et al. (1991); Noë (2005)). From this point of view, the action of listening – as our perception in general – is intrinsically linked to the process of acquiring knowledge and applying this knowledge when interacting with our environment (Merleau-Ponty (1945)). In music making, but also in speech and many other everyday activities, listening plays a particular role in the identification, evaluation, and execution of actions. The intrinsic relationship between action and listening in human cognition has been confirmed by many studies (Liberman and Mattingly (1985); Fadiga et al. (2002); Zatorre et al. (2007)). By extension, Embodied Music Cognition, developed by Leman (Leman (2007)) and Godøy (Godøy (2006)), tends to see music perception as based on actions. Many situations involve people moving while listening to music. In the embodied music cognition framework, these movements are seen as conveying information on the perceived sonic moving forms (Leman et al. (2009)).

While embodied music cognition provides us with a theoretical framework for the study of listening in a musical context and for the study of the link between music perception and human actions, digital music performance requires computational tools to implement experimental breakthroughs. Recent tools coming from the machine learning research field allow for building scenarios and prototypes implementing concepts borrowed from embodied music cognition. Such scenarios are indeed usually best defined from high-level gesture and acoustic descriptions, which cannot generally be easily programmed with other techniques. For examples, the use of machine learning techniques allows for setting the gesture-sound relationships from examples or database.

In this paper we propose a new approach of gesture-*to*-sound mapping that relies on the concept of embodied sound cognition and we report applications that make use of machine learning techniques to implement these scenarios.

The article is structured as follows. In the second section, we review previous works characterizing different listening modes, and reporting on gestural descriptions of sounds. The third section describes our approach for the design of mappings inspired by the different modes of listening. The proposed mappings are presented as real-world applications and stem from our past and current research in this field. In the last section, we discuss the different scenario and mapping strategies.

Describing Sound Gesturally

As mentioned previously, we are interested in examining mapping strategies through the theory of Embodied Music Cognition. In particular we focus on listening processes that might induce gestural representations in order to conceptually invert the process, from gesture to sound, to create the mapping. First, we review in this section works describing different listening modes, that can be related to specific sound properties. Second, we show that these listening modes can be related to different action strategies.

Listening Modes

Sound, as considered here, refers to recorded audio material on a given support. Recorded sound can be played back and processed through various techniques, which, importantly, leads to different listening experiences. A vast body of work is devoted to the mechanisms of listening gathering various research fields such as psychoacoustics, neurosciences, auditory scene analysis, musicology. In this section, we focus on conceptual approaches of listening which originated principally from music theory and ecological perception theory. Our goal is to create a comprehensive overview of listening modes and their functions which will eventually be linked, in the next section, to gestural representations.

First, in the context of *musique concrète*, Schaeffer (Schaeffer (1966)) defined four functions of listening[1]: 1) *Listening* (*écouter*) that focuses on the indexical value of the sound (i.e. the sound source); 2) *Perceiving* (*ouïr*) that is the most primitive mode consisting of receiving the sound through the auditory system; 3) *Hearing* (*entendre*) that refers to the selective process between auditory signals, the attention to inherent characteristics of the perceived sound; and 4) *Comprehending* (*comprendre*) that brings semantics in sounds, treated them as signs. These different functions of listening are non mutually exclusive and operate competitively.

Based on Schaeffer's theoretical taxonomy, and motivated by new concepts from auditory display, Chion (Chion (1983)) proposed a taxonomy comprising three categories, called *modes of listening*: 1) *Causal* listening, that consists in listening to a sound in order to gather information about its cause (or source); 2) *Semantic* listening, that refers to a code or a language to interpret a message; and 3) *Reduced* listening, that focuses on the qualities of the sound itself, independent of its cause and of its meaning[2]. Hence Chion does not consider the low-level aspect of perception called perceiving (*ouïr*).

Modes of listening have also been of interest in the ecological approach to auditory perception. One important application has been the design of sounds in human-computer interaction. In this context, Gaver considered environmental sounds

[1] Note that translating Schaeffer's listening mechanisms is far from trivial. Consequently we chose to report in this article both the translation and the term in its original language.

[2] Note that the reduced listening is a concept that has first been introduced by Schaeffer to motivate the concept of *Sound Object* in *Musique Concrète*.

and proposed to differentiate between two types of listening (Gaver (1993b,a)) defined as: *everyday* listening, in which the perception focuses on events rather than sounds; and *musical* listening, in which perception is centered on the sound characteristics. As mentioned by Gaver, musical listening of environmental sounds can be achieved by listening "to the world as we do music" (Gaver (1993b), p1). Gaver took the example of the work by the american composer John Cage that aims at hearing everyday world as music in some of his compositions.

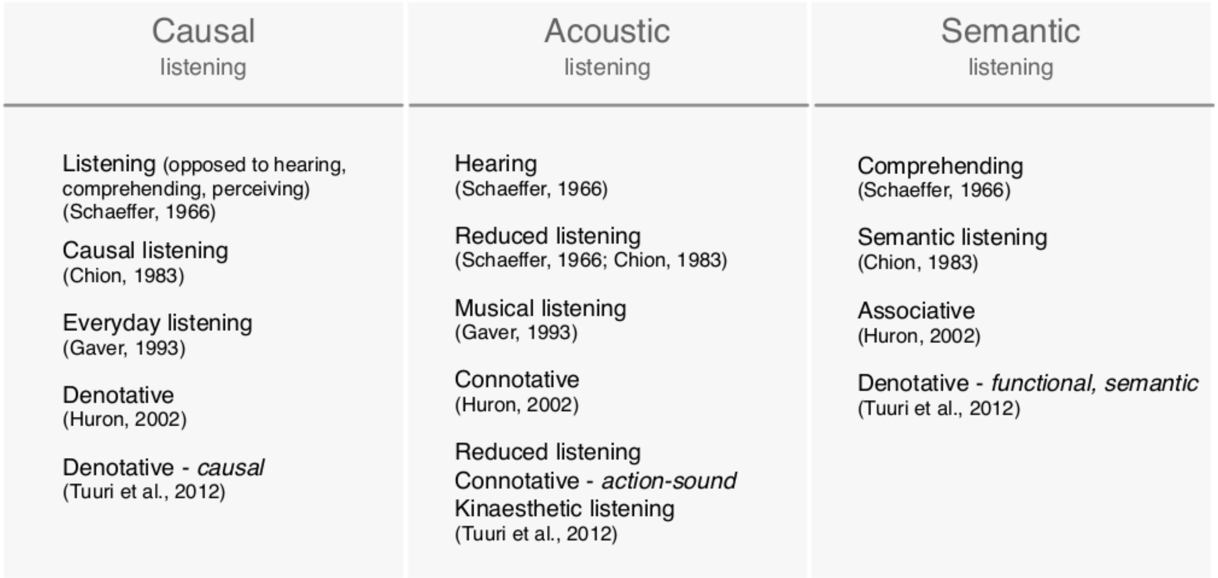
Recent studies proposed to enrich these previous taxonomies by adding an emotional dimension evoked by the auditory stimulus. Huron proposed an analytic framework supporting the idea that emotional experiences may be usefully characterized according to a six-part classification (Huron (2002)): 1) *Reflexive* that refers to fast automatic physiological responses; 2) *Denotative* that allows the listener to identify sound sources; 3) *Connotative* that allows the listener to infer various physical properties about sound sources such as size, proximity, energy, material, and mode of excitation; 4) *Associative* that refers to arbitrary learned associations; 5) *Empathetic* that refers to identify if the sound is produced by an animate agent and if it is a human it refers to his or her "state of mind"; and, finally, 6) *Critical* that refers to conscious cognitive processes by which the intentions of a sound-producing agent are evaluated.

Recently, Tuuri et al. (Tuuri and Eerola (2012)) proposed an extended taxonomy of listening modes. The taxonomy is hierarchical with three levels: *experiential*, *denotative*, *reflective*. The *experiential* level encompasses Huron's reflexive and connotative modes. The connotative mode more precisely focuses on the relation between the action and the external world (object, people, cultural context). In Tuuri et al.'s taxonomy the experiential mode also induces a *kinaesthetic* mode that refers to the inherent movement qualities in the sound (for example characterizing a sound as "wavy"). The second level in the hierarchy is the *denotative* mode. This mode has been first defined by Huron and extended by Tuuri in order to separate between modes focusing on sound sources and sound contexts. Finally, the top level is the *reflective* mode that encompassed the reduced mode as well as the critical mode.

The important point here is to realize that several of the listening modes make reference explicitly or implicitly to motor imagery or action. Both the *Causal* listening mode of Chion and the *Denotative* listening mode of Huron/Tuuri refer to associate the sound to the action that created the sound. Such actions are generally linked to clear interactions and motion between objects (e.g. a stick hitting a cymbal). We will keep the term *causal* listening throughout this article to denote such an association between the action and the sound.

The *Reduced* listening mode of Schaeffer and Chion, the *Connotative* mode of Huron, and *Kinaesthetic* mode of Tuuri refer to acoustic properties of the sound. We will use the term *Acoustic* listening throughout this article for such a type listening. These acoustic aspects could be quantified using a set of sound descriptors from the sound signal. However, a crucial point is to acknowledge that defining the *reduced* listening mode was also linked to sound description such as the Schaeffer's Typo-Morphology of Sonic Objects (Schaeffer (1966)), or later to Temporal Semiotic Units (Unité Sémiotiques Temporelles) (Frey et al. (2009)). As elucidated by Godøy (Godøy (2006)), these descriptions can in many cases be linked to notions of motions and actions.

The last mode of listening encompasses semantic aspects of sound perception and will be called as such. Figure 1 summarizes the three modes of listening *Causal*, *Acoustic*, *Semantic* that will be considered in this paper and that we aim at associating to gestural representations.



Causal listening	Acoustic listening	Semantic listening
Listening (opposed to hearing, comprehending, perceiving) (Schaeffer, 1966)	Hearing (Schaeffer, 1966)	Comprehending (Schaeffer, 1966)
Causal listening (Chion, 1983)	Reduced listening (Schaeffer, 1966; Chion, 1983)	Semantic listening (Chion, 1983)
Everyday listening (Gaver, 1993)	Musical listening (Gaver, 1993)	Associative (Huron, 2002)
Denotative (Huron, 2002)	Connotative (Huron, 2002)	Denotative - *functional, semantic* (Tuuri et al., 2012)
Denotative - *causal* (Tuuri et al., 2012)	Reduced listening Connotative - *action-sound* Kinaesthetic listening (Tuuri et al., 2012)	

Figure 1. Simplified taxonomy of listening modes. Causal listening refers to an explicit association between sound and its producing action. Acoustic listening is related to perceptual acoustic properties of the sound. Semantic listening integrates higher level notions of meaning and interpretation.

Linking Gesture and Listening

In the previous section we reviewed the listening modes as introduced by various authors in the literature that we summarized as a three-mode approach accounting for Causal, Acoustic and Semantic listening. In this section, we posit that these modes of listening can be linked to specific gestural strategies. We base this statement on a review of important works within the field of behavioral approaches in embodied music cognition, that reported on gestural sound description.

Interactions between sound perception and motion have been studied either through a neuroscientific perspective or a behavioral perspective (Zatorre et al. (2007)). Generally, the motor-auditory interaction has been recognized as important to describe sound perception. Neuroscience studies have shown how listeners activate cognitive action representations while listening to music performances whether they are expert musicians or novices (Haueisen and Knösche (2001); Lahav et al. (2005); Zatorre et al.

(2007)).

In a behavioral approach, a common experimental methodology consists in asking participants to perform movements along with the music, while it is played. The movement analysis can reveal important insights into the underlying embodied cognitive processes related to music perception. A wide range of works concern controlled tasking, for instance tapping task on beats (Large (2000); Large and Palmer (2002)).

In systematic musicology, exploratory procedure is more commonly used such as asking participants to spontaneously gesticulate while listening to a sound stimulus or music. For instance, Leman (Leman et al. (2009)) studied participants' movements made with a joystick, along a guqin music performance that they heard. Also, Küssner (Küssner (2013)) considered free tracing movements on a tablet along two Chopin's preludes. Other works concern specifically designed stimuli with well characterized musical parameters. Consequently, it is possible to investigate how the chosen musical parameters affect the resulting movements.

Godøy is one of the pioneer of this type of research and proposed to use the morphology of the sound stimulus based on Schaeffer's typology (impulsive, iterative, sustained) (Godøy et al. (2006)). This methodology was then used by other authors such as Merer (Merer (2011)) and Nymoen (Nymoen et al. (2011)). Recently Küssner (Küssner (2013)) proposed the use of sequences of pure tones while changing the following parameters: pitch, loudness and tempo.

These previous works provide us with promising methodology for the study of gestural description of sounds. Most of these studies rely on exploring *analog* relationships between gestural and sound parameters. We will refer to such an approach as *tracing/analog* experiments where the motion trajectories are associated to acoustic parameter. In addition, in the following we will refer to sound morphology to designate the temporal profile of its acoustic characteristics, e.g. amplitude, pitch, timbre aspects.

In our prior work, we conducted experiments to give evidence on the link between gestural description and both acoustic and causal listening modes. We examined experimentally how participants can associate different types of motion in the *acoustic* and *causal* listening modes. We observed two related strategies Caramiaux et al. (2014): *mimicking* the action related to the sound source (*causal* listening mode) or *tracing* the shape of the sound parameter (*acoustic* listening mode). In particular, we showed that the sound source identification, thus the mode of listening, has a direct consequence on the gesture strategies. If the participants can identify the sound source as an action, they tend to mimic the action. On the contrary, a non-identifiable sound leads participants to trace the profile of perceived sound features.

This experimental study showed a link between causal/acoustic listening modes and analog/mimicking motion strategies. This study brings grounds for establishing mapping strategies based on listening modes and associated motion strategies. Mapping strategies can stem from the reviewed experimental findings as well as they can evoke particular links between listening modes and motion through a scenario and design of interaction. In the next section we describe specific examples illustrating the link between causal, acoustic and metaphoric listening modes and gestural strategies.

From Listening to Controlling

In this section we describe concrete examples that we developed and which were used in different settings, from experiments, demos, interactive installations and performances. All these examples are based on modeling the target sound from a gestural perspective: a prior listening (or evocation) to the sound provides performers with insights on possible gesture control strategies. These strategies are then made possible using machine learning techniques. Similar approaches have been described by Godøy (Godøy (2006)), Van Nort (Van Nort (2009)), Fiebrink (Fiebrink (2011)) and Maes (Maes (2012)).

Our general methodology is as follows. The first step corresponds to listening to recorded sounds from different perceptual perspectives as described in the previous section. This leads to consider scenarios and metaphors where the motion in interaction is linked to the targeted sounds. Mapping strategies are then designed to implement the interaction scenarios. In most cases, the mapping is built using machine learning techniques from examples gathered during a "learning" phase, before the final "playing" phase.

Interaction Scenarios and Mapping Strategies

Shaking

The action-sound mapping of this scenario emerges from the action metaphor of *shaking* associating the performer's shaking movement to the generation of percussive sounds. This scenario is meant to be related to the causal mode of listening since the performer mimics the gesture of shaking. While in music performance this metaphor may refer to percussion instruments such a shaker or maracas, it can also be associated to various non-musical actions and sounds. Consequently, the mapping designed for this scenario can be applied to any sound that is composed of percussive events of varying intensity and to any movement that resembles shaking or waving (i.e. movements that are periodic and modulated in intensity).

The mapping is designed to be a direct relationship between the movement energy and the energy of the sound played. However the sound can be chosen to be any percussive recorded sound. The mapping relies on a first phase called *learning*. During this phase, an offline analysis of a sound database segments the recorded materials into percussive events and describes each segment by its perceived intensity. Each segment is consequently structured according to its intensity level. During the second phase, *playing*, the performer's motion is analyzed in real-time by computing its energy. Sounds are then selected from the database according to the motion's level of energy. The intensity of shaking has a direct relationship to the intensity of the synthesized percussive sound event whereas the performer does not control the rhythmic pattern. Figure 2 illustrates the scenario.

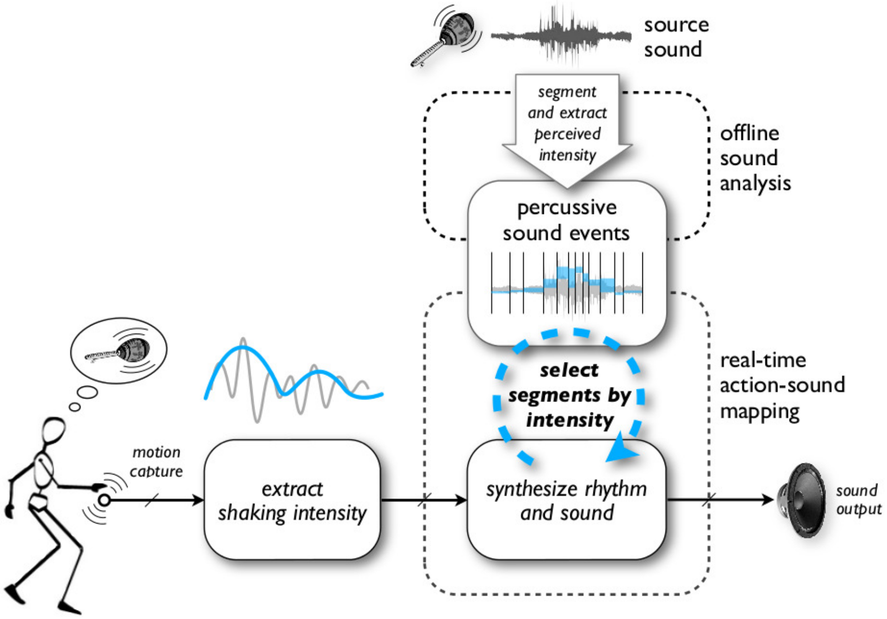

source sound
segment and extract perceived intensity
offline sound analysis
percussive sound events
select segments by intensity
real-time action-sound mapping
motion capture
extract shaking intensity
synthesize rhythm and sound
sound output

Figure 2. Shaking scenario. A recorded rhythmic sound is analyzed and segmented. An incoming gesture is analyzed and its energy is computed and drives the selection of the segment to be played.

Technically, we use accelerometers to sense the performer's motion. Concrete implementations were featured in different performances using the MO (musical object) interfaces [3](Rasamimanana et al. (2011)). The shaking intensity can be obtained by integrating the variations of the measured acceleration magnitude. Audio segmentation is performed by onset detection. A mean loudness measure is computed for ear segment. Both feature spaces, motion and sound, are normalized so that each sound

[3] Performances exhibited at the 2011 Guthman Competition of New Musical Instrument, or the performance of the TEI'13 conference

segment can be associated to a corresponding shaking intensity between the lowest and highest possible intensity. The system used a k-nearest-neighbor (kNN) search algorithm based on a k-D tree to select a sound event of a given intensity among the available segments (Schwarz et al. (2009)).

Shaping

Shaping refers to scenarios where performers control sound morphologies by "tracing" in the air salient sound features they desire to control. It is thus related to *acoustic* listening as we defined previously, where the performer pays attention to perceptual acoustic aspects of the sound, and in particular to its temporal evolution.

The interaction scenario leads the performer to design gestures related to specific recorded sound morphologies. Rather than using a metaphor, the link between gestures and sounds is built by analogy, as the design of gestures needs to reflect tightly the aspects of the sound the performer perceives and intends to affect. The mapping relies on two distinct phases: *learning* and *playing*. The learning phase consists in a prior construction and analysis of a database of sounds. Each sound is analyzed offline to compute the feature representation. The playing phase starts with a gesture executed by a performer. The performer draws gesturally the morphology of a particular sound, and re-plays the sound in real-time, translating the time variations of the input gestures to sound variations. The beginning and the end of the gesture must be marked by the performer (e.g. using buttons on the interface). A sound is selected as soon as the gesture starts thanks to a realtime shape matching algorithm that finds, at each time step, the closest audio feature morphology to the gesture morphology and temporally aligns them. Figure 3 illustrates the scenario.

The implementation is based on machine learning technique is based on Hidden Markov Model (HMM), called *gesture follower*, and presented in Appendix. Since the

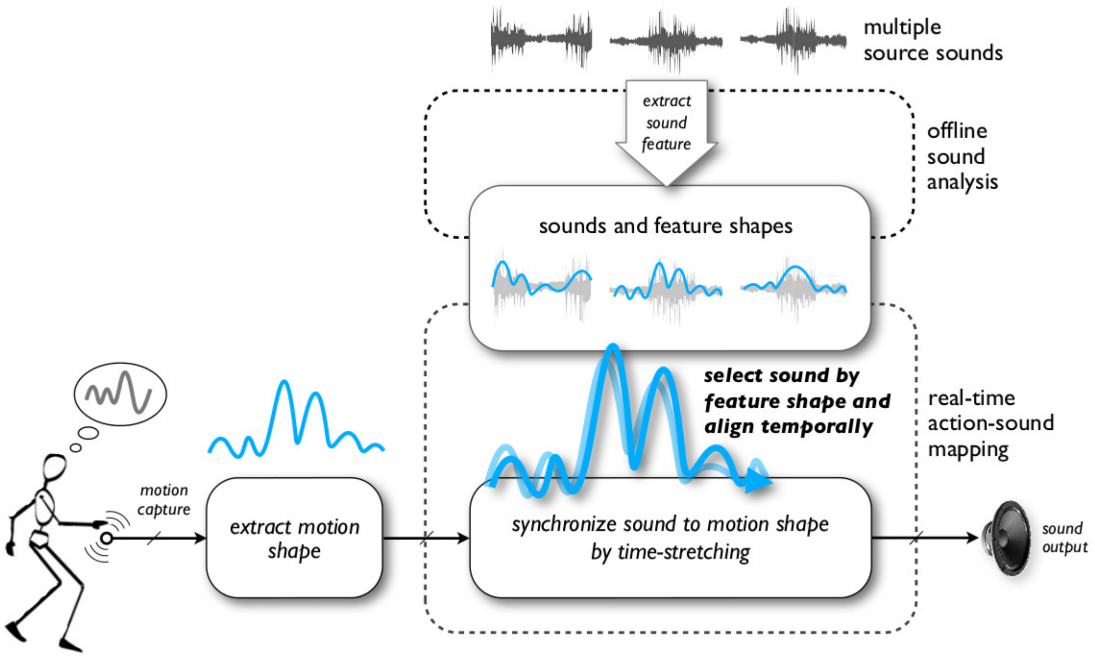

multiple source sounds
extract sound feature
offline sound analysis
sounds and feature shapes
select sound by feature shape and align temporally
real-time action-sound mapping
motion capture
extract motion shape
synchronize sound to motion shape by time-stretching
sound output

Figure 3. *Shaping scenario. Multiple sounds are analyzed by computing feature shapes. On the other hand, the motion shape of a live gesture performance is extracted and used to select and control the sound whose feature shape is the closest to the gesture.*

sound is aligned to the gesture in real-time, it translates the variations in the gesture morphology, such as the speed of execution, to variation in the playback, re-interpreting the recorded sound. In the demonstration presented at the Sound and Music Computing Conference in 2010 (Caramiaux et al. (2010a)), gesture and sound were represented by unidimensional time series, the energy of a gesture controlling the loudness. The energy of a gesture was computed as its absolute speed (an infra-red camera motion capture system was used to capture the gesture). Being of different physical dimensions, the time series were scaled beforehand into the same range of values.

Fishing

The fishing scenario relies on a metaphor where performer mimics an action in order to select and play a specific sound. In other words, the performer virtually "fishes" the sound by mimicking the associated action that supposedly caused the sound. Therefore, the fishing scenario is meant to be related to the causal aspect of listening where a performer focuses on the event that has produced the sound and tries to mimic it.

The application is based on the recognition of the performed action and requires a training phase: a database of actions is built by recording one example of each action to be recognized. An action is a single unit represented as a multidimensional continuous time series of its parameters. In addition, each action has an associated sound meant to illustrate the possible sound produced by the action. During the *playing* phase, the user performs a gesture that, if recognized as an action from the database, will trigger the playback associated sound. Since the system relies on action recognition, both the performed and the pre-defined actions must have a consistent representation, which could imply to have been performed with the same device and consequently with the same set of parameters taking their values into the same range. Figure 4 illustrates the scenario.

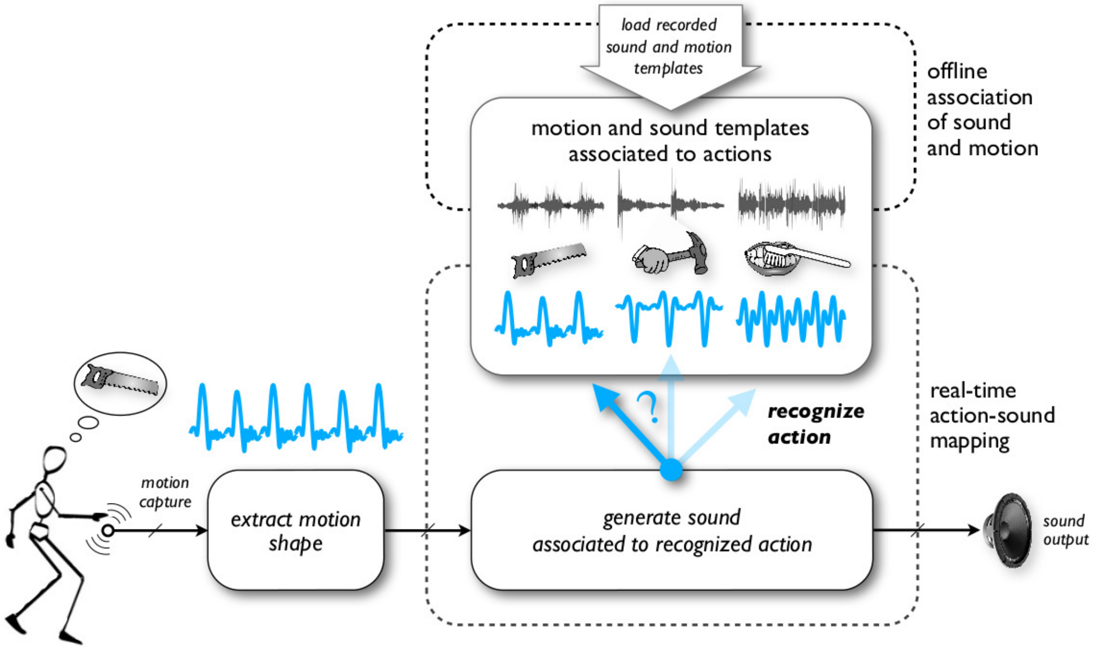

load recorded sound and motion templates
motion and sound templates associated to actions
offline association of sound and motion
?
recognize action
real-time action-sound mapping
motion capture
extract motion shape
generate sound associated to recognized action
sound output

Figure 4. Fishing scenario. A set of recorded sound is loaded together with associated actions that represent the sound. The incoming live gesture performance tries to "fish" a sound by mimicking the associated action. If successful, the sound is played.

The system uses a HMM-based gesture recognition method called *gesture follower* presented in Appendix and also used in the shaping scenario. In the installation version of the system, presented during the SAME project meeting[4], the actions were captured through the use of mobile phones with embedded accelerometers. The training process is part of the design and not seen by the performer. The playing phase was implemented with a gaming scenario. A set of two action–sound from the database was presented to the user in order to be mimicked. The algorithm was set to play the sound associated to an action as soon as this action is recognized. In addition, the algorithm was set to output the time progression in the executed action. If the user reaches 90% of the recognized action, this sound is set to be fished. The user has to do the same with the second action. Once both sounds are successfully fished, another set of pairs is presented.

Shuffling

The *shuffling* scenario consists in re-composing and re-interpreting complex sound sequences gesturally. This is achieved by processing short pieces of recorded sounds put in relationships with gesture segments. The scenario does not involve pre-established metaphors as in the previous examples, but defers the design choices to the performers, allowing them to implement interactively their own metaphors and control strategies.

The mapping is designed by demonstration: the gestures performed by the performer in conjunction with particular sounds are used to train a machine learning model encoding their relationships. When performing a new gesture sequence, sounds are resynthesized and aligned in real-time using phase vocoding. In some aspects, the present scenario generalizes some of the previous examples by allowing the performer to mimic sound-producing actions (cf. *fishing*), to trace sound features (cf. *shaping*), or to combine these approaches sequentially.

Designing the mapping by demonstration involves an interaction loop divided into

[4] Sound And Music For Everyone Everyday Everywhere Everyway http://www.sameproject.eu/

two distinct phases: *learning* and *playing*. During the *learning* phase, the performer begins by selecting sounds and defining their segmentation manually using a graphical editor. Then the performer records one or multiple gestures associated to each sound, for example by recording a template gesture synchronously while listening to a given sound. Additionally, one can specify authorized transitions between each gesture and sound segment. During the *playing* phase, the performer recomposes the original sounds by performing arbitrary sequences of gestural segments. The gestures are recognized and aligned to their reference in real-time to select and replay dynamically the appropriate sequence of sound segments along with the gesture performance. Figure 5 illustrates the shuffling scenario.

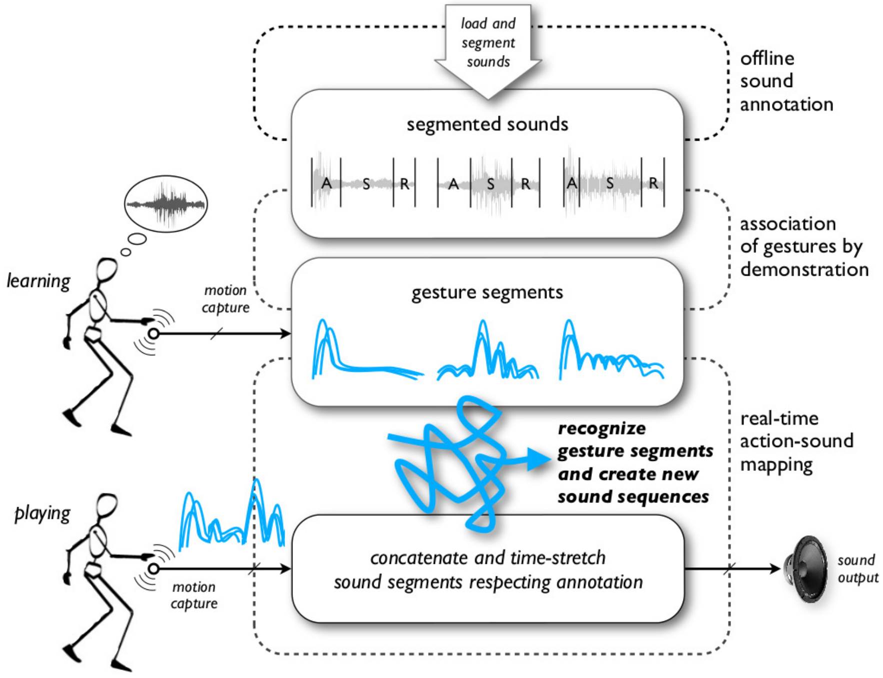

load and segment sounds
offline sound annotation
segmented sounds
A S R A S R A S R
association of gestures by demonstration
learning
motion capture
gesture segments
recognize gesture segments and create new sound sequences
real-time action-sound mapping
playing
concatenate and time-stretch sound segments respecting annotation
motion capture
sound output

Figure 5. Shuffling scenario. A learning phase allows the performer to select a segmented sound and to record one gesture associated to it, for example by recording while listening to a given sound. A playing phase allows the performer to recompose the original sound by performing arbitrary sequences of gestural segments.

Technically, the mapping is based on a hierarchical model for continuous gesture recognition and segmentation, called Hierarchical HMM (see Appendix for details). The model has two levels. The lower level precisely encodes the time structure of the segments, while the higher level governs their sequencing, defining the possible transitions between various points within the gesture. The model can be built from a single segmented example. The recognition is performed in real-time and the model estimates the alignment of the new gesture compared with the reference, allowing for the reinterpretation of the sound with a fine time precision. Thus, the temporal variations of the live gestural performance are translated to sound variations using a phase vocoder (superVP in Max/MSP).

A specific implementation was introduced in (Françoise et al. (2012)). Each gesture and sound morphology is segmented in attack, sustain and release segments, possibly complemented with a preparation phase anticipating the attack of the sound. This decomposition is particularly interesting in two aspects. First, the consistency of the relationships between gesture and sound can be guaranteed by specifying constraints for the sound synthesis on particular segments, e.g. silence during preparation or transient conservation on attack phases. In addition, the features extracted from the performer's gesture in one action segment can be mapped to sound features of the following segments. In this way, the silent trajectory of a preparation gesture can define the features at the beginning of the sound that, in the following, can be shaped by the performer's gesture.[5] Second, it allows for designing strategies that involve various gestural descriptions related to listening: preparation and attack can be related to mimicking, e.g. using a metaphor such as hitting an object, while the sustain and release phases can implement a tracing gestural description.

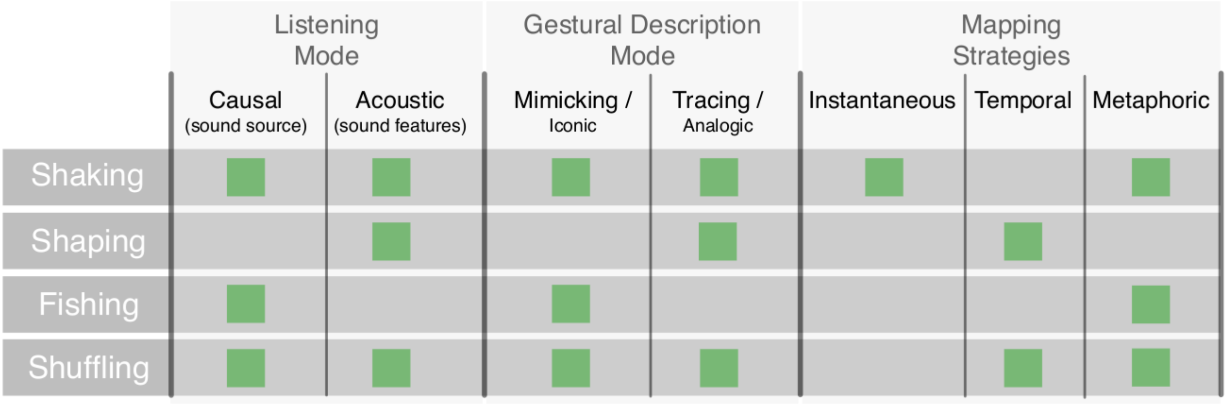
[5] In the design of traditional instruments, similar possibilities are obtained through the instrument's geometry allowing the performer to interact – or not – with different parts of the instrument responding to action in different ways.

Discussion and Conclusion

We presented four mapping examples illustrating our approach based on a perceptual analysis of the target sound. All examples use synthesis techniques to “re-interpret” gesturally the recorded sounds. Each scenario and mapping strategy can be described by a top-down approach. In particular, each can be linked to particular listening modes and gesture strategies presented in Section “Describing Sound Gesturally”.

The Figure 6 summarizes how the examples are related to the different listening modes and gestural strategies discussed in “Describing Sound Gesturally”. In addition, we insist on the different strategies of mapping that are used in the different examples. We distinguish between *instantaneous*, *temporal*, and *metaphoric* aspects that define the relationship between gesture and sound. Instantaneous mapping strategies refer to the translation of magnitudes between instantaneous gesture and sound features or parameters. Temporal mapping strategies refer to the translation and adaptation of temporal morphologies (i.e. profiles, timing, and event sequences) between the gesture and sound data streams. Metaphorical mapping strategies refer to relationships determined by metaphorical or even semantic aspects, that do not necessarily rely on morphological congruences between gesture and sound.

The *shaking* scenario makes principally use of an instantaneous mapping strategy between gesture and sound: the shaking intensity is directly related to the intensity of each percussive sound event. Interestingly, we have observed how performers spontaneously synchronize their shaking movements to the tempo generated by the system. This creates a direct action-perception loop: the sound “feedback” produced is similar to a shaker sound and encourages the player to pursue a shaking movement. The listening mode is causal and there is an metaphorical association between the action and sound. Due to the strong action metaphor the scenario can also use completely



	Listening Mode		Gestural Description Mode		Mapping Strategies		
	Causal (sound source)	Acoustic (sound features)	Mimicking / Iconic	Tracing / Analogic	Instantaneous	Temporal	Metaphoric
Shaking	🟩	🟩	🟩	🟩	🟩		🟩
Shaping		🟩		🟩		🟩	
Fishing	🟩		🟩				🟩
Shuffling	🟩	🟩	🟩	🟩		🟩	🟩

Figure 6. Classification of the scenarios along three dimensions: the listening mode, related to listening processes, the gestural strategy which describes how gestures derive from listening, and the mapping strategies implementing each gestural strategy.

unconventional sounds the performer can *shake*.

In the *shaping* scenario the performers mainly focus on “acoustic” properties of the sound. They must “trace” the temporal profile of a sound feature to be able to select and modify a sound whose morphology matches the motion shape. Relying on temporal morphologies, the mapping of this scenario can be seen as the closest mapping example to previous ideas developed by Godøy (Godøy (2006)) or Van Nort (Van Nort (2009)). The difference with the shaking resides in the precise control of the sound’s temporal evolution, supporting a listening mode focussing on acoustic sound features. Our experiments with this scenario showed us that sonorous profile must be memorized beforehand in order to consciously target one and, eventually, reproduce it with temporal variations.

The *shaping* scenario makes use of a temporal mapping between gesture parameter and sound feature. This mapping allows the performer to re-shape a sound based on the temporal morphology of his or her gesture. The general concept of temporal mapping was previously introduced in (Bevilacqua et al. (2011)) for the cases where temporal relationships between gesture and sound parameter profiles are established.

The *fishing* scenario makes use of a mapping that can be considered as metaphoric: unlike in the *shaking* and *shaping* scenarios, the morphologies of gesture and sound in this example can be incongruent in some cases. Nevertheless, the action-sound relationship is clear from a *causal* listening perspective. As mentioned previously, this scenario has been shown as an installation during the EU Project SAME. Feedbacks from users showed that such a mapping was highly appreciated and characterized as ludic. Indeed, the sounds chosen were easily identified and the action easily reproducible. Although the scenario focuses on a causal mode of listening, an extended version comprising a metaphoric mode of listening can be envisaged and can enrich the scenario.

Finally, the *shuffling* scenario makes use of a mapping strategy that can be characterized as both temporal and metaphoric. The temporal characteristic of the mapping is similar to the shaping scenario while the metaphoric characteristic is enabled by the implementation of a general algorithm for the recognition of action and action sequences. The combined mapping consequently offers additional control opportunities and action-perception loop feedback. It drives performers in both causal and acoustic listening modes, making them conscious of both the sound morphologies (like in shaping) and the control of sound segment through iconic gesture segments (like in fishing). The shuffling example can be seen as an unified approach in the sense that it can be configured in order to activate several modes of listening and several modes of gestural description (and can also easily include the shaking scenario).

The temporal aspects of mapping are particularly important when designing action-sound relationships based on the transformation of recorded sounds. In this case, temporal mapping strategies allow for adapting the temporal morphologies initially present in the recorded sounds to the actions of the performer. Nevertheless, we believe that temporal mapping strategies are equally powerful considering other synthesis methods. They allow for segmenting the performers actions and for defining different action-sound relationships for different segments. A need is for example the distinction between action segments that actually produce sound or induce sound changes and those that do not.

One design choice in the presented examples concerns the motion sensing technology. Any sensing systems provide a partial gesture description, which might impact on the sound controllability. In the four scenarios presented, we used accelerometers. Although these sensors have inherent limitations (e.g. unable to sense spatial information), they are sensitive to small changes in orientations and dynamics. Moreover, the choice has been motivated by other advantages of this technology: low-cost, wireless, good understanding of the signals, sufficient precision for most musical applications.

The scenarios discussed in this article make extensive use of machine learning based methods (k-NN, HMM, Hierarchical HMM). The role of machine learning is to realize the top-down approach of our scenarios based on perceptual or metaphoric action and sound description. All scenarios indeed imply implicit relationships between sound and gesture features. As discussed by Mitchell (Mitchell (2006)), machine learning techniques are efficient to model such implicit relationships. Moreover such approach starts to be implemented and evaluated in different cases in computer music performance (Fiebrink (2011); Gillian (2011); Caramiaux and Tanaka (2013)). The ongoing research in this area examines the use of machine learning for automatically selecting gesture and sound features (Caramiaux et al. (2010b)), for modeling jointly their interactions over time in order to capture implicitly their correlations and the expressive variations emerging in different interpretations (Françoise et al. (2013)) or the use of machine learning as a design tool Fiebrink et al. (2011).

The possibilities arising from the introduction of machine learning techniques into the interaction loop are twofold. First of all, they allow for the instrument integrating notions of recognition and prediction that support the implementation of interactions based on the performer’s listening. While the performer always adapts – spontaneously or by strenuous learning – his or her actions to the behaviour of the instrument, these new instruments for their part adapt themselves to the performer’s behavior, preferences, and playing style. It is worth noticing that ML techniques are prone to errors or may require time to converge to the accurate estimation. Latency involved is inherent. It might be an issue for specific types of control. On the other hand, latency can be handled by design. For instance, in the fishing scenario we chose to use the recognition latency, namely the fact that the user has executed 90% of the action, as a visual progress bar for the user. Interestingly, the latency represented as such, challenged the user during the interaction, enhancing the game play.

In conclusion we proposed a design approach for mapping based on the concept of embodied listening. Building on previous works on listening modes and gestural descriptions we proposed to distinguish three mapping strategies, *instantaneous*, *temporal*, and *metaphoric*. Our approach considers the processes of listening as the foundation – and the first step – in the design of action-sound relationships. In this design process, the relationship between action and sound is derived from actions that can be perceived in the sound. We believe that the described examples are only scratching the surface of the possibilities arising from this approach.

Algorithms

Gesture Follower

Gesture Follower (GF) (Bevilacqua et al. (2010)) is a template-based gesture recognition method based on Hidden Markov Models (HMM). The model is learned from a single gesture example using its whole time series as a template: the model is built by assigning a state to each frame, similarly to Dynamic Time Warping (DTW). The time structure is modeled by a left-right transition structure. A causal forward inference allows for a real-time decoding and returns the currently recognized template as well as the time progression in the template, performing an alignment of the live gesture to the template.

Adaptive Extension

The model has been recently extended to quantify and adapt to gesture variations by using Sequential Monte Carlo inference on the parameters of a non-linear dynamical system. It allows for the continuous adaptation to variations of gesture characteristics (Caramiaux (2012)). Indeed, once the gesture template recorded, a similar live gesture can be performed with variations in speed, scale, rotation, etc. These characteristics can be taken into account explicitly by the method as invariant for the recognition. To that extent, the method continuously estimates the relative characteristics of the gesture variations, which can then be used in continuous interaction scenarios.

Hierarchical Extension

Gesture Follower has been extended to comprehend more complex time structures, allowing for the representation of gestures as ordered sequences of segments. The method is based on Hierarchical Hidden Markov Models (HHMM) (Françoise et al. (2011)) with a two–level structure. The lower level models the fine time structure of a segment using a template–based approach identical to GF. The higher level governs how segments can be sequenced by a high–level transition structure, which probabilities constrain the possible transitions between segments. Thus, the model can be built from a single demonstration of the gesture complemented by prior annotation defining the segmentation. The recognition is based on a forward algorithm allowing for the causal estimation of the performed segment (informed by the high-level transition structure) and the time position within this segment (as detailed in the the previous section). This representation provides both fine–grained and high–level control possibilities, allowing to reinterpret gestures through a segment-level decomposition which can be authored by the performer.

References

Arfib, D., J. M. Couturier, L. Kessous, and V. Verfaille. 2002. "Strategies of mapping between gesture data and synthesis model parameters using perceptual spaces." *Organised Sound* 7(02):127–144.

Bevilacqua, F., N. Schnell, N. Rasamimanana, B. Zamborlin, and F. Guédy. 2011. "Online Gesture Analysis and Control of Audio Processing." *Musical Robots and Interactive Multimodal Systems* :127–142.

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

Cadoz, C. 1988. "Instrumental gesture and musical composition." In *Proceedings of the International computer music conference*. Cologne, Germany.

Caramiaux, B. 2012. "Studies on the relationship between Gesture and Sound in Musical Performance." Ph.D. thesis, University of Paris 6, UMR STMS IRCAM CNRS.

Caramiaux, B., F. Bevilacqua, T. Bianco, N. Schnell, O. Houix, and P. Susini. 2014. "The Role of Sound Source Perception in Gestural Sound Description." *ACM Transactions on Applied Perception (in press)* .

Caramiaux, B., F. Bevilacqua, and N. Schnell. 2010a. "Analysing Gesture and Sound Similarities with a HMM-based Divergence Measure." In *Sound and Music Computing (SMC'10)*. Barcelona, Spain.

Caramiaux, B., F. Bevilacqua, and N. Schnell. 2010b. "Towards a gesture-sound cross-modal analysis." In *In Embodied Communication and Human-Computer Interaction, volume 5934 of Lecture Notes in Computer Science*. Springer Verlag, pp. 158–170.

Caramiaux, B., and A. Tanaka. 2013. "Machine Learning of Musical Gestures." In *New Interfaces for Musical Expression (NIME2013)*.

Castagne, N., C. Cadoz, J.-L. Florens, and A. Luciani. 2004. "Haptics in computer music: a paradigm shift." In *Proceedings of Eurohaptics*. Munich, Germany, pp. 174–181.

Chion, M. 1983. *Guide des objets sonores: Pierre Schaffer et la recherche musicale*. Buchet/Chastel.

Fadiga, L., L. Craighero, G. Buccino, and G. Rizzolatti. 2002. "Short communication: Speech listening specifically modulates the excitability of tongue muscles: A TMS study." *European Journal of Neuroscience* 15:399–402.

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

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

Françoise, J., B. Caramiaux, and F. Bevilacqua. 2011. "Realtime Segmentation and Recognition of Gestures using Hierarchical Markov Models." Master's thesis, Université Pierre et Marie Curie, Ircam.

Françoise, J., B. Caramiaux, and F. Bevilacqua. 2012. "A Hierarchical Approach for the Design of Gesture-to-Sound Mappings." In *Proceedings of the International Conference On Sound and Music Computing*. Copenhagen, Denmark.

Françoise, J., N. Schnell, and F. Bevilacqua. 2013. "A multimodal probabilistic model for gesture–based control of sound synthesis." In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, pp. 705–708.

Frey, A., C. Marie, L. Prod'Homme, M. Timsit-Berthier, D. Schön, and M. Besson. 2009. "Temporal Semiotic Units as Minimal Meaningful Units in Music? An Electrophysiological Approach." *Music Perception: An Interdisciplinary Journal* 26(3):pp. 247–256.

Gaver, W. W. 1993a. "How do we hear in the world? Explorations in ecological acoustics." *Ecological psychology* 5(4):285–313.

Gaver, W. W. 1993b. "What in the world do we hear?: An ecological approach to auditory event perception." *Ecological psychology* 5(1):1–29.

Gillian, N. 2011. "Gesture recognition for musician computer interaction." Ph.D. thesis, School of Music and Sonic Arts, Queen's University Belfast.

Godøy, R. I. 2006. "Gestural-sonorous objects: embodied extensions of schaeffer's conceptual apparatus." *Organised Sound* 11(2):149–157.

Godøy, R. I., E. Haga, and A. R. Jensenius. 2006. "Exploring Music-Related Gestures by Sound-Tracing. A Preliminary Study." In *2nd International Symposium on Gesture Interfaces for Multimedia Systems (GIMS2006)*.

Haueisen, J., and T. R. Knösche. 2001. "Involuntary Motor Activity in Pianists Evoked by Music Perception." *Journal of Cognitive Neuroscience* 13(6):786–792.

Hunt, A., and R. Kirk. 2000. "Mapping Strategies for Musical Performance." In M. M. Wanderley, and M. Battier, (editors) *Trends in Gestural Control of Music*. Ircam - Centre Pompidou, pp. 231–258.

Huron, D. 2002. "A Six-Component Theory of Auditory-Evoked Emotion." In *Proceedings of the 7th International Conference on Music Perception and Cognition*. Sydney, Australia.

Küssner, M. 2013. "Music and Shape." *Literary and Linguistic Computing* :1–8.

Lahav, A., A. Boulanger, G. Schlaug, and E. Saltzman. 2005. "The Power of Listening: Auditory-Motor Interactions in Musical Training." *Annals of the New York Academy of Sciences* 1060(1):189–194.

Large, E. W. 2000. "On synchronizing movements to music." *Human Movement Science* 19(4):527–566.

Large, E. W., and C. Palmer. 2002. "Perceiving temporal regularity in music." *Cognitive Science* 26(1):1–37.

Leman, M. 2007. *Embodied Music Cognition and Mediation Technology*. MIT press Cambridge, Massachusetts.

Leman, M., F. Desmet, F. Styns, L. Van Noorden, and D. Moelants. 2009. "Sharing musical expression through embodied listening: A case study based on Chinese Guqin music." *Music Perception* 26(3):263–278.

Liberman, A. M., and I. G. Mattingly. 1985. "The motor theory of speech perception revised." *Cognition* 21(1):1–36.

Maes, P.-J. 2012. "An empirical study of embodied music listening , and its applications in mediation technology." Phd dissertation, Ghent University.

Merer, A. 2011. "Caractérisation acoustique et perceptive du mouvement {é}voqu{é} par les sons pour le contr{ô}le de la synthèse." Ph.D. thesis, Université de Provence – Aix-Marseille 1.

Merleau-Ponty, M. 1945. “La Phénomenologie de la Perception.” *Gallimard Paris* .

Miranda, E., and M. Wanderley. 2006. “New digital musical instruments: control and interaction beyond the keyboard.” .

Mitchell, T. M. 2006. “The discipline of machine learning.” Technical report, Carnegie Mellon University, School of Computer Science, Machine Learning Department.

Momeni, A., and C. Henry. 2006. “Dynamic Independent Mapping Layers for Concurrent Control of Audio and Video Synthesis.” *Computer Music Journal* 30(1):49–66.

Noë, A. 2005. *Action in Perception*. Cambridge, USA: Massachusetts Institute of Technology Press.

Nymoen, K., B. Caramiaux, M. Kozak, and J. Tø rresen. 2011. “Analyzing Sound Tracings - A Multimodal Approach to Music Information Retrieval.” In *ACM Multimedia, MIRUM 2011*.

Rasamimanana, N., F. Bevilacqua, N. Schnell, F. Guedy, C. Maestracci, B. Zamborlin, J. Frechin, and U. Petrevski. 2011. “Modular Musical Objects Towards Embodied Control Of Digital Music.” In *Proceedings of the Tangible Embedded and Embodied Interaction Conference (TEI)*. pp. 9–12.

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

Schaeffer, P. 1966. *Traité des Objets Musicaux*. Éditions du Seuil.

Schwarz, D., N. Schnell, and S. Gulluni. 2009. “Scalability in Content-Based Navigation of Sound Databases.” In *Proceedings of ICMC*.

Tuuri, K., and T. Eerola. 2012. "Formulating a Revised Taxonomy for Modes of Listening." *Journal of New Music Research* 41(2):137–152.

Van Nort, D. 2009. "Instrumental Listening: sonic gesture as design principle." *Oganised Sound* 14(02):177–187.

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

Varela, F., E. Thompson, and E. Rosch. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, USA: Massachusetts Institute of Technology Press.

Wanderley, M. M. 2002. "Mapping strategies in real-time computer music." *Organised Sound* 7(2):83–84.

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

Wanderley, M. M., and N. Orio. 2002. "Evaluation of Input Devices for Musical Expression : Borrowing Tools from HCI." *Computer Music Journal* 26(3):62–76.

Zatorre, R. J., J. L. Chen, and V. B. Penhune. 2007. "When the brain plays music: auditory–motor interactions in music perception and production." *Nature Reviews Neuroscience* 8(7):547–558.