The Role of Sound Source Perception in Gestural Sound Description

B. CARAMIAUX, F. BEVILACQUA, T. BIANCO, N. SCHNELL, O. HOUIX and P. SUSINI, UMR IRCAM-CNRS, Paris, France

We investigated gesture description of sound stimuli performed during a listening task. Our hypothesis is that the strategies in gestural responses depend on the level of identification of the sound source, and specifically on the identification of the action causing the sound. To validate our hypothesis, we conducted two experiments. In the first experiment, we built two corpora of sounds. The first corpus contains sounds with identifiable causal actions. The second contains sounds where no causal actions could be identified. These corpora properties were validated through a listening test. In the second experiment, participants performed arm and hand gestures synchronously while listening to sounds taken from these corpora. Afterwards, we conducted interviews asking participants to verbalize their experience, watching their own video recordings. They were questioned on their perception of the listened sounds and on their gestural strategies. We showed that for the sounds where causal action can be identified, participants mainly mimic the action that has produced the sound. In the other case, when no action can be associated to the sound, participants trace contours related to sound acoustic features. We also found that the inter-participants gesture variability is higher for causal sounds compared to non-causal sounds. Variability demonstrates that in the first case, participants have several ways of producing the same action whereas in the second case, the sound features tend to make the gesture responses consistent.

Categories and Subject Descriptors: J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Psychology*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Audio Input/Ouput*; H.5.5 [**Information Interfaces and Presentation**]: User Interfaces—*Auditory (non-speech) feedback*; H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Signal analysis, synthesis, and processing*

General Terms: Experimentation, Human Factors

Additional Key Words and Phrases: Gesture, Sound Perception, Environmental Sound, Embodied Cognition

1. INTRODUCTION

Current digital technologies allow for the creation of a large variety of sonic interactive systems, in a wide range of application domains including sound design, music, pedagogy, gaming and rehabilitation. While wearable sensors and realtime high quality sound synthesis processes are commonplace, linking them might be a hard task. In digital music systems, for instance, the relationship between gesture and sound parameters can be designed *a priori* on the contrary to most acoustic musical instruments, leaving the designer with a large range of possibilities [Wanderley and Depalle 2004]. Recent works have focused on sound perception in the design of these systems [Leman 2007; Maes 2013; Godøy 2006; Caramiaux 2012]. Leman [Leman 2007] along with other authors [Godøy 2006] proposed to consider such systems through an approach based on embodied music cognition that studies the role of the body in sound perception and cognition. Our goal is to bring the cognitive link between gesture and sound (namely the role of sound source perception on gestural response) into the context of the design of novel sonic interactive systems.

Several results have been reported on the interaction between sound and gestures (see Zatorre et al. for a review on the link between auditory and motor systems in music performance and perception [Zatorre et al. 2007]). For example, it has been found that music highly influences timing in motor be-

haviors. In particular, experiments have shown that anticipations and asynchronies occurs in tapping task on beats [Large 2000; Large and Palmer 2002].

Focusing on qualities instead of timing, other studies investigated free body gestures related to sound stimuli [Godøy et al. 2006b; 2006a; Nymoen et al. 2010; Caramiaux et al. 2010b; Nymoen et al. 2011]. Godøy et al. [Godøy et al. 2006b] studied how piano performance mimicry can give insight on pianists' musical expression. In another study, Godøy et al. [Godøy et al. 2006a] studied the "traces" that participants performed on a 2-dimensional surfaces in response to sound stimuli. The results showed highly varying strategies and highlighted the importance and difficulty of refining the working set of sound stimuli. Nymoen et al. [Nymoen et al. 2010] focused on abstract synthesized sounds (controlled by the evolution of their acoustic features: loudness, brightness and pitch) and investigated the relationships between sounds and hand gestures. Alternatively, such behavioral approaches can use computational methods for cross-modal gesture–sound analysis. Caramiaux et al. [Caramiaux et al. 2010b] analyzed cross-modal relationships between kinematic gesture and sound features using Canonical Correlation Analysis (CCA). They found consistent control strategies for a restricted set of short and abstract sounds. Other experiments on gesture responses to sound stimuli also showed two intrinsic gesture-sound relationships. First, sound energy or spectral features were correlated with movement's velocity [Caramiaux et al. 2010b]. Second, pitch was correlated with the hand's vertical position [Nymoen et al. 2011]. A recent study also showed how analysis methods can complement themselves in such experimental contexts [Nymoen et al. 2013].

The work presented here follows a similar methodology of behavioral response to sound stimuli. We propose to study free-hand gestures performed during a listening task. Our general goal is to examine how the performed gestures are influenced by the sound perception and, particularly, we focus here on the degree of identification, by the listener, of the actions and materials causing the sound, which is called *sound causality*.

Sound causality has been specifically studied for environmental sound perception. VanDerveer pointed out that environmental sounds are defined and perceived by the identification of their sources: the events having caused the sound [VanDerveer 1980]. This implied that these sources must be identifiable during the listening process. Gaver proposed to differentiate between two types of listening [Gaver 1993a; 1993b] defined as: *musical listening*, focused on the sound qualities of the acoustic signal; and *everyday listening*, focused on the event causing the sound. Gaver insisted on the experiential nature of listening: "the distinction between everyday listening and musical listening is between experiences, not sounds" ([Gaver 1993b], p. 1).

The understanding of the processes that lead a listener to identify the sound sources (either environmental and caused by an action or music) based on acoustic properties remains a challenge. The common methodology relies on asking listeners to sort sounds between categories based on perceptual characteristics, and eventually labeling each category (so-called classification task) [McAdams 1993]. Thus, during a classification task, the different categories formed lead to a understanding of which types of similarity are used and at which level the sounds are identified. Early works on sound events classification showed that such categories mostly reflect the action or the object that caused the sound [VanDerveer 1980]. Later studies showed that during a free sorting task, the event causing the sound (action/interaction, type of excitation, source) was used more often than acoustic qualities of the sound, as a criterion for categorization [Marcell et al. 2000]. Further to this, the event also overrided context and location [Gygi et al. 2007], and overrided any shared physical properties when participants focused on the mental image generated by each stimulus [Scavone et al. 2002]. When sounds were more complex like sequences of environmental sounds, the categorization was more closely related to global judgment like the absence or presence of human activity [Guastavino 2007]. The primary category was composed of traffic noise, and the latter one of subcategories of sounds corresponding to interactions

with the environment through socialized activities. This result was also found at the level of sound events, thereby nonverbal sounds of living things and sounds produced by physical events such as tools, liquids, and dropped objects were processed in different ways by the brain [Lewis 2004]. Classification tasks have shown a distinction between categories of living and non living sounds [Giordano et al. 2010].

However, little attention was paid to the different perception mechanisms directed either to the event causing a sound or a sound's acoustic qualities. Among few studies, Gerard et al. [Gérard 2004] explicitly studied such a difference. He asked one group of listeners to sort together sounds "which they may hear together in the environment" while another group was asked to sort sounds "on the basis of their acoustical characteristics, independently of their meaning". Recently, Lemaitre et al. [Lemaitre et al. 2010] showed that the categorization based on either the action/object or acoustic qualities depends on the sound identification and listener's expertise. This means that when a sound source can be identified, the sound is categorized based on its related action/object. Conversely, when a sound source cannot be identified, the sound is categorized using acoustic properties. Also, experts from sound related fields use acoustic features more frequently for categorization compared to non-experts.

Based on the review of the state of the art, it appears that there is a clear lack of studies bridging sound perception and the aforementioned approaches based on sound embodiment. Our work aims at filling, at least partially, such a gap. Our specific goal is to test the following hypothesis; In the case of *causal sounds*, i.e. when the action causing the sound can be identified, a listener will have the tendency to mimic the action. In the case of *non-causal sounds*, i.e. when the action cannot be identified, the listener might draw trajectories linked to the sound morphologies, or in other words, follow acoustic features. In addition we aim at investigating in more details gesture characteristics related to both types of sounds. To validate the hypothesis, two experiments are presented in this paper.

First, we report in experiment 1 (Section 2) on the building and validation of a non-causal sound corpus based on a controlled causal sound corpus. Second, we investigated in experiment 2 (Section 3) the gestural description of the listener, using the corpora created from experiment 1. Participants were asked to perform a gesture synchronously to specific sounds. They were also asked to describe verbally afterwards their experience and perception. The results are discussed in Section 4 and we conclude in Section 5.

2. EXPERIMENT 1: BUILDING CAUSAL AND NON-CAUSAL SOUND CORPORA

The goal of this experiment is to create a set of non-causal sounds based on a controlled causal sound corpora. We start with a sound corpus where the sound causality was previously established. From this corpus, we create a second corpus of non-causal sounds by blurring the sound characteristics that initially allowed the listener to identify the source. A listening task is conducted to confirm the efficiency of such procedure and to select sounds, from the initial and modified corpora, guaranteeing a high contrast in causal uncertainty.

2.1 Stimuli

The initial sound corpus was taken from the study by Lemaitre et al. [Lemaitre et al. 2010]. The authors chose a set of sounds created by actions on objects from a domestic context (such as kitchen objects) to ensure that the sound sources were likely to be known to all listeners. The sounds were classified according to the causal uncertainty (H_{cu}) [Ballas 1993] which has been used to quantify name agreement among several listeners in identifying sounds in terms of action or object. 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). Contrary to the study Soundnet [Ma et al. 2010] that provides useful information as "the ability for a concept to be conveyed by an environmental sound", this H_{cu} measurement gives a fine level of identification.

From the whole set of 101 sounds used in the Lemaitre et al.' study, we retained the 10 best identified sounds (mean 1.92s, std 1.51s) meaning the 10 sounds with the lowest H_{cu}. The source of each sound is described in [Lemaitre et al. 2010]. These sounds form the causal corpus that are supposed to be easily identified in terms of action or object causing them. A description of the sounds are reported in Table I.

Table I. Sound corpora

Id.	Description [Lemaitre et al. 2010]	Max. level(dB)	Duration(s)
1	Cutting bread	61	1.2
2	Closing a cupboard door	72	1.4
3	Opening a drawer with castors	71	1.7
4	Champagne glasses clink	59	1.0
5	Knife removed from his case	66	1
6	Pouring cereal into a bowl	68	5.1
7	Bottle top	58	0.72
8	Removing a cork stopper	63	1.5
9	Crushing a metallic can	65	1.2
10	Crumpling a plastic bag	67	4.4

Corpus of the ten most identified causal sounds in terms of object or action, from [Lemaitre et al. 2010].

Based on the sounds from this corpus, we created a transformed version of each of them according to the following process. The original sound was analyzed in Mel bands [Stevens et al. 1937] and white noise was convoluted by the analyzed bands. The resulting sound had a similar spectral evolution than the original (for instance preserving the energy, the first Mel coefficient) while avoiding high frequencies. As described in previous works on sound identification based on spectral resolution [Gygi et al. 2004; Shafiro 2008], this process affects the recognition of the sound causality [1].

This second corpus should thus provide us with examples of non-causal sounds meaning that they can not be identified in terms of action or object. The listening experiment, described in the following experiments, was set to evaluate how well the second corpus can not be identified and to choose the sounds where the transformation was the most effective.

2.2 Apparatus

To be consistent with study [Lemaitre et al. 2010], we used a very similar apparatus. The sounds were played diotically by a Mac Book Pro workstation with a MOTU firewire 828 sound card. A pair of YAMAHA MSP5 loudspeakers were used. Participants were seated in a double-walled IAC sound-isolation booth. The sounds were played and the study was run using the software Max/MSP (Cycling'74). The ecological adjustment of sound levels was performed with a Cyrrus sound level meter.

We used the same ecological adjustment of sound levels given in [Lemaitre et al. 2010], that is preferable to simple sound normalization. As pointed out by the authors, the sounds were recorded with different techniques, including near field and far field recordings, and the relative sound levels are not coherent. As in [Lemaitre et al. 2010], the ecological adjustment is performed by asking participants to evaluate the level of the sounds according to a reference. All the sounds from both corpora were monophonic and had 16-bit resolution and a sampling rate of 44.1kHz.

[1] The sounds for both corpora are available online: http://baptistecaramiaux.com/blog/corpora/. In case of broken link, please feel free to contact via email the first author.

2.3 Participants

Twenty-one participants (ten males and eleven females) were recruited outside of the institute for this study which took place at Ircam – Centre Pompidou in Paris. The participants were aged between 21 and 34 years old, with an average age of 26 ($STD = 3.7$). All participants reported to be non musicians. None of the participants reported having hearing problems. All participants gave informed consent to participate in the study. The experiment took approximately thirty minutes, and participants were given a nominal fee.

2.4 Procedure

The experiment inspects the effect of the transformation on the identification of sounds based on well identified sounds. As previously mentioned, the identification uncertainty of the sound cause can be measured by the so-called index of causal uncertainty H_{cu}. However, the measuring procedure of H_{cu} has two majors flaws: it is time-consuming and it requires a precise semantic analysis of listeners' verbalizations of the sound cause. As an alternative, Lemaitre et al. [Lemaitre et al. 2010] proposed to measure the confidence in the identification using a scale between 1 and 5, reported in Table II. They show that the resulting measure is correlated to H_{cu} even if both measures do not provide exactly the same information. While the H_{cu} evaluates the identification of causality, the rating measures the confidence in identification of the sound causal action. So, the chosen measure of confidence in identification does not guarantee if a sound is well identified, contrary to the H_{cu} measure, however this measure guarantees that a sound with a low value of confidence would have also a high value of H_{cu} meaning that it is difficult to identify the cause. This is finally the purpose of Experiment 1.

Table II. Scale for rating the participants' confidence in identifying the sound cause

Statement	Scale
"I do not know at all"	**1**
"I am really not sure"	**2**
"I hesitate between several causes"	**3**
"I am almost sure"	**4**
"I perfectly identify the cause of the sound"	**5**

The twenty-one candidates were split into two groups: one of ten and another of eleven. Each group was randomly assigned to one of the corpora. At the beginning, the participants were told that they would have to listen to sounds corresponding to several actions typically found in a kitchen. The participants sat in the sound-isolation booth and read the instructions alone. The participants were asked to rate their confidence in identifying the sound cause in terms of action using a five-level scale (see Table II). Each sound was rated twice (test and retest). In other words, every sound was played twice in a shuffled order. The participants were told that they may listen to the sounds several times. After making sure that the participants had understood the task, the experimenter demonstrated the task with two sound examples in order to familiarize participants with the interface and the type of the sounds. After this, the participants performed the listening test alone. They heard sequentially the twenty sounds (each sound twice from the assigned corpus). Only the assertions reported in Table II were displayed on the screen together with check boxes. After rating one sound, they could validate their choice by pushing a button and switching to the next sound.

2.5 Results

First we investigated the consistency of the participants' answers between the scores from the test (first listening) and the retest (second listening of the same sound).

For the causal sounds (corpus 1), the confidence scores were found to be equal between the test and retest for 64.6% of the sounds on average (STD$= 14.3\%$) and had less than 1 scale level of difference for 94.6% (STD$= 6.9\%$). For the non-causal sound (corpus 2), the scores given by the test and the retest remained the same for 56.0% on average (STD$= 17.1\%$) and varied by less than 1 scale level for 89.0% (STD$= 9.9\%$). In opposite, the percentage of flipped opinions between test and retest, measured as a difference between confidence scores greater than or equal to 3, is 0% for the causal sounds and 4% (std=7.0%). These results showed that the scores remained close between the test and the retest for half of the sounds. This allowed us to compute the mean values (between test and retest) providing us with one confidence score per participant per sound. The following analysis was conducted using these average values.

The confidence scores were submitted to a repeated-measures analysis of variance (ANOVA) with one within-subject factor (the ten sounds) and one between-subject factor (the two corpora). The analysis revealed that the participants were significantly less confident in identifying the sound cause of the non-causal corpora than the causal corpora ($F(1, 19) = 14.64, p < 0.01$). The analysis also revealed that the confidence in identifying the sound cause depends on the sound ($F(9, 171) = 8.22, p < 0.01$). Finally, an interaction existed between the conditions non-causal and causal and the sound considered ($F(9, 171) = 4.34, p < 0.01$). This means that the confidence in identification depends both on the sound and whether it was transformed. Figure 1 reports the mean values for each sound and both transformations. This illustrates that globally, the confidence scores are lower for the transformed sounds (except from sounds 1 and 2).

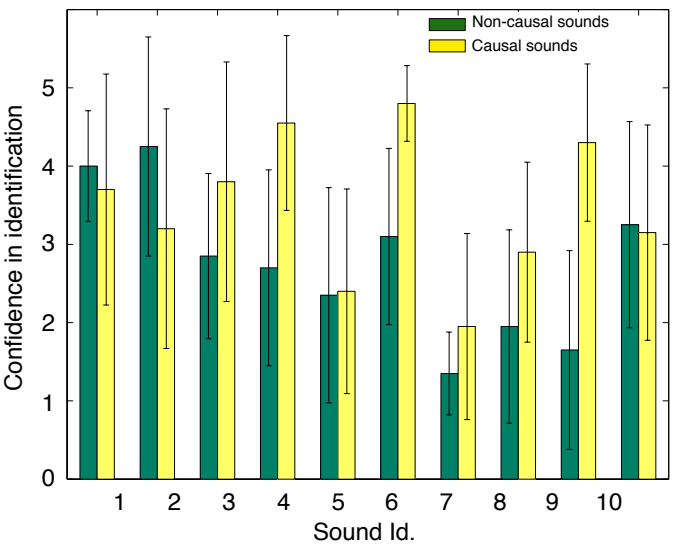

Fig. 1. Experiment 1 scores. Confidence in sound identification in terms of action, as rated by participants. Means are depicted on the figure for both the non-transformed and the transformed sounds.

We then performed a post-hoc analysis by testing, for each sound, if the difference in participants' confidence mean scores were significant between action and non-causal conditions. To compare the means, we applied a Student's t-test with a Boneferroni correction leading to a global significance level of $\alpha = 0.005$. The analysis showed that the mean scores were significantly different for three sounds,

whose index are 4, 6 and 9. Table III reports the three sounds together with their descriptions and their t-value between both versions of the sound.

Table III. Sounds selected after analysis.

Id.	Sound description	t value ($df = 19$)
4	Champagne glasses clink	2.88
6	Pouring cereal into a bowl	3.49
9	Crushing a metallic can	4.89

Three sounds are significantly discriminated ($\alpha = 0.005$) after applying the audio transformation.

2.6 Discussion

The first experiment allows us to build one subset from each corpus considered. Each subset is made of three sounds indexed as 4, 6 and 9. The first subset contains the three original sounds, the second subset contains the transformed versions of these three sounds. Only three sounds out of ten have been discriminated from the listening test, which might seem low. Nevertheless, this is explained by the fact that both corpora have been evaluated independently by their respective group of participants. Hence, each set of sounds was rated according to a confidence scale related to the intra-group and non-inter-group differences. These differences may have increased the contrast, but also changed the tested hypothesis, since the confidence in identification of the non-causal sounds would have been tested relative to the causal sounds.

Regarding their temporal profiles, sound 4 is an impact, sound 6 has a continuous profile without impact and the last is an irregular sound with several articulated impacts. Profiles of the original sounds can be seen in figure 2.

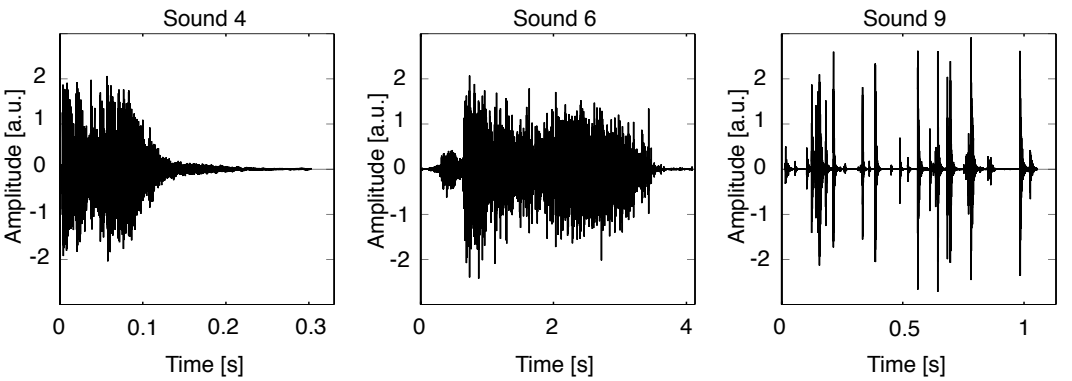

Fig. 2. Waveforms of sounds 4, 6 and 9

3. EXPERIMENT 2: GESTURE RESPONSES

The goal of this second experiment is to investigate the differences between gestural description in response to the causal and non-causal sounds that were validated in the listening test of experiment 1. Our hypothesis is that: First, causal sounds should lead to the gesture mimicking the action causing the sound. Second, non-causal sounds lead to the gesture tracing the sound *spectromorphology* [Smalley 1997]. The term spectromorphology is used to characterize how the sonic content is shaped over

time. Note that such a concept of shape, or morphology can be linked to the concept of *sound object* defined by Schaeffer [Schaeffer 1966], and its link with gestural description is consistent with recent results as mentioned in Section 1 [Godøy 2006; Godøy et al. 2006a; Nymoen et al. 2011; Caramiaux et al. 2010b].

3.1 Stimuli

The stimuli are the sounds from the two corpora created in experiment 1. The first corpus contains the original causal sounds that were selected (sounds 4, 6 and 9). The second corpus contains the transformed versions. As before, the first corpus will be called causal corpus and the second one non-causal corpus.

To compare the implication of causality on gestural responses of sound stimuli, we chose to add an additional sound that was not discriminated by the audio transformation presented in study 1. This sound is added as a "control" sound. Our hypothesis is that the sound does not induce a specific strategy in action ("mimicking the action" vs "trajectories linked to sound morphologies"). We chose to add sound 8 "Removing a cork stopper", which was among the non-discriminated sounds in experiment 1. Globally, the mean sound duration is 2.2s (std=1.9s). The sound descriptions are presented in Table IV.

Table IV. Set of sounds used in the experiment

Id.	Sound description	Duration(s)
4	Champagne glasses clink	1.0
6	Pouring cereal into a bowl	5.1
9	Crushing a metallic can	1.2
8	Removing a cork stopper	1.5

Original sounds corpora used as stimulus in experiment 2. A second corpora contains the transformed versions of the sounds.

3.2 Apparatus

The hand position was captured using an ARTtrack 3D video infra-red motion capture system at 100Hz sample rate. Markers were placed on a glove. As in experiment 1, a pair of YAMAHA MSP5 speakers were used. A video camera recorded each performance. Motion, audio and video were recorded synchronously at each trial using the software Max/MSP (Cycling'74). Participants were standing in a studio in front of a screen displaying a visual countdown that indicated the sounds' start. Two MIDI pedals and a MIDI controller (Berhinger BCF2000) were used for triggering the sounds, and advance throughout the different experiment steps (see details in the next section). The Figure 3 summarizes the whole experiment set up. Finally, the same procedure as described in experiment 1 was used for ecological sound level adjustment.

3.3 Participants

Twenty-two participants (11 women and 11 men) were recruited outside of the institute. The participants were aged between 20 and 50 years old with the mean of 29.9 (STD=7.8). The study took place at Ircam – Centre Pompidou in Paris. None of the participants participated in experiment 1. All participants used their dominant hand (19 participants were right-handed and 3 were left-handed). All participants were non-musician. None of participants reported having hearing problems. All participants gave informed consent to participate in the study. The experiment took approximately one hour, and the participation was retributed with a nominal fee.

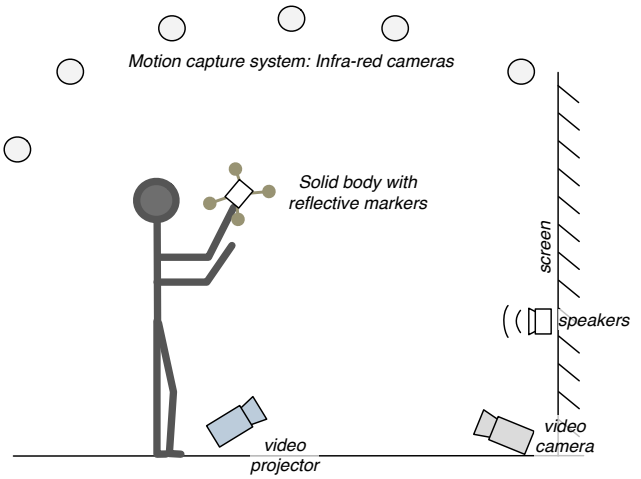

Fig. 3. Schematic of the set-up used in the experiment.

3.4 Procedure

The experiment was designed to examine the influence of the two corpora (between-participants factor) and each sound from a corpora (within-participant factor) on the participants' gestures. The experiment was divided into two phases.

During the first phase, participants were asked to synchronously perform a gesture to a sound they listened to. They were told that they had to perform gestures that were related to the listened sound, either linked to its production or its qualities. Two examples for these two distinct strategies were given by the experimenter ("Striking and igniting a match" and its transformation), nonetheless no example of gestural response was told to the participants. Moreover, participants were not told that the sounds correspond to actions typically found in a kitchen.

The participant stood up in front of a large-scale screen and two MIDI pedals. Each MIDI pedal was used to trigger the current sound. Pushing down a MIDI pedal started a 2-second countdown displayed on the screen. The participants were told that the sound would start when the countdown reached 0. Pushing down a MIDI pedal also created a new track used to store the multimodal recording including audio, gesture data and video. The recording stops 0.7 seconds after the end of the sound.

For each sound of the corpus, there were three sequential steps: *training*, *selecting*, *validating*. The left MIDI pedal was used for the training and validating steps, the right MIDI pedal for selecting. In the training step, the participant could listen to the sound any number of times by pushing the left pedal. Synchronously, any number of rehearsals could be performed in order to find the gesture that was, according to the participant, fitting well to the sound. Note that all the trials were recorded. When the candidates felt confident in having found the gesture they considered adequate, they selected this gesture by using the right pedal that involved the same procedure. The selected gesture is called *candidate gesture*. The final step was the validation of the candidate gesture by performing the same gesture three times. This final step was used to ensure that the candidate gestures were stabilized. The validation trials were also recorded. For each participant, the sound sequence was randomized.

During the second phase, participants sat on a chair in front of the screen. Each participant watched the video of their own candidate gesture, sound after sound, and they were asked questions to stimulate the participants to comment on their own gestures. Overall, a total of four videos were visualized. The

interview was conducted following the interviewing techniques defined by Vermersch et al. [Vermersch 1990] and also used by Tardieu et al. [Tardieu et al. 2009]. This auto-confrontation setting facilitates the description of the sounds and the participant's gesture by an open dialogue. For each video a dialogue was opened according to two pre-defined topics. First, participants were asked to recall what came spontaneously to their mind when they first listened to the sound. Then, questions were asked on the gesture they performed, for example "Was it difficult to find the gesture?", "What are the different steps in your gesture?". The interviews were recorded externally to be further analyzed as explained in the next section 3.5. Importantly, the technique relies on restarts that correspond to keywords that help the participant when she encounters difficulties in the verbalized description.

3.5 Data Analysis

Regarding the interviews, a general grid for the analysis was designed to be filled by the verbal description (verbs, nouns, adjectives) given by the participants. It consisted of two main categories: *causal level* and *acoustic level*. The causal level referred to the sound cause and the acoustic level referred to the sound quality. Following [Lemaitre et al. 2010; Houix et al. 2012], both levels were divided into sub-categories as follows:

—causal level: *action* and *object* (that produced the sound)

—acoustic level: *temporal description*, *profile* and *timbre*

The interviews were analyzed by three independent experts that filled the grid of analysis while listening to the interviews; two of the experts are co-authors of this paper. On purpose, the last expert was not aware of the goal of the study. The three experts have listened independently to all the recorded interviews. The resulting analysis was reported in a general table for both corpora leading to a *portrait* of each sound.

Regarding the gesture data, each gesture trial corresponds to a 3-dimensional time series formed by the positions along the three axes (x, y, z). We computed the norm of the velocity profiles using functional data analysis techniques [Ramsay and Silverman 1997]. As shown in previous similar studies, velocity is one of the relevant information gestures to consider [Leman et al. 2009; Caramiaux et al. 2010a; 2010b]. Velocity profiles are directly related to the kinetic energy that participants engage during their performance.

We used a B-spline basis of order 6 to fit the discrete gesture data. One spline function is placed at each discrete sample. A roughness penalty is applied on the fourth derivative with the penalty coefficient $\lambda = 1e2$. This allowed for the accurate computation of smooth first and second order derivatives [Ramsay and Silverman 1997]. Note that Loehr et al. [Loehr and Palmer 2007] used a similar analysis of the discrete data of pianists' movements.

Finally, each participant's candidate gesture is manually segmented according to three gesture parts corresponding to the *preparation*, the *stroke* and the *release* [Kendon 2004]. These parts are defined as follows:

—*preparation*: gesture occurring before the stroke, used for preparing the movement such as bringing the arms in a specific position.

—*stroke*: gesture synchronous to the sound (e.g. producing the sound). In this paper, stroke does not solely refer to impact but more generally to gesture part intentionally synchronous to the sound.

—*release*: gesture release after the stroke, occurring after the sound's end, e.g. relaxing the arms on both sides of the body.

Some gestures might not exhibit any release part. For consistency, only preparation and stroke gestures were used in the analysis. In addition, preparation and stroke length may vary across the partic-

ipants. This led us to choose a comparison measure that takes into account non-linear time stretching, as further detailed in the next section

3.6 Results: Analyzing the Interviews

We analyzed the verbalization given by the participants for both corpora according to the causal and acoustic levels defined in the analysis grid. We remind the reader that the sounds are: 4 (Champagne glasses clink), 6 (pouring cereal into a bowl), 9 (crushing a metallic can) and 8 (removing a cork stopper).

3.6.1 *Causal sounds.* First, we analyzed the sound verbalization at a causal level. The results are reported in Table V. The participants used verbs to characterize the action causing the sound. These verbs are consistent for sounds 4, 6 and 9: "to hit" (sound 4); "to pour" (sound 6); "to crush" (sound 9). The aforementioned verbs were used to describe the original sounds from the kitchen reported in Table I. Sound 8 did not show any agreement in the description of the action. The verbs used by the participants were "to force", "to rub", "to pull", or "to scrape", showing a higher variability in the description.

Regarding the object description, sound 4 referred to either one object "glass", "bell", "jar" or two objects hitting "glass–glass", "knife–glass". Sound 6 referred either to several small objects as "grains", "marbles", "peas", "coins" or an interaction between small objects and a static one "stones–floor", "rice–bowl". Sound 9 was described in terms of a single object: a consensus was found for a "metal can". Finally, sound 8 referred to objects from distinct lexical fields. Objects are described as "cork stopper", a "zip", "rack-and-pinion". Interaction between objects are "stick–box with sprockets", "ruler–table".

Table V. Verbalization result at the causal level for the four causal sounds

Sound	CAUSAL LEVEL	
Id.	Action	Object
4	to hit	glass, bell, jar glass–glass, knife–glass
6	to pour	grains, marbles, peas, coins stones–floor, rice–bowl
9	to crush	metal can
8	to force, to rub to pull, to scrape	cork stopper, a zip rack-and-pinion stickbox with sprockets rulertable

The second step is the analysis of the sounds at an acoustic level. The results are reported in Table VI. Temporal description of sounds were coarse: "short" (sound 4); "long" (sound 6); "discontinuous" (sound 9). Sound 8 had no temporal description. The profiles were described as: "changes in loudness" (sound 6); "irregular" (sound 9); "continuous" (sound 8). The timbre was not described by the participants, only a rough description of sound 4's timbre as "high–pitched" was asserted.

Finally, the gestures were described by the action, for example: "I was cheering with a glass in my hand" (sound 4); "I am pouring grains of something" (sound 6); "I am crushing a can" (sound 9). Gesture related to sound 8 is described as "pulling", "rubbing".

3.6.2 *Non-causal sounds.* First, we analyzed the sound verbalization at a causal level. Table VII reports the results. Semantic description of transformed sounds was metaphorical in the sense that no precise actions or objects were emphasized but rather there was a qualitative description of what the sound evoked. Sound 4 was characterized by an object that "is falling", "is hitting". The object itself was

Table VI. Verbalization result at the acoustic level for the four causal sounds

Sound	ACOUSTIC LEVEL		
Id.	Temporal	Profile	Timbre
4	short		high-pitched
6	long	changes in loudness	
9	discontinuous	irregular	
8		continuous	

described differently according to the participants but referred to the metaphor, for example “a box”, “a javelin”, “a knife” or “a ball” (petanque) and could be in interaction: “box–tiled floor”, “hammer–sheet metal”. Sound 6 was described as “waves” or more globally the “sea”, “water” that was going “back and forth”, “up and down”. The lexical field used is related to the sea. Sound 9 showed a higher variability between descriptions. Participants described it as “something” that “is walking”, “is hurtling down”. Several objects were reported like “a horse, book pages, rope [lasso], a whip”. Finally, sound 8 was described as either “a geyser, a wagon, a cupboard” or “a flame” that accomplished the action of “being sucked up, being teared up”.

Table VII. Verbalization result at the causal level for the four non-causal sounds

Sound	CAUSAL LEVEL	
Id.	Action	Object
4	Is falling, hitting, tumbling, beating	box, javelin, knife, ball box–tiled floor, hammer–sheet metal
6	Is going back and forth, going up and down	sea, waves, water
9	Is walking, hurtling down	horse, book pages, rope, whip
8	Is disappearing, sucking up	wagon, geyser, flame, cupboard

The acoustic level description showed more detailed features compared to non-transformed sounds. Table VIII reports the results. We examined first the temporal description. Sound 4 was “brief”. Sound 6 was decomposed into three distinct parts: the beginning (“louder, sudden, brutal”), the middle (“jolt”), the end (“slower, descrescendo, quiet”). Sound 9’s temporal evolution was qualified as “rhythmic, jerky”. Sound 8 was “linear” and decomposed into two distinct parts: the beginning (“open, intense”), the end (“abrupt, is closing”). The second sub-category was the sound profile. Sound 4 was “brutal”. Sound 6 had a profile that “goes up and goes down”, with “peaks of intensity” and “irregular, recurrent”. Sound 9 had an “irregular” profile. Finally, sound 8 had a “crescendo/descrescendo” profile (i.e. trajectories of the loudness). The timbre was more verbalized for sounds 9 and 8 than sound 4. Sound 4 timbre was “high–pitched” (same description as in the non-transformed case). Sound 9 had a timbre that was “wide” and “diffuse” while sound 8 had a timbre that was a “switching between high-pitched and low-pitched”.

Finally, gestures were described as follows. The gesture associated to sound 4 is “brief” and “precise” and “acts the impact on an object” or “by an object”. The gesture associated to the sound 6 is “waving”, is “mimicking this object” (in this case the wave) and is describing as “drawing the sound” profile. The gesture associated to sound 9 is “repetitive” and “follows the sound”. Finally, the gesture associated to sound 8 is also described as “mimicking the object” and “temporally following the sound”.

Table VIII. Verbalization result at the acoustic level for the four non-causal sounds

Sound	ACOUSTIC LEVEL		
Id.	Temporal	Profile	Timbre
4	brief	brutal	high-pitched
6	beginning: louder, sudden, brutal middle: jolt end: slower, decrescendo, quiet	goes up and goes down peaks of intensity irregular, recurrent	
9	Rhythmic	Irregular	wide and diffuse
8	Linear Beginning: open, intense End: abrupt, is closing	crescendo/decrescendo	switching between high-pitched and low-pitched

3.7 Results: Variability of velocity trajectories

We remind the reader that each gesture considered is represented as a 1-dimensional time series representing its velocity profile. The rationale behind this choice is based on previous work which highlights the importance of velocity to characterize free hand gestures [Leman et al. 2009; Caramiaux et al. 2010a; 2010b]).

Let us examine the variability of the velocity trajectories across the participants. Figure 4 illustrates all the performances related to each sound from each corpus. Each plot represents from top to bottom: the waveform for the causal sound; the median of all candidate gesture velocity profiles associated to the causal sound (upper bound being the third quartile limit and lower bound being the first quartile limit); the waveform for non-causal sound; the median of all candidate gestures associated with the non-causal sound.

We analyzed the effect of the sound and its transformation on the variability of gesture velocity profile across participants. Variability was measured through Dynamic Time Warping (DTW) between gesture velocity signals. DTW measured the cost to time stretch one given signal to another of different length. This cost also depended on the amplitude difference between signals. Hence for each pair of gesture signals we had a real value between 0 and ∞ that illustrated how close these two signals were (see appendix A for the technical description of the method). Since one velocity profile represented a participant's candidate gesture, the testing procedure relied on computing the DTW between pairs of velocity profiles.

3.7.1 *Preparation gesture.* First, we examined the preparation gestures. A repeated-measure ANOVA with one within-subject factor (the four sounds) and one between-subject factor (the two corpora) was computed based on the warping distances between participants' gestures. The results showed no significant effect of the corpus (in other words, the audio transformation performed on sounds). However, there was a significant effect of the sound considered ($F(3, 324) = 12.88, p < 0.01$). Figure 5 (left) illustrates the mean distance values.

3.7.2 *Stroke gestures.* The warping distance values were submitted to a repeated-measure ANOVA with one within-subject factor (the four sounds) and one between-subject factor (the two corpora). The analysis revealed that the corpus affected significantly the warping distances ($F(1, 108) = 36.67, p < 0.01$) as well as the sounds ($F(3, 324) = 38.95, p < 0.01$). There was interaction between the treatment and the sounds ($F(3, 324) = 13.82, p < 0.01$) meaning that the effect of the corpus (transformation of the sounds) on distance values were not equivalent over the sounds. We examined these differences in more details using a Student's t-test with a Boneferroni correction: the significance level was set to $\alpha = 0.01$. The analysis revealed that the sound transformation affects significantly the gesture variability for sounds 4, 6 and 9 but not for sound 8. As illustrated in Figure 5 (right), the variability was lower for

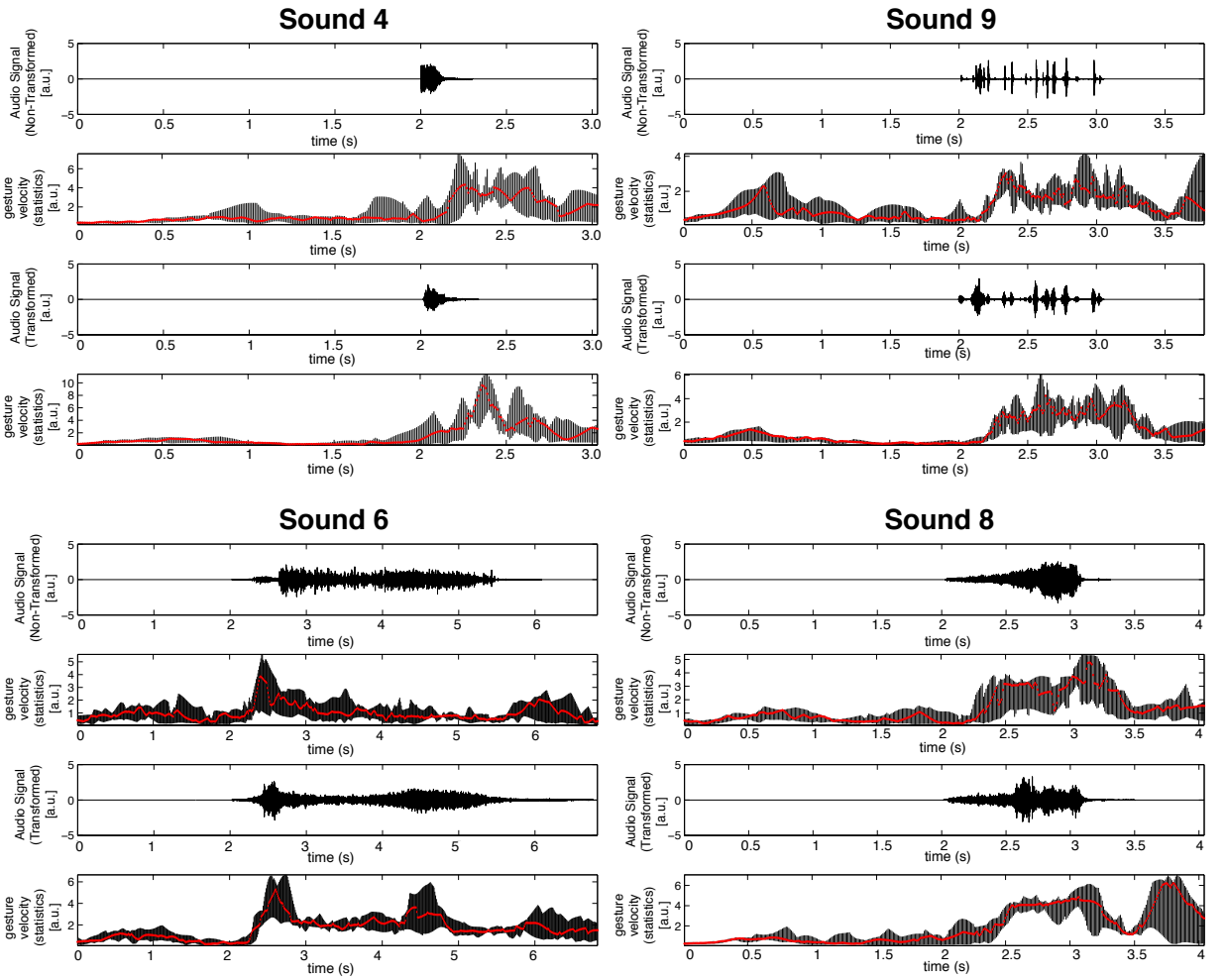

Fig. 4. Gesture velocity associated to each sound from each corpus. Each plot represents from top to bottom: 1) The waveform for the causal sound. 2) The candidate gestures associated to the causal sound: the red curve is the median across all participants, the grey upper bound is the third quartile limit, the grey lower bound is the first quartile limit. 3) The waveform of the non-causal sound. 4) The candidate gestures associated to the non-causal sound: the red curve is the median across all participants, the grey upper bound is the third quartile limit, the grey lower bound is the first quartile limit.

non-causal sounds than for causal sounds in the case of sounds 4 ($0.3 < 2.0$ a.u.), 6 ($4.0 < 6.1$ a.u.), and 9 ($3.0 < 8.3$ a.u.).

4. DISCUSSION

4.1 Qualitative verbalization

For the causal sounds 4, 6, and 9, the participant verbalizations of the action were consistent. They all used verbs to describe the sound in terms of action. In comparison, the description of the object was less consistent. As expected, no action description was found for the “control” sound 8, which was consistent with the listening test presented in section 2. Moreover, for sound 8, the listening test

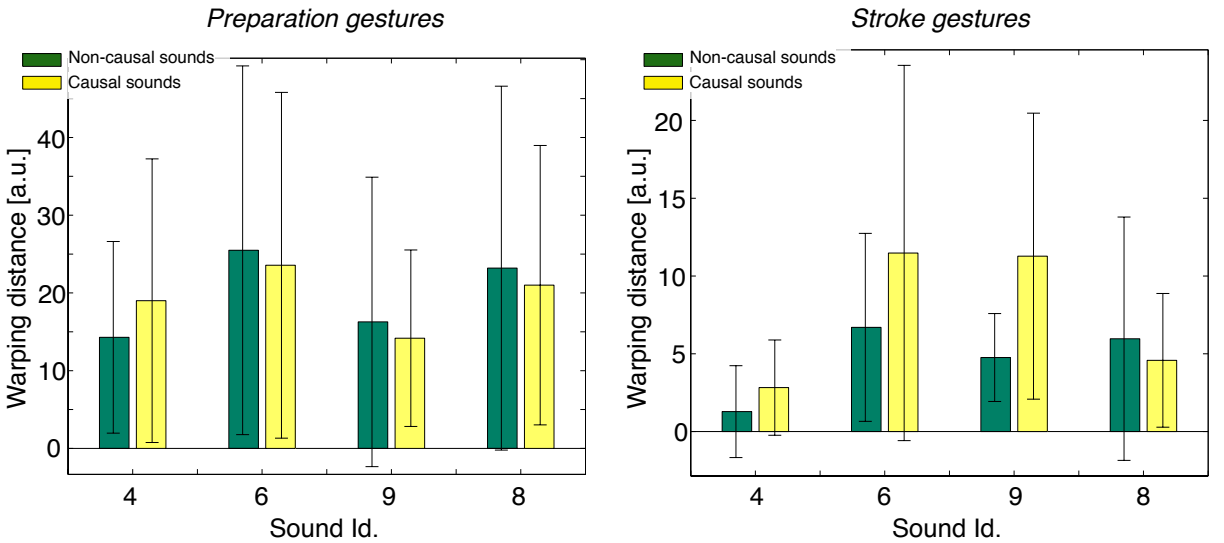

Fig. 5. Mean results of the warping distances for both the preparation gestures (left) and the stroke gestures (right). Means are reported for each sound (columns) and each corpus (two colors).

showed no significant difference between the mean confidence scores in identification between the original and the transformed sound.

In addition, acoustical description of causal sounds was sporadic and poorly detailed. Causal sounds were described sometimes with acoustical descriptions as global properties rather than local acoustical descriptions. This result is consistent with results found by Lemaitre et al. [Lemaitre et al. 2010] where they observe that non-musicians focused more on sound identification in terms of action or object rather than on acoustic qualities (also coherent with [Marcell et al. 2000]). Using Gaver's terminology, we can argue that participants made use of an *everyday listening* strategy [Gaver 1993b; 1993a] since the sounds were easily identified.

Importantly, the gestures were also described in terms of action, and using the same terminology that was used to describe causal sounds. This showed that, for this type of causal sound, there was a direct relationship between the verbalization of the gestures and of the sounds.

Non-causal sounds were described using acoustic description, characterizing temporal evolution and morphologies, rather specifying the any actual sound cause. Precisely, the causal description was subjective metaphorical descriptions of sounds in terms of action and object. Action was not described by infinitive verbs, such as "to do" something, but rather indicating something that "is doing" the action. The emphasis was translated from the action itself to the object or the event corresponding to such an action. This also could be linked to the difference in perception between action sounds (ex. clapping) and non action sounds (ex. water boiling) involving different processes in the brain [Pizzamiglio et al. 2005].

In addition, terminology used to describe the object was highly variable with changes in the lexical field used across all the metaphorical descriptions given by the participants. These metaphors could be generally related to the sound's acoustic qualities. For example the metaphor of the wave for the transformed sound 6 refers to the continuous and oscillating profile of the sound's loudness. These acoustic qualities were more detailed than for causal sounds. They verbalized the temporal descrip-

tion, highlighting different parts in sounds (e.g. 6 and 8), and in the profiles. We argue therefore that participants made use of another listening strategy, referred by Gaver as the *musical listening* strategy [Gaver 1993b; 1993a].

Importantly, gestures associated to non-causal sounds were verbalized as related to sound tracing and drawing, describing profiles of acoustic parameters, using descriptions such as “following”, “drawing” the sound.

These results confirm our hypothesis. If a sound is perceived as being caused by an identifiable action, the associated gesture mimics this action. On the other hand, if the sound is perceived through its acoustic qualities, and no causal action can be identified, the associated gesture follows its acoustic features.

4.2 Quantitative gesture data analysis

The results from the analysis of inter-participants’ gesture velocity variability showed that strokes were more consistent across participants when associated with non-causal sounds rather than causal sounds. Indeed the variability in terms of warping distance was lower after treatment for sounds $4, 6$ and 9. Sound 8 did not induce difference in variability for both measures. This meant that if the audio transformation did not discriminate between causal and non-causal sounds through a listening test, the gesture responses to these sounds could not be discriminated based on gesture consistency measured across participants.

Our interpretation of gesture variability related to the considered corpora is as follows. Participants mimic the action causing the sound when the action can be identified. This leads to variations due to subjective strategies (linked to idiosyncrasy) for executing the action. This type of gesture can be related to the *iconic gestures* postulated by McNeil in [McNeill 1996] defined as the non-verbal elements used in communication to represent an object or a concrete action.

In the other cases, participants mimic the sound object (i.e. follow the acoustic qualities of the sound) leading to a common reference which is the sound itself. The participants relied on the temporal evolution of the sound to perform the gesture. As discussed previously, they are often derived from a metaphorical image of the stimulus. In that sense, they can be related to the *metaphorical gestures* defined by McNeil in [McNeill 1996] as the gestures linked to an abstract idea.

We also inspected whether these results hold for preparation gestures. First, one could observe larger preparation gesture when the causal action was identified, compared to non-causal sounds, meaning that participants prepared the stroke gesture for mimicking causal sound. This highlights two distinct strategies in the movement planning according to the sound played.

4.3 Aspects related to auditory cognition and neurocognition

The present results have shown that verbalizations are more consistent for actions than for objects. In a recent study, verbalizations obtained to describe different classes of environmental sounds were more heterogeneous for causal sounds than for living sounds - sounds which are part of the body of living being [Giordano et al. 2010]. It is likely if living sounds were based on symbolic information which are more consistently coded. In the same way, it could be hypothesized that for causal sounds, actions are more consistently coded than objects or material which is coherent with recent results, in the one hand, by [Lemaitre and Heller 2012] showing that listeners are faster at identifying the action than the material, and on the other hand, by [Houix et al. 2012] showing that listeners use sound information related to physical actions across different objects to sort causal sounds in different sub clusters.

In the present study, causal sounds lead to gestures mimicking the action causing the sound while non causal sounds lead to gestures that follow the sound’s acoustic contours. On the other hand, it

has been shown that neurocognitive processing of environmental sounds can lead to the activation of a motor plan for the generation of a passively heard sound. Indeed, there are neurons in the monkey (F5 area) that discharge both when specific actions ("peanut breaking") are carried out and when the same actions are only heard. However, there is no discharge when sounds are non-causal [Kohler et al. 2002]. Results of the present could be related to a such neurocongnitive process.

4.4 Applications

The experiments reported in this article have been motivated by the need to include embodied cognition aspects in the design of sonic interactive systems. On the one hand, control of causal sounds must rely on gestural input that mimics the action causing the sound. The concrete implementation might involve gesture recognition techniques. A typical sound synthesis engine might be based on physical models. On the other hand, control of non-causal sounds must rely on continuous gestural input that modulates sound parameters (e.g. loudness, brightness, etc.). Typical implementation could be gesture tracking systems whose outputs are mapped to continuous sound synthesis parameters.

5. CONCLUSION

In this paper, we presented experiments to characterize gestural response to environmental sound stimuli. The sound stimuli were either causal sounds, where the action that has produced the sound is clearly identified, or transformed versions of the same sounds where the sound source could no longer be identified (non-causal sounds). We validated the following hypothesis: gesture associated to causal sound mimics the action causing the sound while gesture associated to non-causal sound follows the sound's acoustic contours. Verbalization from participant interviews confirm our finding.

In addition, we showed that gesture data varied less between strokes (gestures linked to the sound) for non-causal sounds than for causal sounds. Preparation gestures showed consistency across participants for transformed sounds but not for the original sounds. We gave the following interpretation: sound causality as action is represented by an iconic gesture that can be performed under distinct forms (depending on the participants' habits in doing the action). The participants have distinct references. On the other hand, when the sound cause is difficult to identify, the participants perform a metaphoric gesture that follows the acoustic energy contour of the sound. In this case, the common reference is the sound itself.

Experiment 1 and 2 have their limitations that encourage follow-up studies in further research. First, the set of sounds is small. We could consider more than 4 sounds by keeping sounds in-between the two extreme cases considered. In this way, one would be able to investigate further the intermediate strategy between mimicking and tracing. Another aspect is the ease of mimic. In the presented experiments, we use kitchen sounds as stimuli in order to limit the effect of level of difficulty in mimicry because these sounds are related to very common everyday actions. Nevertheless, it would be interesting to investigate how the mimicking strategy would be affected by the level of difficulty in mimicking actions related to sound production.

REFERENCES

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

CARAMIAUX, B. 2012. Études sur la relation geste–son en performance musicale. Ph.D. thesis, Université Pierre et Marie Curie (Paris 6).

CARAMIAUX, B., BEVILACQUA, F., AND SCHNELL, N. 2010a. Mimicking sound with gesture as interaction paradigm. Tech. rep., IRCAM - Centre Pompidou.

CARAMIAUX, B., BEVILACQUA, F., AND SCHNELL, N. 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, 158–170.

GAVER, W. 1993a. How do we hear in the world? explorations in ecological acoustics. *Ecological psychology 5,* 4, 285–313.

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

GÉRARD, Y. 2004. Mémoire sémantique et sons de l'environnement. Ph.D. thesis, Université de Bourgogne.

GIORDANO, B., MCDONNELL, J., AND MCADAMS, S. 2010. Hearing living symbols and nonliving icons: Category specificities in the cognitive processing of environmental sounds. *Brain and Cognition 73,* 1, 7–19.

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

GODØY, R. I., HAGA, E., AND JENSENIUS, A. R. 2006a. 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)*.

GODØY, R. I., HAGA, E., AND JENSENIUS, A. R. 2006b. Playing "air instruments": Mimicry of sound-producing gestures by novices and experts. In *Lecture Notes in Computer Science*. Springer-Verlag.

GUASTAVINO, C. 2007. Categorization of environmental sounds. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale 61,* 1, 54.

GYGI, B., KIDD, G., AND WATSON, C. 2004. Spectral-temporal factors in the identification of environmental sounds. *The Journal of the Acoustical Society of America 115*, 1252.

GYGI, B., KIDD, G., AND WATSON, C. 2007. Similarity and categorization of environmental sounds. *Attention, Perception, & Psychophysics 69,* 6, 839–855.

HOUIX, O., LEMAITRE, G., MISDARIIS, N., SUSINI, P., AND URDAPILLETA, I. 2012. A lexical analysis of environmental sound categories. *Journal of Experimental Psychology: Applied 18,* 1, 52–80.

KENDON, A. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

KOHLER, E., KEYSERS, C., UMILTA, M., FOGASSI, L., GALLESE, V., AND RIZZOLATTI, G. 2002. Hearing sounds, understanding actions: action representation in mirror neurons. *Science 297,* 5582, 846.

LARGE, E. 2000. On synchronizing movements to music. *Human Movement Science 19,* 4, 527–566.

LARGE, E. AND PALMER, C. 2002. Perceiving temporal regularity in music. *Cognitive Science 26,* 1, 1–37.

LEMAITRE, G. AND HELLER, L. M. 2012. Auditory perception of material is fragile while action is strikingly robust. *The Journal of the Acoustical Society of America 131*, 1337.

LEMAITRE, G., HOUIX, O., MISDARIIS, N., AND SUSINI, P. 2010. Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental Psychology: Applied 16,* 1, 16–32.

LEMAN, M. 2007. *Embodied Music Cognition and Mediation Technology*. Massachusetts Institute of Technology Press, Cambridge, USA.

LEMAN, M., DESMET, F., STYNS, F., VAN NOORDEN, L., AND MOELANTS, D. 2009. Sharing musical expression through embodied listening: A case study based on chinese guqin music. *Music Perception 26,* 3, 263–278.

LEWIS, J. W. 2004. Human brain regions involved in recognizing environmental sounds. *Cerebral Cortex 14,* 9, 1008–1021.

LOEHR, J. AND PALMER, C. 2007. Cognitive and biomechanical influences in pianists finger tapping. *Experimental brain research 178,* 4, 518–528.

MA, X., FELLBAUM, C., AND COOK, P. R. 2010. Soundnet: investigating a language composed of environmental sounds. In *Proceedings of the 28th international conference on Human factors in computing systems*. ACM, 1945–1954.

MAES, P.-J. 2013. An empirical study of embodied music listening, and its applications in mediation technology. Ph.D. thesis, Ghent University.

MARCELL, M., BORELLA, D., GREENE, M., KERR, E., AND ROGERS, S. 2000. Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology 22,* 6, 830–864.

MCADAMS, S. 1993. Recognition of sound sources and events. *Thinking in Sound: The Cognitive Psychology of Human Audition, Oxford University Press, Oxford 1993*.

MCNEILL, D. 1996. *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

NYMOEN, K., CARAMIAUX, B., KOZAK, M., AND TØRRESEN, J. 2011. Analyzing sound tracings - a multimodal approach to music information retrieval. In *ACM Multimedia – MIRUM 2011 (accepted)*.

NYMOEN, K., GLETTE, K., SKOGSTAD, S., TORRESEN, J., AND JENSENIUS, A. 2010. Searching for cross-individual relationships between sound and movement features using an svm classifier. In *Proceedings, New Interfaces for Musical Expression, NIME 2010 Conference*.

NYMOEN, K., GODØY, R. I., JENSENIUS, A. R., AND TORRESEN, J. 2013. Analyzing correspondence between sound objects and body motion. *ACM Transactions on Applied Perception (TAP) 10,* 2, 9.

PIZZAMIGLIO, L., APRILE, T., SPITONI, G., PITZALIS, S., BATES, E., D'AMICO, S., AND DI RUSSO, F. 2005. Separate neural systems for processing action-or non-action-related sounds. *Neuroimage 24,* 3, 852–861.

RAMSAY, J. AND SILVERMAN, B. 1997. *Functional Data Analysis*. 2nd edition, Springer Science.

SCAVONE, G. P., LAKATOS, S., AND HARBKE, C. R. 2002. The sonic mapper: an interactive program for obtaining similarity ratings with auditory stimuli. In *Proceedings of the International Conference on Auditory Display*.

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

SHAFIRO, V. 2008. Identification of environmental sounds with varying spectral resolution. *Ear and hearing 29,* 3, 401.

SMALLEY, D. 1997. Spectromorphology: explaining sound-shapes. *Organised Sound 2,* 2, 107–126.

STEVENS, S., VOLKMANN, J., AND NEWMAN, E. 1937. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America 8,* 3, 185–190.

TARDIEU, J., SUSINI, P., POISSON, F., KAWAKAMI, H., AND MCADAMS, S. 2009. The design and evaluation of an auditory way-finding system in a train station. *Applied Acoustics 70,* 9, 1183–1193.

VANDERVEER, N. 1980. Ecological acoustics: Human perception of environmental sounds. Ph.D. thesis, ProQuest Information & Learning.

VERMERSCH, P. 1990. Questionner l'action: l'entretien d'explicitation. *Psychologie française 35,* 3, 227–235.

WANDERLEY, M. M. AND DEPALLE, P. 2004. Gestural control of sound synthesis. *Proceedings of the IEEE 92,* 4, 632–644.

ZATORRE, R., CHEN, J., AND PENHUNE, V. 2007. When the brain plays music: auditory–motor interactions in music perception and production. *Nature Reviews Neuroscience 8,* 7, 547–558.

A. DYNAMIC TIME WARPING (DTW)

The DTW performed a temporal alignment between two curves (here we considered two 1-dimensional velocity curves). The DTW is based on the Euclidean distance. Consider two velocity profiles, denoted $\mathbf{v}_1$ and $\mathbf{v}_2$, of same length L. The method first computes a cost matrix $C(v_1, v_2)$ such as:

$$C(v_1, v_2) = (c(v_1, v_2)_{ij})_{1 \leq i,j \leq L}$$

Where

$$c(v_1, v_2)_{ij} = \sqrt{\sum_{i=1}^{L} \sum_{j=1}^{L} (v_{i,1} - v_{j,2})^2}$$

Where $v_{i,1}$ is the i-th element of the mulitdimensional vector $\mathbf{v}_j$. Then a dynamic programming based algorithm finds the optimal path (i.e. minimizing the cumulative cost) in the matrix $C(v_1, v_2)$ leading to the temporal alignment between both velocity profiles. The cumulative sum's end point is the global cost and was used as the measure value.