Mapping Through Listening

Baptiste Caramiaux,* Jules Françoise,† Norbert Schnell,† and Frédéric Bevilacqua†

*Department of Computing
Goldsmiths College
University of London
New Cross London, SE16 6NW, UK
b.caramiaux@gold.ac.uk
†STMS Lab IRCAM-CNRS-UPMC
Institut de Recherche et Coordination Acoustique/Musique
1 place Igor Stravinsky
75004 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 work on listening modes and gestural descriptions, we propose to distinguish between three mapping strategies: instantaneous, temporal, and metaphorical. 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.

In digital musical instruments, gestural inputs obtained from motion-sensing systems, image analysis, or sound analysis are commonly used to control or to interact with sound processing or sound synthesis (Miranda and Wanderley 2006). This has led artists, technologists, and scientists to investigate strategies for mapping 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 has led to a flourishing of research work dealing with: taxonomy (Wanderley 2002); the study of various strategies based, for example, on perceptual spaces (Arfib et al. 2002), mathematical formalization (Van Nort, Wanderley, and Depalle 2004), or dynamical systems (Momeni and Henry 2006); and evaluation procedures based on user studies and other tools borrowed from the field of human–computer interaction (Hunt and Kirk 2000; Wanderley and Orio 2002).

It has often been discussed that for digital music instruments, unlike most acoustic instruments (Cadoz 1988; Wanderley and Depalle 2004), there is no direct coupling between the gesture energy and the acoustic energy. More 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 unprecedented opportunities to create various types of mapping that can be seen as part of the creative endeavor to build 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. As a complement to approaches that focus on building active haptic feedback to enhance the action–perception loop (Castagne et al. 2004), we propose a methodology rooted in the concept of embodied music cognition. This methodology considers listening as a process from which gestures and interactions, defining key elements for the design of mappings, emerge.

Computer Music Journal, 38:3, pp. 34–48, Fall 2014
doi:10.1162/COMJ_a_00255

Our approach is anchored in advances in cognitive sciences and rooted in embodied cognition (Varela, Thompson, and Rosch 1991). The enactive point of view on perception and the idea of embodied cognition cover aspects of cognition as shaped by the body, which consitute the perceptual and motor systems (Varela, Thompson, and Rosch 1991; Noë 2005). From this point of view, the action of listening—as is the case with 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—as well as 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, Chen, and Penhune 2007). By extension, embodied music cognition, developed by Marc Leman (2007) and Rolf Inge Godøy (2006), tends to see music perception as based on actions. Many situations involve people moving while listening to music. In the framework of embodied music cognition, these movements are seen as conveying information about the perceived sonic moving forms (Leman et al. 2009).

Although 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 field of machine learning research 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 example, the use of machine-learning techniques allows one to set the gesture–sound relationships from examples or from a database.

In this article 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 following section, we review previous work characterizing different listening modes and how they relate to gestural descriptions of sounds. We then describe our approach for the design of mappings inspired by these 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 final section, we discuss the different scenarios 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, going from gesture to sound, to create the mapping. In this section we first review work describing different listening modes that can be related to specific sound properties. Then we show that these listening modes can be related to different action strategies.

Listening Modes

Sound, as considered here, refers to recorded audio material. Recorded sound can be played back and processed using various techniques, which, importantly, leads to different listening experiences. A vast body of work is devoted to the mechanisms of listening, gathering together various research fields such as psychoacoustics, neurosciences, auditory scene analysis, and musicology. In this section, we focus on conceptual approaches of listening that principally originated from music theory and the theory of ecological perception. 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*, Pierre Schaeffer (1966) defined four functions of listening. (Note that the translation of Schaeffer's terms is far from trivial. For this reason, in this article we use both our translation and the original French term.) These functions are: (1) *listening* (*écouter*), which focuses on the indexical value of the sound (i.e., the sound source); (2) *perceiving* (*ouïr*), the most primitive mode, consisting of receiving the sound through the auditory system; (3) *hearing* (*entendre*), referring to the selective process between auditory signals, the attention to inherent characteristics of the perceived sound; and (4) *comprehending* (*comprendre*), which brings semantics into sounds, treating them as signs. These different functions of listening are not mutually exclusive and operate competitively.

Based on Schaeffer's theoretical taxonomy, and motivated by new concepts from auditory display, Michel Chion (1983) proposed a taxonomy comprising three categories, called modes of listening: (1) *causal* listening, consisting of listening to a sound in order to gather information about its cause (or source); (2) *semantic* listening, refering to a code or a language to interpret a message; and (3) *reduced* listening, focusing on the qualities of the sound itself, independent of its cause and of its meaning. (Note that reduced listening is a concept that was first introduced by Schaeffer to motivate the concept of the "sound object" in *musique concrète*.) 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, William Gaver (1993a, b) considered environmental sounds and proposed a differentiation between two types of listening 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 noted by Gaver (1993b, p. 1), musical listening to environmental sounds can be achieved by listening "to the world as we do music." Gaver used, as examples, compositions by the American composer John Cage that aim at hearing the everyday world as music.

Recent studies have proposed to enrich these previous taxonomies by adding an emotional dimension, evoked by the auditory stimulus. David Huron (2002) proposed an analytic framework supporting the idea that emotional experiences may be usefully characterized according to a six-part classification, categorized as follows: (1) *reflexive*, referring to fast, automatic physiological responses; (2) *denotative*, allowing the listener to identify sound sources; (3) *connotative*, allowing the listener to infer various physical properties about sound sources such as size, proximity, energy, material, and mode of excitation; (4) *associative*, referring to arbitrary learned associations; (5) *empathetic*, referring to auditory empathy that allows the listener to detect emotion from the sound (such as fear in a voice) coming from an animate agent (be it human or animal); and, finally, (6) *critical*, referring to conscious cognitive processes by which the intentions of a sound-producing agent are evaluated.

Recently, Kai Tuuri and colleagues proposed an extended taxonomy of listening modes (Tuuri and Eerola 2012). The taxonomy is hierarchical with three levels: experiential, denotative, and 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 (i.e., object, people, and cultural context). In this 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 was first defined by Huron and extended by Tuuri in order to separate between modes focusing on sound sources and those focusing on sound contexts. Finally, the top level is the *reflective* mode, encompassing Chion's reduced mode as well as Huron's 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 Chion's causal listening mode and the denotative

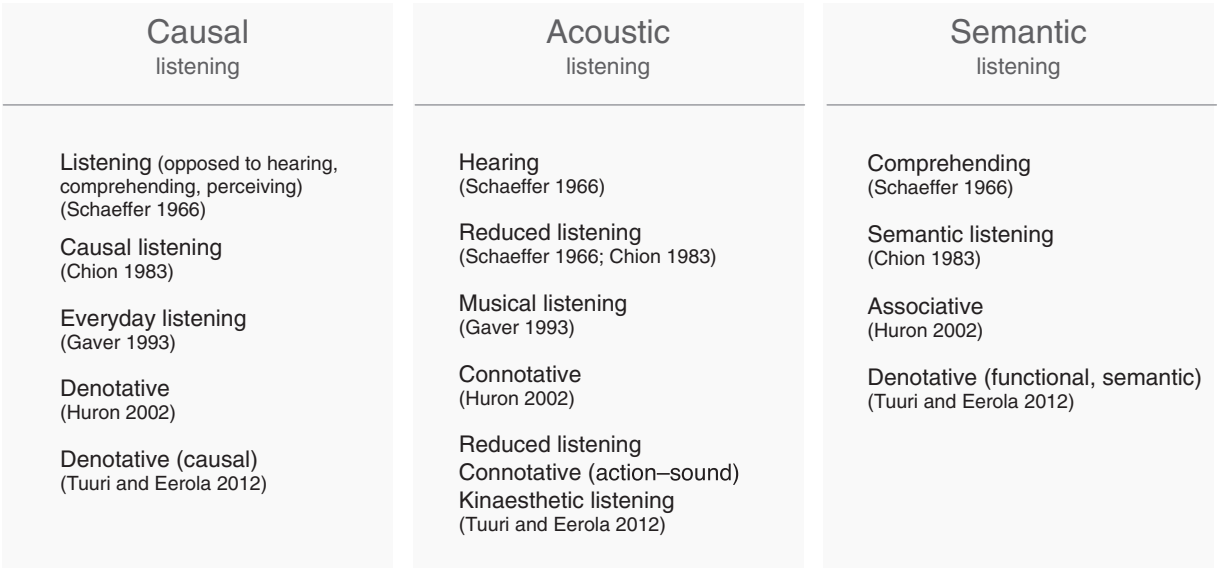


Figure 1. A simplified taxonomy of listening modes. Causal listening refers to an explicit association between sound and its producing action. Acoustic listening is related to acoustic qualities of the sound. Semantic listening integrates higher level notions of meaning and interpretation.

listening mode, used by both Huron and Tuuri, refer to associating a 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, Huron's connotative mode, and Tuuri's kinaesthetic mode 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. A crucial point, however, is to acknowledge that defining the reduced listening mode is also linked to sound descriptions such as the Schaeffer's typomorphology of sonic objects (Schaeffer 1966), or later to temporal semiotic units (*unités sémiotiques temporelles*, cf. Frey et al. 2009). As elucidated by 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 is named accordingly. Figure 1 summarizes the three modes of listening—causal, acoustic, and semantic—that we will consider in this article, with the goal of associating them with gestural representations.

Linking Gesture and Listening

In the previous section we reviewed the listening modes as introduced by various authors in the literature. These were summarized as an approach using three modes, 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 work 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, Chen, and Penhune 2007). Generally, the motor–auditory interaction has been recognized as important for describing 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, Chen, and Penhune 2007).

In a behavioral approach, a common experimental methodology consists of asking participants to perform movements along with 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 work concerns controlled tasking, for instance, the task of tapping on beats (Large 2000; Large and Palmer 2002).

In systematic musicology, exploratory procedure is more commonly used. Examples include asking participants to spontaneously gesticulate while listening to a sound stimulus or music. For instance, Leman and co-workers (2009) studied participants' movements made with a joystick while listening to a performance of guqin music. Also, Mats Küssner (2013) considered free tracing movements on a tablet while two of Frédéric Chopin's preludes were played. Other works concern specifically designed stimuli with well characterized musical parameters. Consequently, it is possible to investigate how the chosen musical parameters affected the resulting movements.

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

This previous work provides us with a 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* (or *analog*) experiments, where the motion trajectories are associated with acoustic parameters. In addition, in the following we will refer to sound morphology to designate the temporal profile of the acoustic characteristics of sound (e.g., amplitude, pitch, and timbral aspects).

In prior work (Caramiaux et al. 2014), we conducted experiments to give evidence regarding 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: 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 identification of sound sources (i.e., the mode of listening) has a direct consequence on the gestural strategies. If the participants can identify the sound source as an action, they tend to mimic the action. On the other hand, a sound that cannot be identified leads participants to trace the profile of perceived sound features.

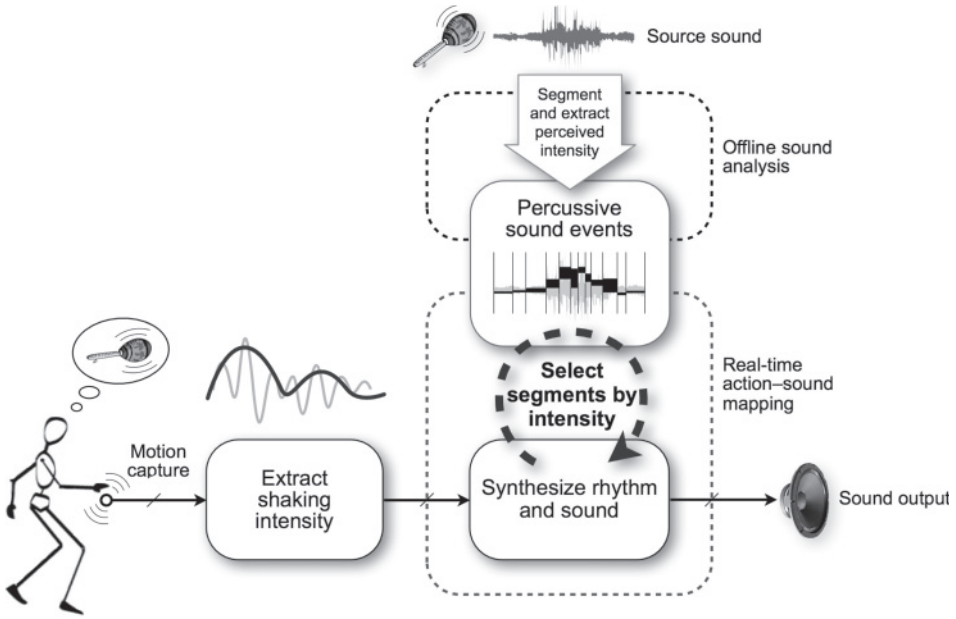
This experimental study showed a link between acoustic (or causal) listening modes and analog (or mimicking) motion strategies. This study provides a rationale for establishing mapping strategies based on listening modes and associated motion strategies. Mapping strategies can stem from the reviewed experimental findings, and 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 metaphorical listening modes and gestural strategies.

From Listening to Controlling

In this section we describe concrete examples that we developed and that were used in different settings, from experiments and demonstrations to interactive installations and performances. All these examples are based on modeling the target sound from a gestural perspective: a prior listening to (or evocation of) the sound provides performers with insights into possible strategies for gesture control. These strategies are then made possible using machine-learning techniques. Similar approaches have been described by Godøy (2006), Doug Van Nort (2009), Rebecca Fiebrink (2011), and Pieter-Jan 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 one to consider scenarios and metaphors where the motion

Figure 2. The 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.



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

Four interactions have been created that implement distinct mapping strategies illustrating the approach. These four scenarios are shaking, shaping, fishing, and shuffling.

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. Although this metaphor may refer, in music performance, to percussion instruments such as a shaker or maracas, it can also be associated with various nonmusical 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 it can be applied to any movement that resembles shaking or waving (i.e., movements that are periodic and modulated in intensity).

This mapping is designed to be a direct relationship between the movement energy and the energy of the sound played. The sound can, however, 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.

We use accelerometers to sense the performer's motion. Concrete implementations were featured in different performances using the musical object interfaces (e.g., performances at the 2011 Margaret Guthman Musical Instrument Competition or the 2013 International Conference on Tangible, Embedded and Embodied Interaction, cf. 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 each segment. Both feature spaces, motion and sound, are normalized, so that each sound segment can be associated with a corresponding shaking intensity lying between the lowest and highest possible values. The system used a *k*-nearest neighbor (*k*-NN) search algorithm based on a *k*-dimensional (*k*-D) tree to select a sound event of a given intensity from among the available segments (Schwarz, Schnell, and Gulluni 2009).

Shaping

Shaping refers to scenarios where performers control sound morphologies by "tracing" in the air those salient sound features they desire to control. It is thus related to acoustic listening as we defined previously, where the performer pays attention to acoustic qualities 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 tightly reflect the aspects of the sound the performer perceives and intends to affect. Again, the mapping relies on two distinct phases: learning and playing. The learning phase consists of 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 gesturally draws the morphology of a particular sound and replays 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, using a real-time shape-matching algorithm that finds, at each time step, the audio-feature morphology closest to the gesture morphology and aligns the two morphologies temporally. Note that the algorithm can be configured to allow transitions between gestures, which enables the algorithm to switch between sounds during the execution of a gesture. Figure 3 illustrates the scenario.

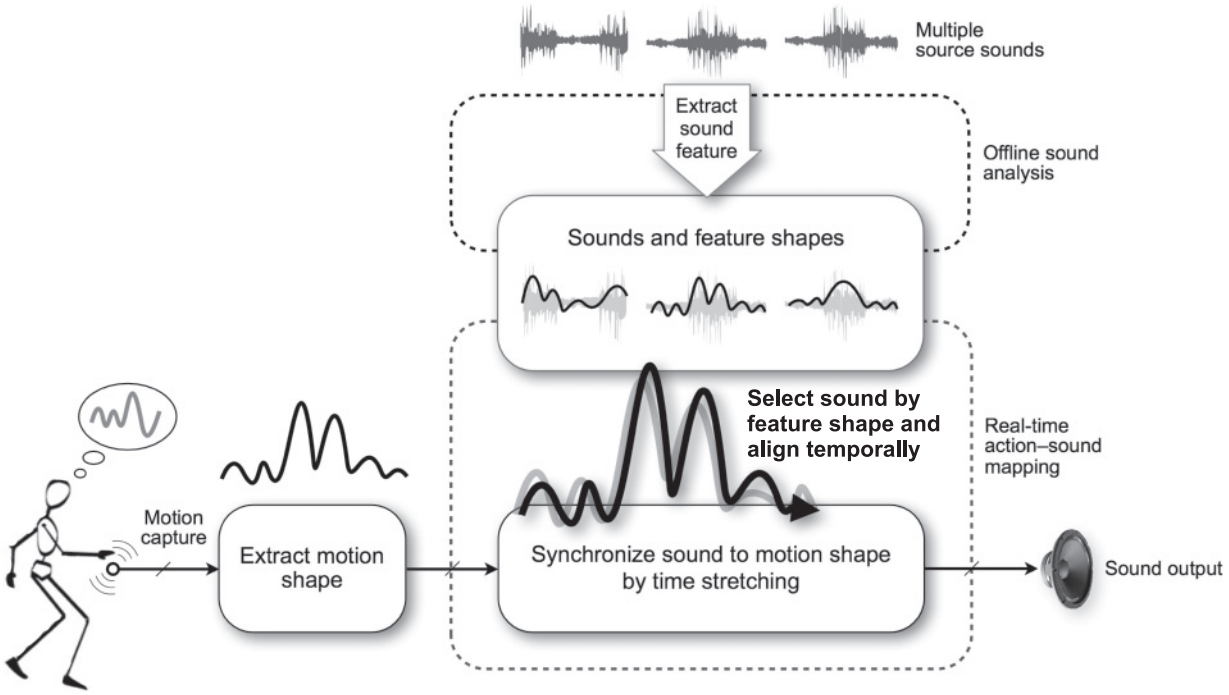
The implementation, called the gesture follower, is based on a machine-learning technique using hidden Markov models (HMMs) and is presented in the Appendix. Because the sound is aligned to the gesture in real time, it translates the variations in the gesture morphology, such as the speed of execution, to variation in the playback, reinterpreting the recorded sound. In a demonstration presented at the 2010 Sound and Music Computing Conference (Caramiaux, Bevilacqua, and Schnell 2010a), gesture and sound were represented by a unidimensional time series, the energy of a gesture controlling the loudness. The energy of a gesture was computed as its absolute speed (an infrared 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 the performer mimics an action in order to select and play a specific sound. In other words, the performer virtually "fishes" for 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 learning 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

Figure 3. The 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.



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 playback of theassociated sound. Because the system relies on action recognition, both the performed and the predefined actions must have a consistent representation, which could imply they were 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.

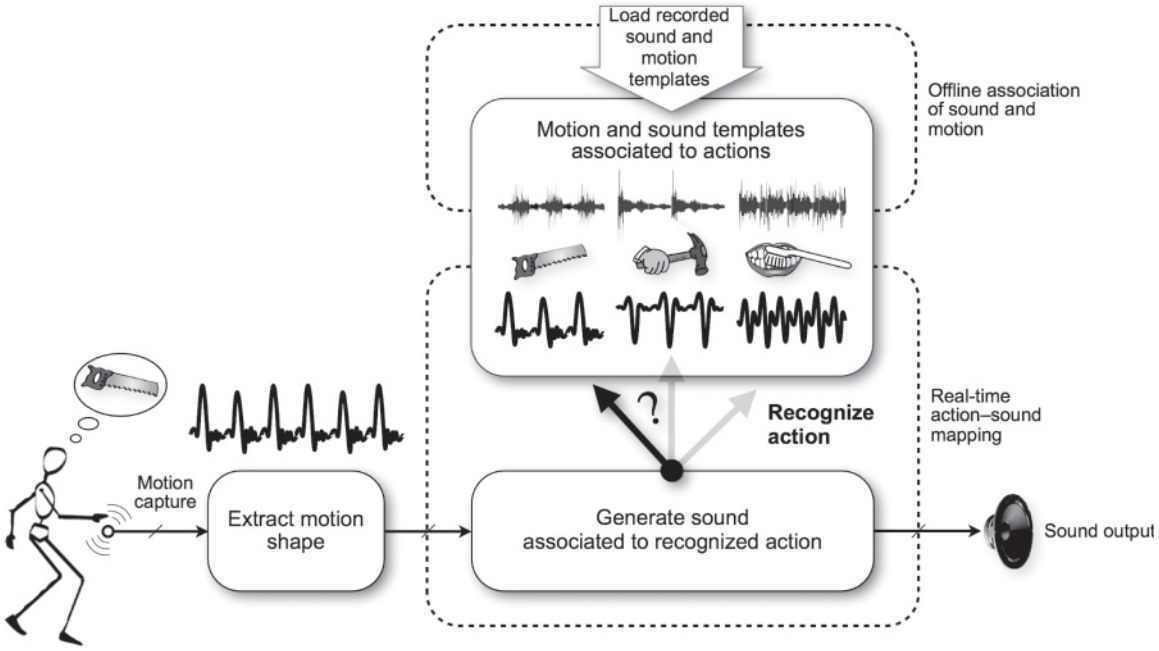
The system uses the same algorithm (the gesture follower) as the shaping scenario presented in the Appendix. In the installation version of the system, presented during a meeting of Sound and Music for Everyone Everyday Everywhere Everyway project (SAME, www.sameproject.eu/), the actions were captured 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 pairs from the database was presented to the user in order to be mimicked. The algorithm was set to play the sound associated with an action as soon as this action is recognized. In addition, the algorithm was set to output the time progression in the executed action. When the user reached 90% of the recognized action, the sound was 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 gesturally recomposing and reinterpreting complex sound sequences. 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 interactively implement 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

Figure 4. The fishing scenario. A set of recorded sounds 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.



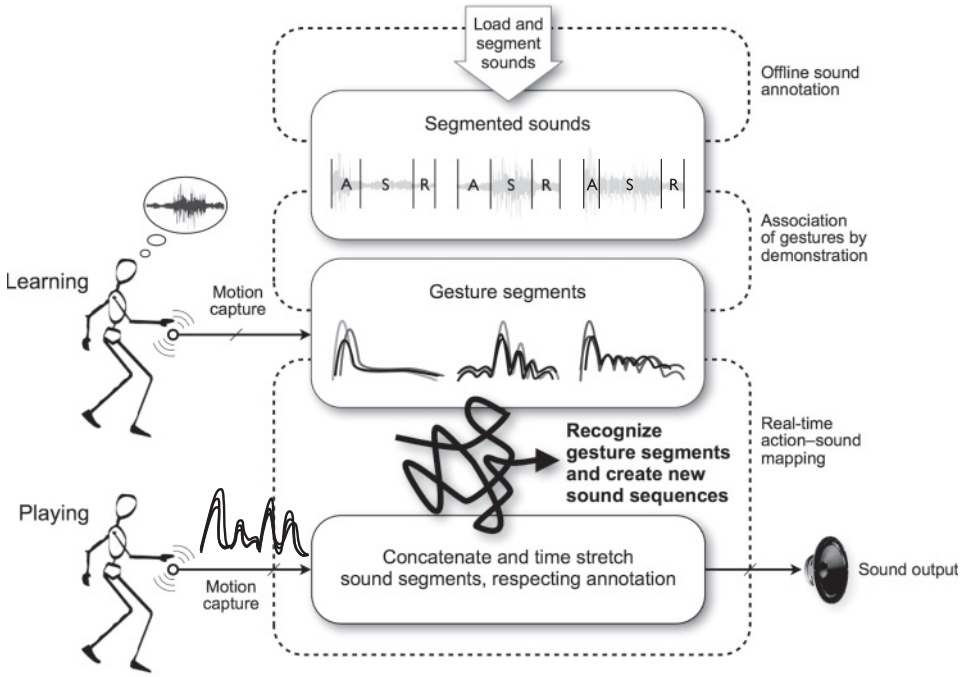
machine-learning model that encodes their relationships. When the performers perform a new gesture sequence, sounds are resynthesized and aligned in real time, using phase vocoding. In some aspects, the present scenario generalizes some of the previous examples by allowing the performer to mimic sound-producing 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 two distinct phases: learning and playing. During the learning phase, the performer begins by selecting sounds and manually defining their segmentation using a graphical editor. Then the performer records one or multiple gestures associated with 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 dynamically select and replay the appropriate sequence of sound segments along with the gesture performance. Figure 5 illustrates the shuffling scenario.

The mapping is based on a hierarchical model for continuous gesture recognition and segmentation, called a hierarchical HMM (see the Appendix for details). The model has two levels. The lower level precisely encodes the time structure of the segments, and 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 by Françoise, Caramiaux, and Bevilacqua (2012). Each gesture and each sound morphology is segmented

Figure 5. The 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.



into attack, sustain, and release segments, possibly complemented by a preparation phase anticipating the attack of the sound.

Two aspects of this decomposition are particularly interesting.

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 segments, can be shaped by the performer's gesture. (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.)

Second, this decomposition allows for designing strategies that involve multiple gestural descriptions related to listening: For example, 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.

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 gesturally "reinterpret" 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 the section "Describing Sound Gesturally."

Figure 6 summarizes how the examples are related to the different listening modes and gestural strategies we have discussed. In addition, we require the different strategies of mapping that are used in the different examples. We distinguish between instantaneous, temporal, and metaphorical aspects that define the relationship between gesture and

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.

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

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, which do not necessarily rely on morphological congruences between gesture and sound.

The shaking scenario makes use, principally, 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 a metaphorical association between the action and sound. Owing to the strong action metaphor, the scenario can also supply completely unconventional sounds for the performer to shake.

In the shaping scenario 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 (2006) or Van Nort (2009). The difference with shaking resides in the precise control over the sound's temporal evolution, supporting a listening mode focussed on acoustic sound features. Our experiments with this scenario showed that a sonic profile must be memorized beforehand in order to consciously target it and, eventually, to reproduce it with temporal variations.

The shaping scenario makes use of a temporal mapping between gesture parameters and sound features. This mapping allows the performer to reshape a sound based on the temporal morphology of his or her gesture. The general concept of temporal mapping was previously introduced by 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 metaphorical: Unlike the shaking and shaping scenarios, the morphologies of gesture and sound in this example can be incongruent in some cases. The action–sound relationship is, nevertheless, clear from the perspective of causal listening. As mentioned previously, this scenario has been shown at an installation during the EU Project SAME. Feedback 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 metaphorical 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 metaphorical. The temporal characteristic of the mapping is similar to the shaping scenario, and the metaphorical characteristic is enabled by the implementation of a general algorithm for the recognition of actions 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 (as in shaping) and the control of sound segments through iconic gesture segments (as in fishing). The shuffling example can be seen as an unified approach in the sense that it can be configured to activate several modes of listening and several modes of gestural description (and it 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. We believe, nevertheless, that temporal mapping strategies are equally powerful when considering other synthesis methods. They allow one to segment the performers' actions and to define different action–sound relationships for different segments. There is, for example, a need for a distinction between action segments that actually produce sound or induce sound changes, and those that do not.

One design choice in the examples presented here concerns the motion-sensing technology. Any sensing system provides a partial gesture description, which might impact sound controllability. In the four scenarios presented, we used accelerometers. Although these sensors have inherent limitations (e.g., they are unable to sense spatial information), they are sensitive to small changes in orientation and dynamics. The choice, moreover, has been motivated by other advantages of this technology: low cost, wireless, well-understood signal characteristics, and sufficient precision for most musical applications.

The scenarios discussed in this article make extensive use of methods based on machine learning (*k*-NN, HMM, hierarchical HMM). The role of machine learning is to implement the top–down approach of our scenarios based on perceptual or metaphorical action and sound description. Indeed, all scenarios imply implicit relationships between sound and gestural features. As discussed by Tom Mitchell (2006), machine-learning techniques are effective for modeling such implicit relationships. Moreover, such an approach has started 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, Bevilacqua, and Schnell 2010b), for jointly modeling their interactions over time to implicitly capture their correlations and the expressive variations emerging in different interpretations (Françoise, Schnell, and Bevilacqua 2013), or the use of machine learning as a design tool (Fiebrink, Cook, and Trueman 2011).

The possibilities arising from the introduction of machine-learning techniques into the interaction loop are twofold. First of all, they allow the instrument to integrate notions of recognition and prediction that support the implementation of interactions based on the performer's listening. As the performer always adapts his or her actions to the behavior of the instrument—either spontaneously or by strenuous learning—these new instruments, for their part, adapt themselves to the performer's behavior, preferences, and playing style. It is worth noticing that machine-learning techniques are prone to errors or may require time to converge to an accurate estimate. Latency is inherently involved, and it may 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 percent of the action, as a visual progress bar for the user. Interestingly, with latency represented in this manner, it challenged the user during the interaction, enhancing the game play.

In conclusion, we propose a design approach for mapping based on the concept of embodied listening. Building on previous work on listening modes and gestural descriptions we propose to distinguish three mapping strategies: instantaneous, temporal, and metaphorical. 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 only scratch the surface of the possibilities arising from this approach.

Acknowledgments

This work is supported by the mixed research lab Sciences and Technologies for Music and Sound (STMS), the Institut de Recherche et Coordination Acoustique/Musique (IRCAM), the Centre National de la Recherche Scientifique (CNRS), the Université Pierre et Marie Curie (UPMC), and the Legos project (ANR 11 BS02 012).

References

Arfib, D., et al. 2002. "Strategies of Mapping Between Gesture Data and Synthesis Model Parameters Using Perceptual Spaces." *Organised Sound* 7(02):127–144.

Bevilacqua, F., et al. 2010. "Continuous Realtime Gesture Following and Recognition." In S. Kopp and I. Wachsmuth, eds. *Gesture in Embodied Communication and Human–Computer Interaction*. Berlin: Springer, pp. 73–84.

Bevilacqua, F., et al. 2011. "Online Gesture Analysis and Control of Audio Processing." In J. Solis and K. Ng, eds. *Musical Robots and Interactive Multimodal Systems*. Berlin: Springer, pp. 127–142.

Cadoz, C. 1988. "Instrumental Gesture and Musical Composition." In *Proceedings of the International Computer Music Conference*, pp. 1–12.

Caramiaux, B., F. Bevilacqua, and N. Schnell. 2010a. "Analysing Gesture and Sound Similarities with a HMM-Based Divergence Measure." In *Proceedings of the Sound and Music Computing Conference*. Available online at smcnetwork.org/files/proceedings/2010/9.pdf. Accessed March 2014.

Caramiaux, B., F. Bevilacqua, and N. Schnell. 2010b. "Towards a Gesture–Sound Cross-Modal Analysis." In S. Kopp and I. Wachsmuth, eds. *Gesture in Embodied Communication and Human-Computer*. Berlin: Springer Verlag, pp. 158–170.

Caramiaux, B., and A. Tanaka. 2013. "Machine Learning of Musical Gestures." In *Proceedings of the Conference on New Interfaces for Musical Expression*. Available online at baptistecaramiaux.com/blog/wp-content/uploads/2013/05/nime2013_mlrev.pdf. Accessed March 2014.

Caramiaux, B., et al. 2014. "The Role of Sound Source Perception in Gestural Sound Description." *ACM Transactions on Applied Perception* 11(1):1–19.

Caramiaux, B., et al. In press. "Adaptive Gesture Recognition with Variation Estimation for Interactive Systems." *ACM Transactions on Iterative Intelligent Systems*.

Castagne, N., et al. 2004. "Haptics in Computer Music: A Paradigm Shift." In *Proceedings of the Eurohaptics Meeting*, pp. 174–181.

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

Fadiga, L., et al. 2002. "Speech Listening Specifically Modulates the Excitability of Tongue Muscles: A TMS Study." *European Journal of Neuroscience* 15(2):399–402.

Fiebrink, R. A. 2011. "Real-Time Human Interaction with Supervised Learning Algorithms for Music Composition and Performance." PhD Thesis, Princeton University, Department of Computer Science.

Fiebrink, R. A., 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*, 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 (Paris VI).

Françoise, J., B. Caramiaux, and F. Bevilacqua. 2012. "A Hierarchical Approach for the Design of Gesture-to-Sound Mappings." In *Proceedings of the Sound and Music Computing Conference*. Available online at smcnetwork.org/system/files/smc2012-203.pdf. Accessed March 2014.

Françoise, J., N. Schnell, and F. Bevilacqua. 2013. "A Multimodal Probabilistic Model for Gesture-Based Control of Sound Synthesis." In *Proceedings of the ACM International Conference on Multimedia*, pp. 705–708.

Frey, A., et al. 2009. "Temporal Semiotic Units as Minimal Meaningful Units in Music? An Electrophysiological Approach." *Music Perception* 26(3):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." PhD Thesis, Queen's University Belfast, School of Music and Sonic Arts.
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 *Proceedings of the International Symposium on Gesture Interfaces for Multimedia Systems*, pp. 27–33.
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, eds. *Trends in Gestural Control of Music*. Paris: Institut de Recherche et Coordination Acoustique/Musique, pp. 231–258.
Huron, D. 2002. "A Six-Component Theory of Auditory-Evoked Emotion." In *Proceedings of the International Conference on Music Perception and Cognition*, pp. 673–676.
Küssner, M. 2013. "Music and Shape." *Literary and Linguistic Computing* 28(3):1–8.
Lahav, A., et al. 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*. Cambridge, Massachusetts: MIT Press.
Leman, M., et al. 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 évoqué les sons pour le contrôle de la synthèse." PhD Dissertation, Université de Provence Aix-Marseille 1.
Merleau-Ponty, M. 1945. *La Phénomenologie de la Perception*. Paris: Gallimard.
Miranda, E., and M. Wanderley. 2006. *New Digital Musical Instruments: Control and Interaction beyond the Keyboard*. Middleton, Wisconsin: A-R Editions.
Mitchell, T. M. 2006. "The Discipline of Machine Learning." Technical Report CMU-ML-06-108. Pittsburgh, Pennsylvania: 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, Massachusetts: MIT Press.
Nymoen, K., et al. 2011. "Analyzing Sound Tracings: A Multimodal Approach to Music Information Retrieval." In *Proceedings of the International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies*, pp. 39–44.
Rasamimanana, N., et al. 2011. "Modular Musical Objects Towards Embodied Control of Digital Music." In *Proceedings of the International Conference on Tangible, Embedded, and Embodied Interaction*, pp. 9–12.
Rovan, J., et al. 1997. "Instrumental Gestural Mapping Strategies as Expressivity Determinants in Computer Music Performance." In *Proceedings of Kansei: The Technology of Emotion Workshop*, pp. 68–73.
Schaeffer, P. 1966. *Traité des Objets Musicaux*. Paris: Éditions du Seuil.
Schwarz, D., N. Schnell, and S. Gulluni. 2009. "Scalability in Content-Based Navigation of Sound Databases." In *Proceedings of the International Computer Music Conference*, pp. 13–16.
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 Conference on New Interfaces for Musical Expression*, pp. 87–91.
Varela, F., E. Thompson, and E. Rosch. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, Massachusetts: MIT 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.

Appendix: Algorithms

In this appendix we describe the algorithm used in the interaction scenarios. Rather than a full technical specification, we outline the model used and how the model has been adapted to the context.

Gesture Follower

The gesture follower (GF, cf. Bevilacqua et al. 2010) is a template-based gesture-recognition method based on HMMs. The model is learned from a single example gesture, using its whole time series as a template. The model is built by assigning a state to each frame, similarly to dynamic time warping. The time structure is modeled by a left-to-right transition structure. A causal forward inference allows for decoding in real time 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 nonlinear dynamic system. It allows for the continuous adaptation to variations of gesture characteristics (Caramiaux et al. in press). Indeed, once the gesture template is recorded, a similar live gesture can be performed with variations in speed, scale, rotation, etc. These characteristics can be explicitly taken into account by the method as invariant for the recognition. To that extent, the method continuously estimates the relative characteristics of the gestural variations, which can then be used in continuous interaction scenarios.

Hierarchical Extension

The gesture follower has been extended to comprehend time structures that are more complex, allowing for the representation of gestures as ordered sequences of segments. The method is based on hierarchical HMMs with a two-level structure (Françoise, Caramiaux, and Bevilacqua 2011). The lower level models the fine time structure of a segment using a template-based approach identical to the GF. The higher level governs how segments can be sequenced by a high-level transition structure, whose 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 one to reinterpret gestures through a segment-level decomposition that can be authored by the performer.