Gestural Embodiment of Environmental Sounds: an Experimental Study

B. Caramiaux, P. Susini, T. Bianco, F. Bevilacqua, O. Houix, N. Schnell, N. Misdariis

IMTR and PDS Team
Ircam - CNRS
1 place Igor Stravinsky
75004, Paris, France

baptiste.caramiaux@ircam.fr

ABSTRACT

In this paper we present an experimental study concerning gestural embodiment of environmental sounds in a listening context. The presented work is part of a project aiming at modeling movement-sound relationships, with the end goal of proposing novel approaches for designing musical instruments and sounding objects. The experiment is based on sound stimuli corresponding to "causal" and "non-causal" sounds. It is divided into a performance phase and an interview. The experiment is designed to investigate possible correlation between the perception of the "causality" of environmental sounds and different gesture strategies for the sound embodiment. In analogy with the perception of the sounds' causality, we propose to distinguish gestures that "mimic" a sound's cause and gestures that "trace" a sound's morphology following temporal sound characteristics. Results from the interviews show that, first, our causal sounds database lead to consistent descriptions of the action at the origin of the sound and participants mimic this action. Second, non-causal sounds lead to inconsistent metaphoric descriptions of the sound and participants make gestures following sound "contours". Quantitatively, the results show that gesture variability is higher for causal sounds that non-causal sounds.

Keywords

Embodiment, Environmental Sound Perception, Listening, Gesture Sound Interaction

1. INTRODUCTION

In the context of music playing as well as music listening, movements and actions related to musical stimuli can be seen as the embodied manifestation of sound/music perception and cognition [14, 7]. In the cognitive neuroscience literature, previous works have shown some evidences for music embodiment in the auditory-motor systems interaction during music performance (see [17] for a review). For instance, people naturally tap the beat while listening to a piece of music and often anticipate the rhythmic accents [11, 12]. In [4], the authors investigate a more abstract relationship between body motion and music that is examining whether changes in musical parameters evoke corresponding changes in listeners' spatial and kinetic imagery. In parallel, a need for a coherent typology of music-related gestures or actions has emerged [2]. A movement reacting to sonorous stimuli can be qualified as sound-accompanying gestures [10], i.e. gestures that are not involved in the physical production of sound but rather are reflecting some important aspects in sounds.

Godøy et al. have conducted two experimental studies showing two sub-categories of sound-accompanying gestures: gestures that mimic instrumental performances [9] and sound-tracing gestures [8]. While the first study is concerned by musical piece stimuli, the second involves a larger set of sounds from musical instruments, electronic sounds or environmental sounds (taken as *concrete* sounds in the sense of Schaeffer [15]).

Through their explorative works, Godøy et al. have highlighted two interesting strategies in music embodiment: *mimicking* and *tracing*. However, they were studied independently with two distinct experimental protocols. We believe that both strategies constitute an important dichotomy in gestural sound embodiment and precisely when considering environmental sounds. To that extent, it seems pertinent to considered them jointly. The experiment presented in this paper aims to characterize both *mimicking* and *tracing* strategies through an unique experimental protocol.

Computational characterization of these two strategies related to environmental sounds can be insightful for the design of virtual instruments and sound design tools. Mimicking can be transcribed as the excitation of a specific physical model while tracing can be transcribed as the instantaneous mapping between gesture features and audio features. Both strategies can lead to a wide range of applications for sonic interaction design as well as future theoretic studies.

The paper is organized as follows. The next section aims at placing our contribution in the state of the art. As far as we know, very few works exist on characterization of embodied listening of environmental sounds. Therefore, the related work is focused on sound perception and listening strategies. Our methodology is reported in section 3. This is the starting point for our experimental study that is divided into two steps. First we present the sound stimuli in section 4 then we define an experimental protocol to evaluate our hypothesis in section 5. Results are presented inside of each section. Finally we conclude in section 6 giving some ongoing short-term perspectives.

NIME'11, 30 May–1 June 2011, Oslo, Norway.

2. RELATED WORKS

In environmental sound perception, Gaver in [5, 6] has proposed the distinction between *musical listening* and *everyday listening*. In musical listening the listener focuses on acoustic qualities and other musical aspects of sound while in everyday listening the listener focuses on causal aspects. In the task of categorization of sounds, this suggests that some people will consider as similar two sounds with the same acoustic characteristics and others will consider as similar two sounds with the same cause. Following these previous studies, Lemaitre et al. in [13] have shown the categorization of environmental sounds is influenced by the listener's expertise and the sound identification (i.e. if the cause that has produced the sound is identifiable or not). They showed that in categorization task, people will more frequently base their choice on acoustic characteristics if the identification of the cause is difficult (i.e. the causal uncertainty is high). On the other hand, people will frequently use the sound's cause as categorization criterion if the causal uncertainty is low.

Our contribution is to propose an experiment that analyses how people embody musical or everyday listening of environmental sounds. The methodology is exposed in the next section.

3. HYPOTHESIS

Previous works [9, 8, 3] roughly depict two categories for gesture embodiment of environmental sounds: gestures mimicking the action that has produced the sound and gestures following (or tracing) the temporal evolution of the perceived sound features. In the following we will use the terms *symbolic* referring to the gestures from the first category and *morphologic* referring to the gestures from the second category. This terminology emphasizes the distinction between the symbol and the shape. These two terms are not established and a deeper discussion about their use is part of our prospective works.

Consider the following experimental methodology. We propose to consider causal sounds and to synthetically take off the causality by an audio process (that roughly corresponds to retain the global energy evolution whereas timbre characteristics are flatten). Then we ask for people to associate gestures while listening these causal and non-causal sounds.

The goal is to analyze the gesture and sound data to explore the following hypothesis: *causal* sounds imply *symbolic* gestures and *non-causal* sounds induce *morphologic* gestures?

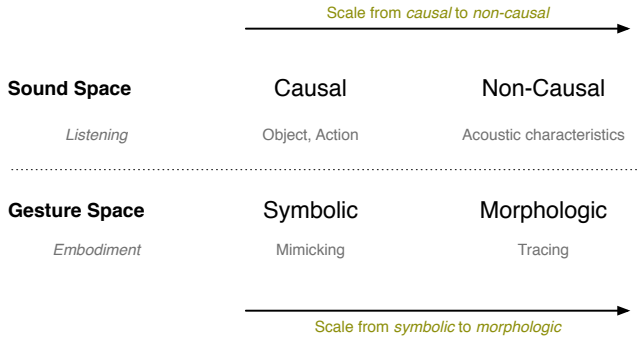

Figure 1: Scheme of the global working context of the experimental study

4. SOUND SELECTION PROCEDURE

The sounds used in the study belong to a domestic context (usual objects found in a kitchen), to ensure that the sources of the sounds were likely to be known to all listeners [13]. Each sound identification is calculated through the causal uncertainty index (noted H_{cu}) [1, 13] that measures the identification of the cause in terms of action and/or object verbalized description. Each sound has a H_{cu} index scaled between 0 (i.e. all the participants provided the same description of the sound in terms of action or object) and 4.75 (all the participants provided a different description in terms of action or object). However, the procedure of measuring H_{cu} is very time-consuming and needs for a precise semantic analysis of verbalizations. Instead the authors propose to measure the confidence in the identification by an usual scale between 1 and 5.

1. "I don't know at all"
2. "I am really not sure"
3. "I hesitate between several causes"
4. "I am almost sure"
5. "I perfectly identify the cause of the sound"

Lemaitre et al show that the resulting measure is correlated to H_{cu} even if both measures do not provide exactly the same information. From this previous study, we have selected the ten most identified sounds (low H_{cu}) in the kitchen sounds database in order to define a first corpus, namely the "causal" sounds. Having the corpus of causal sounds, we build a second corpus by applying an audio process transforming the sounds taken from the first corpus. We design a sound transformation that takes the original causal sound and returns a sound with the same energy evolution but having occulted some of the timbre aspects. The transformation is convolution-based and is illustrated by figure 2. In this figure, the reader can see that the temporal evolution of the mel cesptrum remained whereas timbre characteristics of original sounds are flatten.

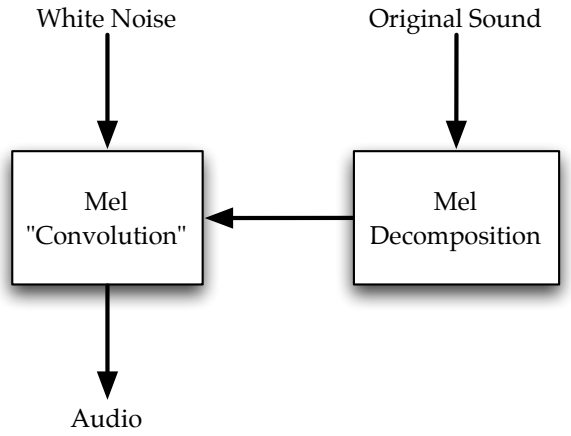

Figure 2: Scheme of the sound transformation used in the pre-experiment. Original sound is analyzed according Mel decomposition. Then, mel coefficient evolutions are used to convolve with white noise. It results an audio stream with the same energy distribution as the original sound but without timbre.

The experiment accounted for 21 non-expert candidates that have rated on a scale from 1 to 5 their confidence in identifying the action that have produced the sounds. Eleven candidates were assigned to the non-transformed sound corpus and the other ten to the transformed one.

The results can be seen on figure 3. The figure shows the statistics for each sound. Plot on the left corresponds to the original sounds. Plot on the right corresponds to the transformed sounds. Black solid horizontal lines are the median rate and boxes are illustrating the deviation between the first and the third quartile. Black dashed lines are reporting min and max values.

The results show that for some sounds of the transformed corpus the candidates are still confident in their identification of the sound cause (e.g. sounds 1 and 2). However, other sounds are efficiently non-identifiable (e.g. sounds 8 and 9). We select four sounds that represent the best the effect of the audio transformation and having different temporal profiles. The resulting corpus contains 8 sounds corresponding to: NT 4, NT 6, NT 8, NT 9 and T 4, T 6, T 8, T 9 (where NT=non-transformed and T=transformed) corresponding to:

(4) glass impact
(6) pouring rice
(8) screwing a bottle cap
(9) squeezing a can

Median rates for the selected sounds are: 5 (NT4), 2.5 (T4); 5 (NT6), 3 (T6); 2.5 (NT8), 1.5 (T8); 4.5 (NT9), 1 (T9).

5. EXPERIMENTAL PROTOCOL

In this section we present the experimental protocol. Since we have two corpuses, the same protocol is used for each corpus and each candidate participates to the experiment for only one of the two corpuses.

5.1 Method

5.1.1 Task

The task is presented as follows. "*You must perform a gesture associated to the sound you will listen to. Here "associated" means performing gestures that mimic the action producing the sound or that follow temporal evolution of the sound*". Two fixed examples for the different strategies that can be adopted in the performance are illustrated by the examiners. The strategies are explicitly told to the participants to avoid participants to be lost when being faced to such a non-usual experience. The experiment continues with two phases: the performance and the interview.

5.1.2 Phase 1: Performance

Only one of the two corpuses is used per candidate. The participants are asked to perform gestures synchronously to the sound they are listening. For each sound of the corpus, there are three sequential steps: *training*, *selecting*, *validating*. In the first step, the participant can listen to the sound any number of times. Synchronously, any number of rehearsals can be performed in order to find the gesture that is, for the participant, well associated to the sound. When the candidates feel confident, they select the associated gesture (so-called *candidate gesture*). The final step is the validation of the candidate gesture. The participant must perform three times exactly the same gesture. This step validates that the candidate gesture is stabilized. The whole performance phase is recorded by a video camera.

5.1.3 Phase 2: Interview

The interview is an *auto-confrontation* of the participants with their performance [16]. Together with the participants we sequentially visualize the videos corresponding to each sound. Only the candidate gestures are watched (i.e. four videos). For each candidate gesture we ask questions that allow the participants to verbalize their action. First we discuss what came spontaneously to their mind when they first listened to the sound. Then we discuss the gesture they performed (e.g. *was it difficult to find the gesture? what are the different steps in your gesture?* etc.). Finally, we discuss the relationships between the performed gesture and the listened sound (e.g. *did you try to be synchronous?* etc.).

The aim is to help the analysis of the data collected during the experiment. Verbalization given by the participants informs us on their intentions during the performance: for instance if they tried to mimic a specific action or to follow acoustic features; how they can describe the listened sound; if they were comfortable with the interface etc.

5.1.4 Data collection

Participants. Twenty-two non-musician subjects participated to the experience, which took place at Ircam between August and October 2010. In a mixed between-within design, two groups, of 11 subjects each, performed either on the Non-Transformed, or on the Transformed sound corpus stimuli. The experiment took approximately one hour, and the participation was retributed with a nominal fee.

Material. The hand's position was captured by tracking on-hand placed markers with an ARTtrack motion capture system at 100Hz sample rate. No other motion capture interface was used during the experiment. The sound stimuli were monophonic and had 16-bit resolution and a sampling rate of 44.1kHz. A video camera recorded each performance. Motion, audio and video were recorded synchronously at each trial using the real time programming environment Max/MSP.

5.2 Results

5.2.1 Interviews: mimicking and tracing

First, we examine the interviews for participants having listened to the non-transformed corpus. Globally, the participants do not succeed to describe the sounds' characteristics but rather describe the action that has produced the sound. Sound descriptions show that gestures associated to the sounds focus on the action in interaction with an object. While they do not accurately describe the object, they are more consistent on the actions. The terminology used to describe each sound can be synthesized as: sound 1, *to hit* (70%); sound 2, *to pour* (85%); sound 3, no clear terminology *to pull*, *to scrap*, *to push*; sound 4, *to squash* (85%). The gesture associated to the sounds corresponds to the action described. Finally, all participants have imagined manipulating an object while they were performing their gesture.

Second we examine the interviews for participants having listened to the transformed corpus. It appears that the participants have not precisely recognized an action or an object. The cognitive representation associated to the sounds is often metaphorical and with large variations across the candidates. Gestures associated to the sounds are described as representations of the corresponding metaphors. The time evolution of the sound characteristics are often referred in the descriptions. To conclude, the interviews reveal that the metaphor associated to a sound emanates from the sound characteristics.

5.2.2 Performed gesture characterization

We are interested in analyzing the gesture variability for each sound from each corpus: *non-transformed* and *transformed*. We choose in a first step to take into account the velocity, found in a previous study as one of the important gesture parameter. Temporal evolution of sound and sensation of energy in our body are linked by the gestural representation of sound during the experiment. Considering

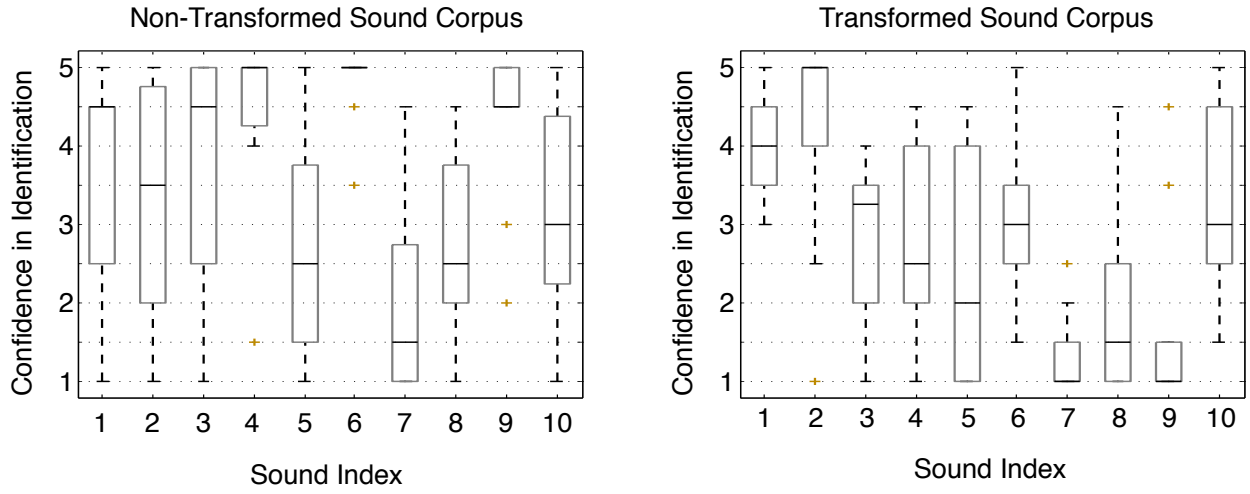

Figure 3: Building corpus. Rates are given for each sound (from sound 1 to sound 10 either for causal sounds or non-causal sounds). Plots depict statistics on resulting rates. Crosses are outliers. Black solid horizontal lines are the median rate and gray boxes are illustrating density between first quartile and the third quartile. Black dashed lines are reporting min and max values.

velocity allow us to be position and direction independent as well as focusing on kinetic energy. A further detailed analysis that consider other gesture parameters (like accelerations or jerks) is left as future work.

Figure 4 illustrates all the performances for each sound from each corpus. Each plot represents from top to bottom: The waveform for the non-transformed sound i; The *candidate* gestures associated to the non-transformed sound i by all the participants: upper bound is the third quartile limit, lower bound is the first quartile limit and the curve is the median evolution; The corresponding transformed sound i; The *candidate* gestures associated to the transformed sound i by all the participants. Gesture variability is computed as the mean and variance of the density range defined as the upper bound minus the lower bound. Results are given in table 1.

	Sound #1	Sound #2	Sound #3	Sound #4
NT	0.920 ±0.285	1.235 ±0.134	0.778 ±0.146	1.144 ±0.256
T	0.591 ±0.180	0.829 ±0.095	0.863 ±0.423	0.605 ±0.111
$\frac{NT-T}{NT}$ (%)	*-35.8*	*-32.9*	*+11.0*	*-47.2*

Table 1: Global cumulative variance

One can see that the gestures performed while listening to the transformed sounds 1, 2 and 4 are less varying than the ones associated to non-transformed sounds. The means are significantly distinct (according to a t-test with α level set to .01). However, there is no significant difference in variability between gestures associated to non-transformed and transformed sound 3 (that is *screwing a bottle cap*). Actually, sound 3 (referring to sound 8 in figure 3) was the less contrasted from the set of selected transformed sounds: the median of confidence rate for non-transformed sound 3 was 2.5 while the median of confidence rate for its transformed version was 1.5. A greater gesture variability for causal sounds than non-causal sounds could be interpreted as follows. When participants identify the sound as its cause, each participant has their own manner to represent the cause. Otherwise, when participants identify the sound by its acoustic characteristics, each participant has a common reference to gesturally represent the sound.

6. CONCLUSION

The aim of this study was to better understand the dichotomy that can exist in gestural environmental sound embodiment. We establish a methodology based on two environmental sound corpuses (non-causal and causal sounds) used as stimuli for candidates. They had to associate a gesture for each sound from one corpus and verbalized their action during an interview. Results show that verbal description of the causal sounds are consistent and they comment their gestures as mimicking the cause whereas verbalization for non-causal sounds do not show a particular consensus in the sound identification. Interestingly, quantitative analysis on gesture data shows that gesture variability is lower for non-causal sounds than for causal sounds. A first interpretation is that people are consistent in the identification of action but the gestural representation of action is highly subjective (because some of these actions are commonly used in the everyday life). On the contrary, when the mental image of the sound cause is confused, the reference becomes the sound itself that is common to all the participants.

Prospective works will go further in the analysis of the terminology and gesture analysis as well as comparing gesture data to sound data. Another short-term perspective is the analysis of a second phase (not described in this paper) that consists in gestures performed on concatenation of the sounds taken from the two corpuses of causal and non-causal sounds presented in this paper.

7. ACKNOWLEDGMENTS

We acknowledge partial support from the project Interlude -ANR -08-CORD-010 (French National Research Agency).

8. REFERENCES

[1] J. Ballas. Common factors in the identification of an assortment of brief everyday sounds. *Journal of Experimental Psychology: Human Perception and Performance*, 19(2):250–267, 1993.

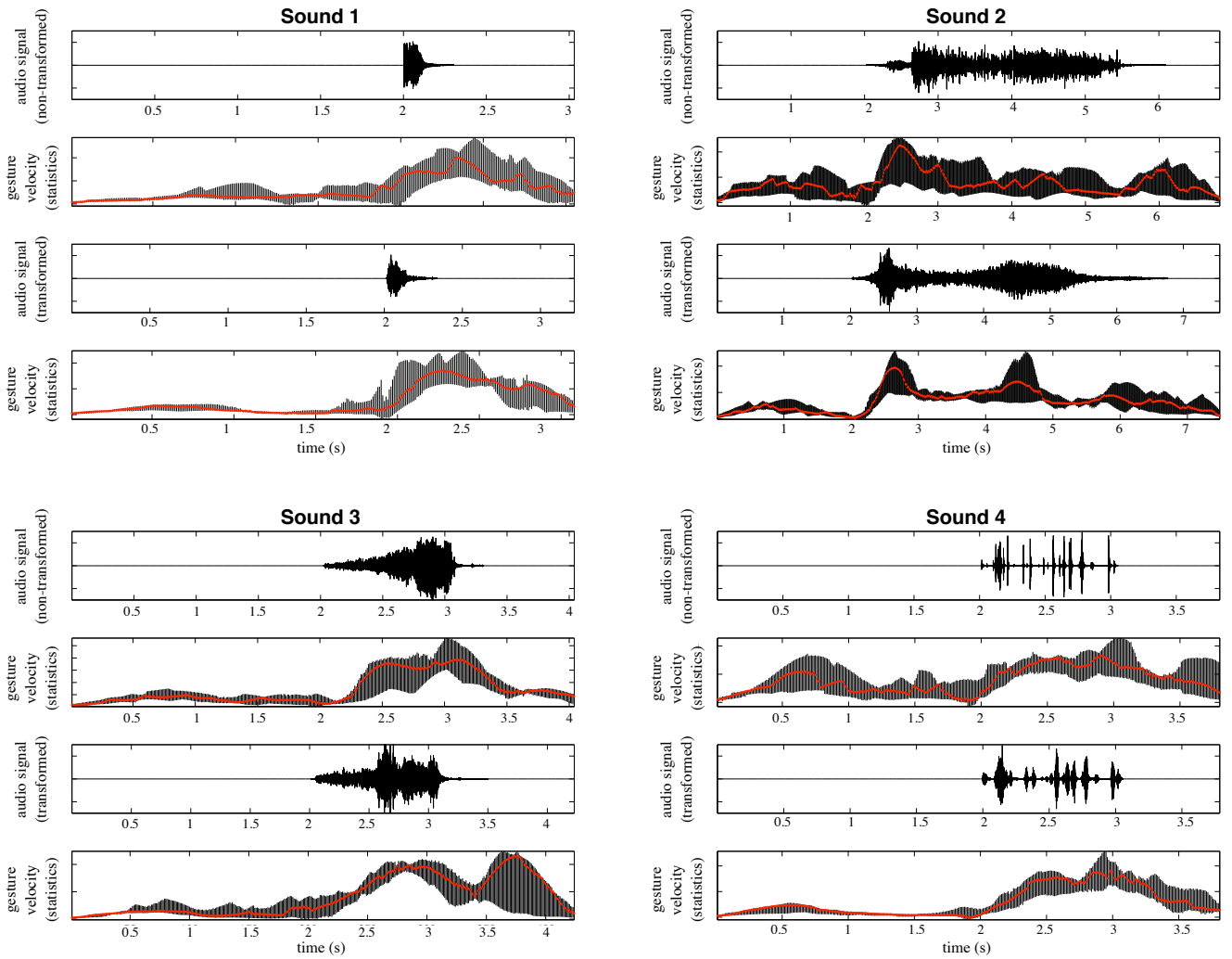

Figure 4: Gestures' velocity associated to each sound from each corpus. Each plot represents from top to bottom: The waveform for the non-transformed sound i; The *candidate* gestures associated to the non-transformed sound i by all the participants: upper bound is the third quartile limit, lower bound is the first quartile limit and the curve is the median evolution; The corresponding transformed sound i; The *candidate* gestures associated to the transformed sound i by all the participants.

[2] C. Cadoz and M. M. Wandereley. Gesture-music. *Trends in Gestural Control of Music*, 2000.

[3] B. Caramiaux, F. Bevilacqua, and N. Schnell. Mimicking sound with gesture as interaction paradigm. Technical report, IRCAM - Centre Pompidou, 2010.

[4] Z. Eitan and R. Granot. How music moves: Musical parameters and listeners' images of motion. *Music perception*, 23(3):221–248, 2006.

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

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

[7] R. Godoy, A. Jensenius, and K. Nymoen. Chunking in music by coarticulation. *Acta Acustica united with Acustica*, 96(4):690–700, 2010.

[8] R. I. Godøy, E. Haga, and A. R. Jensenius. Exploring music-related gestures by sound-tracing: A preliminary study. In *Proceedings of the COST287-ConGAS 2nd International Symposium on Gesture Interfaces for Multimedia Systems (GIMS2006)*, 2006.

[9] R. I. Godøy, E. Haga, and A. R. Jensenius. Playing "air instruments": Mimicry of sound-producing gestures by novices and experts. In *Lecture Notes in Computer Science*. Springer-Verlag, 2006.

[10] A. R. Jensenius, M. Wanderley, R. I. Godøy, and M. Leman. Musical gestures: concepts and methods in research. In *Musical gestures: Sound, Movement, and Meaning*. Rolf Inge Godoy and Marc Leman eds., 2009.

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

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

[13] G. Lemaitre, O. Houix, N. Misdariis, and P. Susini. Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental Psychology: Applied*, 16(1):16–32, 2010.

[14] M. Leman. *Embodied Music Cognition and Mediation Technology*. Massachusetts Institute of Technology Press, Cambridge, USA, 2008.

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

[16] J. Tardieu. *De l'ambiance à l'information sonore dans un espace public*. PhD thesis, Université Pierre et Marie Curie, 2006.

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