Journal of Experimental Psychology: Applied
2010, Vol. 16, No. 1, 16–32

1076-898X/10/$12.00 DOI: 10.1037/a0018762

Listener Expertise and Sound Identification Influence the Categorization of Environmental Sounds

Guillaume Lemaitre, Olivier Houix, Nicolas Misdariis, and Patrick Susini
STMS-Ircam-CNRS, Paris, France

The influence of listener's expertise and sound identification on the categorization of environmental sounds is reported in three studies. In Study 1, the causal uncertainty of 96 sounds was measured by counting the different causes described by 29 participants. In Study 2, 15 experts and 15 nonexperts classified a selection of 60 sounds and indicated the similarities they used. In Study 3, 38 participants indicated their confidence in identifying the sounds. Participants reported using either acoustical similarities or similarities of the causes of the sounds. Experts used acoustical similarity more often than nonexperts, who used the similarity of the cause of the sounds. Sounds with a low causal uncertainty were more often grouped together because of the similarities of the cause, whereas sounds with a high causal uncertainty were grouped together more often because of the acoustical similarities. The same conclusions were reached for identification confidence. This measure allowed the sound classification to be predicted, and is a straightforward method to determine the appropriate description of a sound.

Keywords: environmental sounds, categorization, similarity, causal uncertainty, identification confidence

Most recent experimental studies on sound perception premised that listeners are able to focus on several different aspects of sounds. The idea was initially introduced by Gaver (1993a, 1993b), who proposed the distinction between *musical listening* (when the listener focuses on the qualities of the acoustic signal) and *everyday listening* (when the listener identifies the event causing the sound and its properties: type of interaction, material, shape of the objects interacting, etc.), in a widely cited discussion inspired by Gibson's ecological approach to perception (Gibson, 1966).

Illustrations of these different modes of listening are found in different fields of research. The notion of timbre, for instance, investigated in depth over the years, has two main definitions: first, according to its psychoacoustical definition, timbre is an attribute of the sound signal that allows two sounds to be distinguished (i.e., musical listening); second, according to its ecologically inspired definition, timbre is what allows an instrument to be identified (i.e., everyday listening: see Hajda, Kendall, Carterette, & Harshberger, 1997; see also Lemaitre, Susini, Winsberg, Letinturier, & McAdams, 2007, for a review). Sorting tasks are another example where the different modes of listening are of particular importance: just as listeners can focus on different types of information in the sound, they can group together sounds on the basis of different types of similarity. These experimental tasks have been very commonly used in research on sound perception, to study the organization of cognitive representations, the identification of the properties of the sound sources (everyday listening, see the following paragraphs), or as an alternative to similarity judgments (musical listening, see Bonebright, Miner, Goldsmith, & Caudell, 2005). Depending on the purpose of the study, one or the other modes of listening is required. For instance, when using a sorting task to study the timbre of a set of sounds, it is expected that participants will make groups of sounds that are acoustically similar. Experimental instructions are traditionally written in such a way that they orient the participants toward the proper mode of listening. Only by a posteriori screening of the results can the experimenter estimate if the participants have properly followed the instructions. Every experimental study using sorting tasks therefore faces the methodological problem that different strategies are available to the participants, and that it is not clear how to favor one strategy. Moreover, little is known about the psychological mechanisms that underlie these different modes of listening. For instance, very few studies have experimentally addressed the influence of the identifiability of the sounds, and of the expertise of the listeners, both factors that can be fairly suspected of influencing the mode of listening. The influence of these factors on the categorization therefore raises methodological and conceptual problems.

In addition to these basic questions, many applications are concerned with the different types of information that a user is capable of focusing on. Indeed, most audio-processing algorithms (instrument recognition, musical genre recognition, music information retrieval, environmental sound classification, auditory scene recognition, surveillance, etc.) require one to compute similarity measures between sound signals at some stage. One of the

Guillaume Lemaitre, Olivier Houix, Nicolas Misdariis, and Patrick Susini, STMS-Ircam-CNRS, Paris, France.

Guillaume Lemaitre is now at the Department of Psychology at Carnegie Mellon University.

This work was funded by the FP6-NEST-PATH European project no. 29085 CLOSED (Closing the loop of sound evaluation and design). Portions of this work were presented in July 2008 during the Acoustics 08 Conference held in Paris, France.

Correspondence concerning this article should be addressed to Guillaume Lemaitre, Department of Psychology, Carnegie Mellon University, Baker Hall 354Q, 5000 Forbes Ave., Pittsburgh, PA 15213. E-mail: guillaum@andrew.cmu.edu

problems of such measures is that sounds can be similar in many different ways: two sounds can be considered as similar because the *acoustical signals* are similar (acoustical similarity), or because the *sources* that have caused the signal are also similar (similarity of the sources). Two similar sources can produce similar signals, but can as well produce very different signals. The similarity of two sounds therefore depends on which aspect one chooses to focus. Different listeners might pay more attention to one or the other of these aspects (i.e., use different modes of listening), depending on their listening habits, abilities and skills. This problem becomes critical in the case of automatic classifiers. Such algorithms are used for instance in massive sound databases for automatic classification of sounds into predefined categories (i.e., to recognize and label an unknown sound), whether the sounds are musical excerpts, sound effects or environmental sounds (see the definition in next paragraph). It is therefore very important for these applications to translate operationally the notions of different modes of listening into different kinds of similarities.

To provide a first systematic insight into these methodological and applied issues, the studies reported in this article aim at experimentally addressing the different strategies used by listeners to sort a set of environmental sounds. More precisely, the article focuses on two aspects: the influence of sound identifiability and listeners' expertise on the different strategies. The following paragraphs first review the literature on the perception of environmental sounds, experimental classification of sounds, and the difference between expert and nonexpert listeners, before defining the outline of the experimental studies reported in this article.

Perception of Environmental Sounds

Environmental sounds were first defined by Vanderveer (1979, pp. 16–17):

> ". . . any possible audible acoustic event which is caused by motions in the ordinary human environment (. . .). Besides 1) having real events as their sources (. . .) 2) [they] are usually more "complex" than laboratory sinusoids, (. . .) 3) [they] are meaningful, in the sense that they specify events in the environment (. . .) 4). The sounds to be considered are not part of a communication system, or communication sounds, they are taken in their literal rather than signal or symbolic interpretation."

This definition emphasizes that environmental sounds specify the events that have caused the sounds. Vanderveer also investigated how listeners identify and describe the cause of environmental sounds. The results showed that they mostly describe the action, the object of the action, or the place where the action took place. Thereafter, many publications have reported the listeners' ability to recover auditory properties of events causing sound. Some were related to the objects causing the sound: the *length* of wooden rods dropped on the floor (Carello, Anderson, & Kunkler-Peck, 1998), the *thickness* of struck bars made of wood or metal (Lakatos, McAdams, & Caussé, 1997), the *shape* (square, rectangular or circular) and the *materials* of struck hung plates (Kunkler-Peck and Turvey, 2000), or synthesized struck clamped bars (Klatzky, Pai, & Krotkov, 2000), the size and the speed of rolling balls (Houben, Kohlrausch, & Hermes, 2004), the *shape* of a ball dropped on a plate (Grassi, 2005), the categories of *materials* (metal and glass vs. wood and Plexiglas) of recorded struck plates (Giordano & McAdams, 2006). Others were related to the action: discrimination between *bouncing* or *breaking* events (Warren & Verbrugge, 1984) or the ability of blindfolded participants to *fill a vessel* to a normal drinking level or to the brim (Cabe & Pittenger, 2000).

One important question raised by these results is how to identify the acoustic information used (or needed) by the listeners to recover these properties. McAdams, Chaigne, and Roussarie (2004) identified perceptual dimensions correlated with physical parameters of synthesized sounds (physical models) of struck bars with different materials. Another series of experiments again using synthesized sounds of struck bars (Lutfi, 1995; Lutfi & Oh, 1997; Lutfi, 2001; Lutfi, Oh, Storm, & Alexander, 2005) showed that listeners do not optimally use the available acoustic information to decide upon the material or the hollowness of the struck bars. Using recorded sounds, it has also been sometimes difficult to identify a clear correlation between acoustical properties and the perceived event properties (Freed, 1990). For instance, stereotypical relationships between acoustical properties and listeners' responses have been highlighted: slow, loud and low frequency sounds systematically associated with male hand-clappers (Repp, 1987) or walkers (Li, Logan, & Pastore, 1991).

Therefore, it appears that not only the acoustical properties of the sound are responsible for the recognition of the source (i.e., the information present within the sound): the context and the knowledge of the listener may also influence the identification. This question has been explored thoroughly in a series of studies (Howard & Ballas, 1980; Ballas, 1993; Ballas & Howard, 1987; Ballas & Mullins, 1991). The main idea of these authors was that the perception of environmental sounds shares similarities with the perception of language (though the parallels have to be considered carefully, according to Ballas & Mullins, 1991). Howard and Ballas showed that the syntax and the semantics of sound sequences influence their memorization (organized and meaningful sound sequences are better memorized). Ballas and Howard (1987) reported homonym-type sounds: sounds being discriminated, but confused when listeners have to identify their cause. In this case, the context helped listeners to choose among the alternative causes of the sounds (Ballas & Mullins, 1991). Ballas (1993) has reported an imposing series of experiments showing that the identification performance is influenced by several factors, including acoustical variables, ecological frequency (the frequency with which a listener encounters a specific sound event in everyday life), causal uncertainty (measured as the amount of reported alternative causes for a sound) and sound typicality. Actually, acoustical variables accounted for only about half of the variance in identification time and accuracy.

Experimental Classifications of Environmental Sounds

More generally, the question of the identification of objects and events is one of the important phenomena for which the different theories of cognition try to account. Identification can be considered as a cognitive act of categorization. Identifying the source of a sound can therefore be viewed as connecting auditory perception to concepts, and concepts to language (Goldstone & Kersten, 2003; McAdams, 1993): "a concept, roughly speaking, is knowledge about a particular category" (Barsalou, Simmons, Barbey, & Wilson, 2003, p. 84). The cognitive processes of categorization are assumed to be reflected in experimental classifications of items. Experimental studies have therefore used sorting tasks and forced-choice experiments to explore different cognitive models of categorization (Smits, Sereno, & Jongman, 2006). Only a few experimental classifications (i.e., the result of a sorting task) of environmental sounds have, however, been reported. The following paragraphs report the results of five studies. For the sake of comparison, only the experimental procedures and results are reported, without mention of either the goals and assumptions of these studies or the interpretations drawn by the authors.

Vanderveer (1979) reported the results of two free sorting tasks. In each experiment, 20 participants listened to 20 sounds recorded on a tape, wrote down descriptions of the sounds on gummed labels, and then had to "sort the items based on the similarity of sounds" (p. 205). The experimenter provided the participants with some examples of "obvious" groupings (e.g., "filing" and "sawing" sounds). From a rough analysis, she concluded that the basis for sorting the sounds was "twofold: acoustical similarity (or temporal patterning in particular) and relatedness of source events (or meanings)" (p. 214). Indeed, the participants appeared to have sorted the sounds because they were caused by similar events (e.g., "drop pen," "drop can," "drop wood"), or because they shared acoustical similarities (e.g., sounds of "pin box," "sawing," "filing" shared obvious acoustical patterns).

Guyot (1996) reported a classification experiment on 25 "domestic" noises (roughly 3 seconds long). Participants had first to group together noises, "according to their perceptual similarities" (p. 114). Afterward, they had to characterize the categories verbally. Two participants made categories on the basis of the acoustical properties of the sounds and explained the categories by describing the signal properties. One made three categories on the basis of the type of excitation (mechanical, electrical, electronic). The others made categories based on the type of movement creating the sound, on the identified source, and so forth Guyot interpreted the sorting data as the result of two modes (levels) of categorization: either the sounds are grouped together because they are made (or belong to) by the same source, or they are grouped together because they are made by the same movement/interaction (scratching, rubbing, etc.).

Marcell, Borella, Greene, Kerr, and Rogers (2000) aimed at building a normalized corpus of unambiguously identified and named sounds. They asked 37 listeners to sort freely a set of 120 environmental sounds, and to describe the categories they made. "Participants were told that categorization involves placing something with other objects that have similar characteristics and are members of the same group" (p. 853). Then, two independent judges reviewed the descriptions and grouped those judged as equivalent in meaning. They found 23 categories. The categories were very general, describing large categories of sources ("air transportation," "tool"), locations ("kitchen," "bathroom," "nature") or abstract ideas ("hygiene," "sickness," "sleep").

In a study of semantic memory, Gérard (2004) reported two experimental classifications of environmental sounds. In the first one, 30 participants had to group together 24 sounds "which they may hear together in the environment." In the second one, 30 other participants had to group together sounds "on the basis of their acoustical characteristics, independently of their meaning." The results distinguished between the sounds of inanimate objects and animate beings. These categories were divided into thematic subcategories: sounds that happen inside a house, transportation noises, sounds made by animals evoking leisure, and sounds made by farm animals. The categories built from the sorting data from the second experiment grouped together sounds sharing acoustical similarities: same rhythmic, pitch, amplitude patterns.

Guastavino (2007) reported the results of free sorting of soundscapes sequences. Twenty-six participants had to group together and name 16 sequences "according to their perceived similarities." The results showed a clustering of sequences including traffic noise, and sequences including sounds of human activity. The latter categories were subdivided into the different types of activities.

Discussion

Comparing these experimental classifications clearly shows that listeners used different similarities, when they had to sort a set of sounds. They grouped together sounds, because:

• they shared some acoustical similarities (same timbre, same duration, same rhythmic patterns): Vanderveer (1979); Guyot (1996), Guastavino (2007);

• they were made by the same kind of action or interaction, the same type of excitation (electrical, electronic, mechanical), the same source: Vanderveer (1979), Guyot (1996), Marcell et al. (2000); Gérard (2004),

• they occurred in the same place or at the same occasion or activity: Marcell et al. (2000), Gérard (2004), Guastavino (2007).

Therefore, it can be concluded that some strategies used to categorize the sounds were based on the signal similarities, or on the similarities of the physical cause of the sound, and required little interpretation. Some others needed one to identify precisely what made the sound, to infer the situation in which the sounds occurred, who was responsible for the event, for what reason the listener attributed a specific meaning to the sound, and so forth Attributing this meaning thus relied on the knowledge of the listeners, required their interpretation and was influenced by the context.

From the results reported earlier, three *types of similarities* can be defined:

(1) the similarity of acoustical properties: *acoustical similarity*;

(2) the similarity of the identified physical event causing the sound: *causal similarity*;

(3) the similarity of some kind of knowledge, or meaning, associated by the listeners to the identified object or event causing the sound: *semantic similarity.*

These studies did not propose any explanation of the different strategies used by the participants. The differences in the experimental procedures, however, suggest several assumptions. First, the instructions given by the experimenter to the participants were not exactly the same. For instance, the (written) instructions provided by Vanderveer described examples of sounds grouped together both because of the similarity of the physical event and of the sounds themselves ("sawing" and "filing"). Participants also made classes of sounds mainly on the basis of these two criteria. By changing the instruction (first experiment: categories of sounds usually heard together; second experiment: categories based on acoustical similarities), Gérard changed the results of classifications as well. Thus, in these examples, the instructions clearly indicated which kind of similarity the participants had to use. The influence of the methodological procedure has been further demonstrated in a recent study: Aldrich, Hellier, and Edworthy (2009) showed that the methodological features of the grouping task influence the types of similarities used to group the sounds: whereas dissimilarity judgments encouraged participants to use acoustical information, free sorting procedure emphasized categorical information.

The different strategies might also have been influenced by the ability of the listeners to analyze sounds. The participants in Vanderveer's and Guastavino's experiment were university students. Those in Guyot's experiment were members of her laboratory. Marcell et al. worked with psychology college students, and Gérard did not provide any biographical data on the participants. None of these authors seems to have recorded the expertise of the participants with sounds, or tried to assess the influence of the expertise on the individual results.

Expert and Nonexpert Listeners

Only a few studies have reported differences between expert and nonexpert listeners. Studies of timbre perception have often compared musicians and nonmusicians, but have rarely reported systematic differences. Faure, McAdams, and Nosulenko (1996) found no difference between musicians and nonmusicians who had to rate synthetic instrumental sounds on semantic scales. Using the MDS technique, Marozeau, Cheveigné, McAdams, and Winsberg (2003), and Caclin, McAdams, Smith, and Winsberg (2005) found no difference between musicians and nonmusicians who had to judge differences of timbres of natural and synthetic instrumental sounds. Using the same technique, McAdams, Winsberg, Donnadieu, Soete, and Krimphoff (1995) did not find clear differences between nonmusicians, amateur musicians and professional musicians. Their results, however, suggested that musicians make more precise and coherent judgments. This is coherent with the results of von Bismarck (1974), who found that nonmusicians provide less reliable ratings than musicians in rating synthetic sounds on semantic scales. Whereas quantitative neural changes associated with musical training have been reported (Peretz & Zatorre, 2005), Bigand and Poulin-Charronnat (2006) reported a number of cognitive and emotional tasks in which musically trained and nontrained listeners perform similarly: perception of musical tensions and relaxations, anticipation of musical sequences, learning of new compositional systems, content of emotional experience. Preis and Chuzdzicka (2004) found no difference between experts and nonexperts evaluating excerpts of Bach's violin pieces.

When differences between experts and nonexperts have been reported, they are related to tasks that correspond precisely to the skills of the expert listeners. For instance, Solomon (1958) let trained sonarmen rate sonar sounds on semantic scales. The results showed that the attributes reflected in the scales were very specific to sonarmen (e.g., sonarmen used the word "heavy" not to describe a "heavy" sound, but a sonar signal representative of heavy ships). Kendall and Carterette (1991) had participants rating timbre similarities between pairs of sounds made by two instruments playing together (dyads). Whereas nonmusicians tended to rate the timbre similarities of the dyads continuously, musicians tended to cluster dyads according to the instruments they recognized. Köster, Jessen, Khairi, and Eckert (2007) found that nonexperts produced more errors when evaluating the characteristics of voice (breathiness, roughness, tremolo, etc.) than experts in the field.

Concerning expertise and categorization, Tanaka and Taylor (1991) showed, in the framework of Rosch's prototypical approach to categorization (Rosch, 1978), that the basic level of categorization depended on the expertise of the participants. Experts had a more specific knowledge of the objects in their domain of expertise than novices, and these differences of knowledge affected the categorization of the objects.

This review shows that differences between expert and nonexpert listeners occurred when the task required specific knowledge. In the case of free sorting tasks, it can therefore be expected that the expertise of listeners will influence the way they categorize the sounds. Indeed, as noted by Rosch (1978) experts in a field develop specific taxonomies for the objects in the field. It can be therefore expected that sound engineers, for example, have developed in their training listening techniques that allow them to focus on specific aspects of the sounds, label them with a technical vocabulary, and categorize them in a technical taxonomy (typical training for sound engineers consist in "hearing" the different frequencies that compose a sound, and mapping them to different labels, e.g., the frequency bands of an equalizer). These kinds of technical categorizations are very unlikely to be found in lay persons.

Outline of the Study

The review of the literature reported earlier confirms the hypothesis that both participants' expertise and sound identifiability might influence the categorization strategies. To investigate this hypothesis, three studies are designed. To assess the identifiability of the sound sources, the causal uncertainty of a set of sounds is measured in Study 1, on the basis of the procedure developed by Ballas (1993). In Study 2, two groups of participants (expert and nonexperts) have to freely sort selection of these sounds. Both Studies 1 and 2 therefore allow us to assess the influence of sound identifiability and listeners' expertise on the classification of the sounds. In Study 3, a procedure is developed to assess the confidence of the listeners in identifying the sounds, again based on Ballas (1993), in order to provide an alternative to the causal uncertainty measured in Study 1.

The sounds used in the study belong to a domestic context (the usual objects found in a kitchen), to ensure that the sources of the sounds were likely to be known to all the listeners. Similarly, the listeners are made aware that the sounds have all been recorded in a kitchen, to reduce the possible interpretations of the sounds, and to provide the same framework of interpretation to all the listeners.

Furthermore, the context of a kitchen provides a large variety of sounds caused by a number of different events (machines, solid, liquid interactions, gas, electronic alarms, etc.).

Study 1: Causal Uncertainty

Method

Participants. Twenty-nine participants (14 women and 15 men) volunteered as listeners and were paid for their participation. They were aged from 20 to 47 years old (median: 35 years old). All reported normal hearing. None of the participants reported being a professional cook or having a professional activity related to cookery. The participants were all native French speakers.

Stimuli. The sounds were 101 monophonic recordings of activities usually occurring in a kitchen, chosen from different commercial sound libraries : Hollywood Edge Premiere Edition I, II and III, Sound Ideas General Series 6000 and Blue Box Audio Wav. They were selected on the basis of a questionnaire filled in by the members of the laboratory, describing the sounds they usually hear in their kitchen. They had 16-bit resolution and a sampling rate of 44.1kHz. The list of sounds is provided in Appendix A.

Apparatus. The sounds were played by a Macintosh Mac Pro (Mac OS X v10.4 Tiger) workstation with a MOTU firewire 828 sound card. The stimuli were amplified diotically over a pair of YAMAHA MSP5 loudspeakers. Participants were seated in a double-walled IAC sound-isolation booth. The study was run using the PsiExp v3.4 experimentation environment including stimulus control, data recording, and graphical user interface (Smith, 1995). The sounds were played with Cycling"74's Max/MSP version 4.6.

Ecological adjustment of sound levels. The sounds were recorded with different techniques, including near field and far field recordings. Therefore, the relative levels of the sounds cannot be assumed to be coherent: some sources usually at a very low level (e.g., an "ice cube"), played at too high a level may be perceived louder than sources normally louder (e.g., "water flow"), or simply may become unidentifiable, because of an "acoustical zooming" effect situation.

Therefore, the level of the sounds was adjusted to reproduce the "usual" or "ecological" level of the sounds in a kitchen during a preliminary study. Six participants were presented with pairs of sounds (none of these participants was used in the following studies). Each pair was made of the same reference sound followed by the sound to be adjusted. They were required to adjust the level of the second sound to what it would sound like in the kitchen, compared with the first sound. They had to move a cursor changing the level of the second sound. The interface was implemented under PsiExp v3.4 (Smith, 1995). The reference sound was sound # 99 ("filling a sink with water"; see Appendix A), played at 70 dB. The sound was chosen because it is easily recognizable, has a rather long duration (6 s), and is expected to have a high ecological level. The descriptions in Appendix A were provided for each sound to ensure that the participants had identified the sound to be adjusted.

Procedure. At the beginning of the study, the participants were told that they would have to listen to sounds recorded in different kitchens. This information was provided to avoid individual differences owing to a possible association with a different context during the study.

The study was divided into three steps. First, participants were provided with five examples of sounds (sounds # 25, 26, 63, 71, and 100 in Appendix A not appearing in the rest of the studies) to get accustomed to the interface. Second participants heard all the remaining 96 sounds. The order of presentation was randomized for each participant. Third participants heard every sound, typing for each sound the cause of the sound. For this part, participants had only two trials. They were simply asked to indicate the cause (object and action) of each sound by typing a "noun" and a "verb." The "verbs" were indeed supposed to indicate the action, and the "nouns" were supposed to refer to the description of the object. Participants were asked not to employ metaphorical descriptions, not to report their preference, and to make simple descriptions. This procedure is similar to the procedure used by Ballas (1993).

Results

Two participants were excluded from the analyses because they could not complete the task during the maximum time allowed (1 hour).

Analysis of the verbalizations. Overall, the participants used 523 different nouns and 289 different verbs. "Water" was the most cited noun (141 occurrences), and "closing" was the most cited verb. Forty-seven nouns and 44 verbs were cited by more than 14 participants, and 295 nouns and 117 verbs were cited only once.

Before calculating the causal uncertainty for each sound, three persons sorted the 96*27 descriptions, to assess which ones had to be considered as equivalent. They were required to sort the verbalizations for each sound into categories of similar objects and into categories of similar actions. Note that in the procedure adopted by Ballas (1993), the sorters had to sort globally the responses (noun and verb) into categories of similar events. Two of the three sorters were the first two authors, called "expert" sorters A and B, and the other sorter (C) was familiar with analyses of verbalizations, but not involved in the laboratory, and not aware of the goals of the study. The three sorters used the same set of written rules. These rules defined what had to be considered as describing a same object or action (Appendix B). Then, for each sound i the joint proportion p_{ijk} of verbalizations describing the combination of an object j, and an action k, was calculated.

Calculation of the Hcu. Following Ballas (1993), the causal uncertainty of a sound was measured as the entropy contained in the set of descriptions provided by the participants. Because the descriptions were here separated into the descriptions of the "objects" and the "actions" causing the sounds, causal uncertainty had to be defined by the *joint* entropy of the two descriptions (see Legendre and Legendre (1998) for a discussion of joint entropy):

$$Hcu_i = -\sum_{j}^{m}\sum_{k}^{n} p_{ijk} log_2 p_{ijk} \quad (1)$$

where $H\,cu_i$ is the causal uncertainty of a sound i, p_{ijk} is the proportion of verbalizations describing an object j and an action k, n is the number of categories of similar objects, and m is the number of categories of similar actions. A sound had a minimum causal uncertainty when all the participants provided the same description: Hcu = 0. The maximum uncertainty occurred when every participant provided a different description: Hcu = 4.75.

Each sorter needed three days to achieve the sorting task. They found it difficult always to use the same criteria during the coding, especially for "action." The reliability of the three sorters was significant, $r_{A,B}(101) = .84, p < .01, r_{A,C}(101) = .83, p < .01, r_{B,C}(101) = .81, p < .01$, although weaker than the correlations found by Ballas (1993) (.87–.9). Following Ballas, the median uncertainty value for each sound was used in the subsequent analyses.

The Hcu values ranged from 0 to 4.61, with a distribution skewed toward greater values of uncertainty (Figure 1). Among these 96 sounds, 60 were selected, to provide a reasonable number of sounds in the following studies. The selection was made by randomly excluding sounds. Figure 1 represents the distribution of the Hcu values before and after selection.

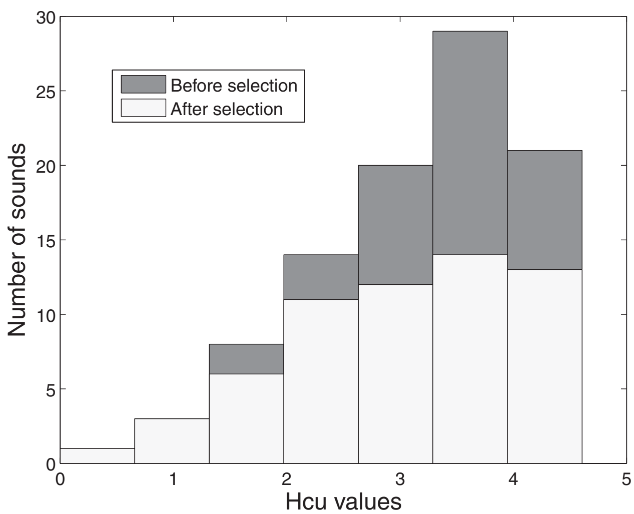

Figure 1. In gray: distribution of the causal uncertainty (Hcu) values for the 101 sounds used in Study 1. In white: distribution of the Hcu values for the subselection of 60 sounds used in Studies 2 and 3.

Study 2: Classification of Environmental Sounds

Method

Participants. Thirty participants (12 women, 18 men) volunteered as listeners and were paid for their participation. They were aged from 19 to 64 years old (median = 32 years old). All reported having normal hearing. They were all native French speakers. None of these participants had previously taken part in Study 1. The participants were preliminarily selected on the basis of questionnaires they had filled in for previous studies, telephone interviews, and a questionnaire about their sound expertise filled before the study. In the light of their answers, they were labeled as “expert” or “nonexpert.” The experts had to fulfill the following requirements:

• Being a professional musician, or having a major musical education,

• Being a professional artist, regularly working with sounds (sound installations, performances, etc.),

• Being a professional or semiprofessional sound engineer or recording engineer,

• Being a scientist working in the fields of sound perception, acoustics, or sound signal processing.

The selection of participants was made to balance the number of expert and nonexpert participants (15 experts, 15 nonexperts).

Stimuli. Sixty sounds were used, selected from the results of Study 1.

Apparatus. The same hardware equipment as in Study 1 was used. The software used to run the study and to implement the graphical interface, however, was Matlab 7.0.4.

Procedure. In the first step of a three-stage procedure, the participants sat alone in a sound-attenuated booth in front of a computer display. They were all given written instructions (in French) explaining the sorting task. They saw a white screen, on which red dots labeled from 1 to 60 were drawn, each dot corresponding to a sound. The labeling was different for each participant. They could hear the sound by double-clicking on a dot. Participants were asked to move the dots to group together the sounds. They were allowed to form as many groups as they wished and to put as many sounds in each group as they desired.

After they had made the categories, they had to type the properties shared by the sounds, which they had used to make this category.

Finally, they were told that three types of similarities had been identified that are usually used to group together sounds. They were given a written description (in French) of these similarities:

> According to you, the sounds within this class:
>
> 1. Are similar because they “sound” the same? For instance, the sounds may be grouped together because:
>
> • They all are low in pitch,
>
> • They all are short rhythmic sounds,
>
> • They all have a very low level, and so forth
>
> 2. Are similar because they all have been caused by the same physical event?
>
> For instance, the sounds may be grouped together because:
>
> • They are all sounds made by impacts,
>
> • They are all made by water drips,
>
> • They are all made by wooden objects,
>
> • They are all made by electrical motors, and so forth
>
> 3. Are similar for a more abstract reason?
>
> For instance, the sounds may be grouped together because:
>
> • They all happen during breakfast,
>
> • They all happen in a restaurant kitchen,
>
> • They are all sounds related to food preparation, and so forth
>
> 4. Are similar for some other reasons?
>
> 5. You do not know why you made this class.

Then the participants were required to indicate, for each category, if they thought they had used one of these similarities. The first three similarities described correspond to the three aforementioned types of similarities (1. Acoustical similarities; 2. Causal similarities; 3. Semantic similarities). In addition, two other answers allowing participants to indicate the irrelevance of the proposed similarities were proposed (4. Other reasons; 5. Don’t know).

Corroboration of the Participants’ Labeling of the Similarities

To assess the reliability of the indications provided by the participants on their strategies to make the groups, this labeling was corroborated by six independent judges: two of the coauthors of the paper, two members of the group of the coauthors, but not aware of the details of the study, and two external sound engineers, specifically hired for the corroboration. The independent judges were provided with the results of every participant. For each participant, they were provided with the same interface, on which the sounds were sorted according to the participant, and for each category, they could read the description of each class provided by the participant. They could listen to all the sounds as many times as they wished. They were also provided with the same instructions as the initial participants. They were required, for each participant and each category, to indicate the strategy they thought best fitted the category and the description. They could not see the participants' labeling nor communicate with the other judges. Each judge had a total of 346 categories to corroborate (185 for the nonexperts, and 161 for the experts). On average, the task took about two hours to complete.

The agreement between the judges was moderate: Fleiss $\kappa = 0.49$ for the corroboration of the experts' categories, and $\kappa = 0.48$ for the nonexperts' categories (Fleiss, 1971). To assess the agreement between the judges and the participants, all the labelings made the participants in each group (experts, nonexperts) were aggregated and compared with the labelings made by independent judges. The Cohen's κ (measuring the agreement between each pair of judges) are reported in Table 1.

These coefficients report a fair agreement between the independent judges and the participants. This agreement was smaller than the agreement between the independent judges themselves. Part of the decrease in agreement can probably be attributed to the procedure of comparing the judges to all the participants altogether (thus incorporating possible disagreements between the participants). No difference, however, was found between the two groups of participants.

To analyze in more details the disagreements between the independent judges and the participants, the contingency tables are reported in Table 2 for the agreement between participants and independent judges.

These tables show where the disagreements between judges and participants occurred. They are graphically represented in Figure 2.

The disagreements between expert listeners and independent judges (bars that are off the main diagonals of Figure 2) were massively localized in the cases where the independent judges reported "causal similarities" whereas participants reported "acoustical similarities." The disagreements between nonexperts and judges seemed to be more spread over all off-diagonal cells of the table, but a closer inspection showed that whereas judges and nonexperts have agreed in 50 cases when reporting acoustical similarities, they have disagreed in 130 cases when the participants have reported acoustical similarities (the judges have then reported causal similarities in 73 cases, semantic in 21 cases, other kind of similarity in 14 cases, and they did not know in 23 cases). Together these results suggest that, both for expert and nonexpert participants, it is difficult for the judges to interpret the classifications and the verbalizations when the participants have made groups of sounds that they found acoustically similar. A possible explanation (suggested during interviews with the judges) is that, because of the lack of technical vocabulary specifically dedicated to the description of sounds in French, most of the descriptions used various kinds of linguistic devices (metaphors, metonymies, etc.),

Table 2
Contingency Tables (Independent Judges vs. Expert Participants, and Independent Judges vs. Nonexpert Participants) for the Labeling of Categories in Study 2

Judges	Acoustical	Causal	Semantic	Other	Don't know
			Expert participants		
Expert					
Acoustical	372	39	17	17	7
Causal	212	125	21	12	4
Semantic	20	19	12	9	2
Other	27	8	5	4	4
Don't know	23	6	0	0	1
			Nonexpert participants		
Nonexpert					
Acoustical	50	15	3	4	0
Causal	73	452	86	11	4
Semantic	21	88	72	6	1
Other	14	8	1	2	4
Don't know	22	7	0	1	21

even when the participants intended to describe the sound itself, which might have confused the judges (even if they also listened to the sounds in each group). This result suggests that the judges' interpretations miss in some cases the intention of the participants. Therefore, in the following analyses, only the judgments of the participants themselves will be considered.

Results and Discussion

The study design had one between-subjects factor (expertise) and one within-subject variable (Hcu), with the reported similarity as the dependent variable. The dependent variable had five modalities: "acoustical," "causal," "semantic," "other," and "don't know."

Raw results. On average, the participants made 11.5 categories (from 3 to 34). They reported having difficulties in describing the categories. They also reported that the proposed similarities were usually relevant to their own strategies. On average, they grouped together 32.2% of the sounds according to acoustical similarities, 45% of the sounds according to the similarities of their physical cause, 12.5% of the sounds because of semantic similarities, 3.3% of the sounds for other reasons, and 0.8% of the sounds because they did not know how to group them. This is coherent with the first hypothesis that participants have used different strategies, in terms of similarity, to make the classes.

Influence of expertise. Figure 3 represents the percentages of judgments in each category for the groups of expert and nonexpert participants. A chi-square test revealed that the distributions were significantly different, $\chi^2(4, N = 1800) = 66.1$, $p < .01$. Both subsamples of participants chose the categories "other" and "don't

Table 1
Cohen's κ, Measuring the Agreement Between Each Judge and Each Group of Participants

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6
Experts	0.30 (58%)	0.32 (57%)	0.28 (59%)	0.34 (60%)	0.36 (59%)	0.32 (60%)
Nonexperts	0.39 (68%)	0.33 (60%)	0.31 (58%)	0.32 (61%)	0.30 (60%)	0.30 (60%)

Note. The values in parentheses represent the percentage of agreement.

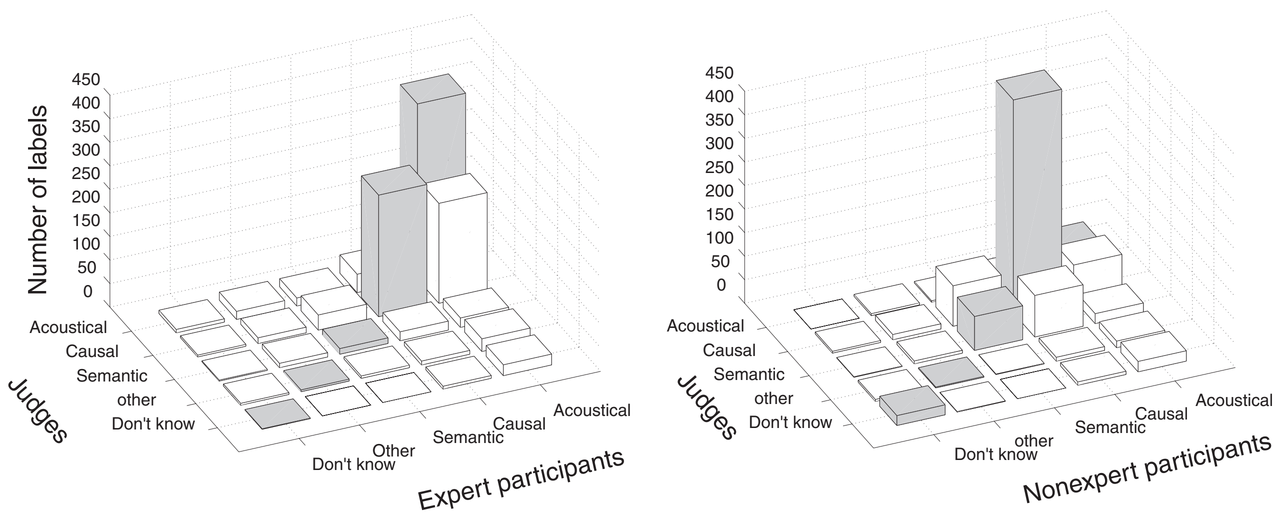

Figure 2. Graphical representation of the contingency tables (Table 2). Each bar represents the number of sounds that have received the label provided by the participants (indexed on the *x*-axis), and by the judges (indexed on the *y*-axis). The left panel represents the data for the expert participants, and the right panel the data for the nonexpert participants. The gray bars along diagonal of each panel correspond to the cases of agreement between the participants and the judges.

know" only rarely. They will not be considered in further analysis. Expert participants have chosen more often the category "acoustical," while nonexpert participants have used more often the category "causal." The relationship between the probability of choosing these two categories and the causal uncertainty values are studied in the next paragraph.

Influence of causal uncertainty. To analyze how causal uncertainty might influence the categorization strategy, the data were submitted to two binary logistic regression analyses, with the categorization strategy as the dependent variable, the expertise of the participants as an independent factor, and the causal uncertainty Hcu as an independent variable. Interaction between Hcu and expertise was also included in the model. The first analysis fitted a model that predicted the odd ratio of the probability of the category "acoustical" over the probability of the other categories ($p(acoustical)/1 - p(acoustical)$), whereas the second analysis fitted a model that predicted the odd ratio of the probability of the category "causal" over the probability of the other categories.

($p(causal)/1 - p(causal)$). Both analyses used Generalized Estimating Equations as an underlying model to allow for analysis of repeated measurements or other correlated observations (Hardin & Hilbe, 2002). In this model, an independent structure was chosen for the working correlation matrices in both analyses by comparing corrected quasi-likelihoods under independence model criterion (QICC). The significance of the coefficients of the model is summarized in Table 3.

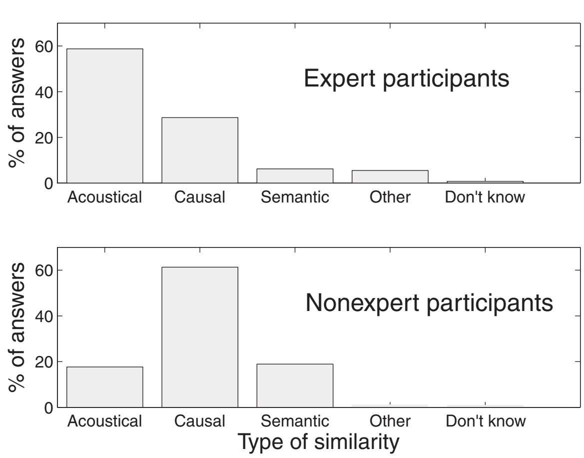

Figure 3. Percentages of judgments (over the 60 sounds) made by expert and nonexpert participants in Study 2, for each type of similarity.

Table 3
Logistic Regression Coefficients Predicting the Odds Ratio of the Probability of "Acoustical" Over the Probability of the Other Categories, and the Probability of "Causal" Over the Probability of the Other Categories

Variable	b	*SD*	Wald	*p*
	log ($p(acoustical)/1-p(acoustical)$)			
Constant	−0.215	0.3358	0.411	.522
Nonexpert	−1.982	0.4917	16.248	.000**
Hcu	0.191	0.0648	8.662	.003**
Hcu (nonexpert)	0.023	0.0856	0.071	.790
	log ($p(causal)/1-p(causal)$)			
Constant	−0.040	0.3247	0.015	.902
Nonexpert	1.173	0.4917	16.248	.033*
Hcu	−0.297	0.0825	12.927	.000**
Hcu (nonexpert)	0.075	0.1105	0.466	.495

Note. Hcu = the measure of causal uncertainty; b = value of the model coefficient; p = probability of the null hypothesis (b = 0).
* $p < .05$. ** $p < .01$.

The main effects of expertise and causal uncertainty were both significant in both analyses, though with opposite signs. The probability of choosing "acoustical" decreased for nonexperts (compared with experts), whereas the probability of choosing "causal" increased. The probability of choosing "acoustical" increased when causal uncertainty increased, whereas the probability of choosing "causal" decreased. The interaction of expertise and causal uncertainty was not significant, indicating that the rate of increase or decrease of both odd ratios did not depend on the group of participants. Figure 4 represents the logarithms of the experimental odd ratios as a function of the Hcu and of the expertise, as well as the logarithms of the odd ratio predicted by the regression model.

Discussion. The corroboration of the categorization strategies by a set of independent judges illustrates the difficulty of interpreting the categorizations a posteriori. Therefore, the strategies self-reported by the participants seem more reliable and are used to fit a logistic regression model that shows how expertise and causal uncertainty influence these strategies.

This analysis clearly showed that nonexpert participants and expert participants used different similarities to categorize the sounds: while nonexpert participants spontaneously (they received no particular instruction) grouped together sounds mainly because they identified them as caused by the same physical event, expert participants spontaneously grouped together sounds on the basis of their acoustical similarities. It may be concluded that judging the sounds according to their acoustical properties requires having been trained (implicitly or explicitly) to do so (or required specifically to do so). This is coherent with the observation made by several authors (see, for instance, Vanderveer, 1979). When listeners have to describe a sound, they describe mainly the cause (when they are able to identify it), and not the sound. Furthermore the nonexpert participants used the semantic similarities much more often to sort the sounds than the experts. This is also coherent with the hypotheses of this study.

The other major conclusion of this analysis is that the causal uncertainty of the sounds strongly influences the strategy used to categorize the sounds. Sounds with a high causal uncertainty, that is, sounds the cause of which it is hard to identify precisely, are grouped together according to acoustical similarities much more often than sounds whose cause is easy to identify. This is coherent with the hypotheses of this study. When, however, the sounds are

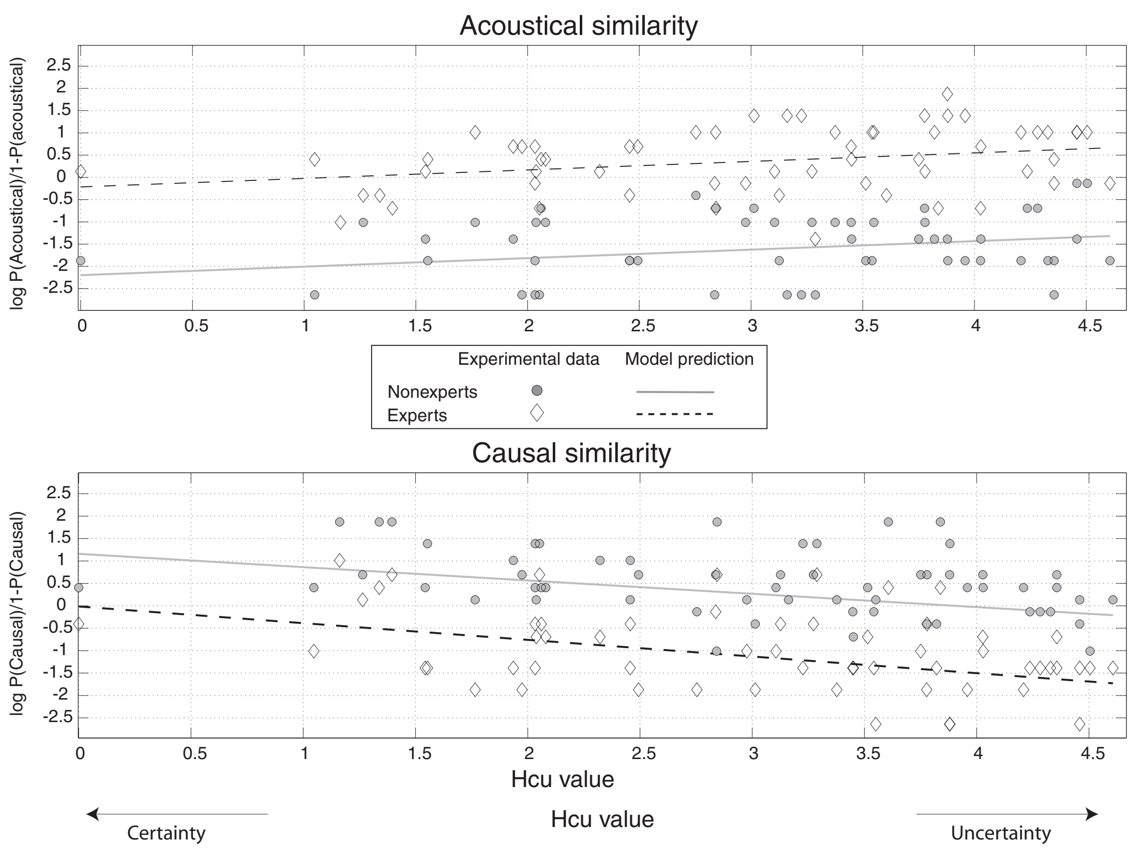

Figure 4. Logarithms of the experimental (data points) and predicted (straight lines) odd ratios, as a function of causal uncertainty (Hcu) and expertise. The odd ratios are the ratios of the frequency with which the participants in Study 2 have reported "acoustical" similarity over the frequency of the other categories of answer (upper panel), or the ratios of the frequency with which they have reported "causal" similarity over the frequency of the other categories of answer (lower panel). Negative ratios indicate that more participants have reported "acoustical" (respectively "causal") similarity for the other categories of answer, positive ratios indicate the opposite.

identifiable, nonexpert listeners group them together mainly because of the similarities of the events that they have identified, while expert listeners use both causal and acoustical criteria to group them together. Furthermore, the logistic regression analysis indicates that these two factors (expertise and Hcu) are independent.

Study 3: Measuring the Confidence in the Identification

The expertise of the participants combined with the causal uncertainty measured in Study 1 accounted for the different similarities used by the listeners to categorize the sounds in Study 2. The procedure for measuring causal uncertainty, however, was very time-consuming and erred on the side of caution, for it required a semantic analysis of verbalizations. Furthermore, the corroboration of the categorization strategies by independent judges reported in the previous section illustrated the difficulty of interpreting other participants' verbalizations, even when rules were clearly defined. To define an index leading to similar results, with an easier and more straightforward experimental procedure, we focus in Study 3 on how confident the participants are about what they identify. Ballas (1993) has shown that both measures are correlated. It is therefore expected that the measure of confidence can be used as an alternative to the laborious Hcu measure.

Method

Participants. Thirty-eight participants (23 women and 15 men) volunteered as listeners and were paid for their participation. They were aged from 21 to 63 years old (median = 29 years old). All reported having normal hearing. None of the participants reported being a professional cook or having a professional activity related to cookery. None of them was a professional musician, sound engineer, or acoustician. The participants were all native French speakers.

Stimuli. The same 60 sounds as in Study 2 were used in this Study.

Apparatus. The same apparatus as in Study 1 was used in this study.

Procedure. Participants had first to read the instructions. The participants were split in two groups. In the first group, the participants had to indicate their confidence in identifying the object causing the sound. In the second group, they had to indicate their confidence in identifying the action causing the sound. Each sound was judged twice (test/retest). The participants were not told that each sound was to be played twice during the session. For each participant, the sequence of sounds was randomized. The participants had to indicate their confidence in identifying the cause (action or object) of each sound by choosing one of the five categories along a horizontal scale (from left to right, translated from the French):

- "I don't know at all"
- "I am really not sure"
- "I hesitate between several causes"
- "I am almost sure"
- "I perfectly identify the cause of the sound"

The results were coded according to the following procedure: 0 when the answer was "*I don't know at all*" to 4 when the answer was "*I perfectly identify the cause of the sound*."

Results and Discussion

Consistency of the judgments (test/retest). For the group rating their confidence in identifying the action causing the sound, the confidence score between the test and retest remained the same for 72.1% of the sounds on average (ranging from 54.1% to 85.2% across the participants), and was different with less than one category for 93.2% of the sounds (from 80.3% to 96.7%). For the other group, the score between the test and retest remained the same for 68.9% of the sounds (from 50.8% to 82.0%), and was different with less than one category for 89.6% of the sounds (from 77.0% to 98.4%). Overall, this showed that the participants have succeeded in maintaining consistent criteria along the study. Therefore, the two scores (test and retest) were averaged. The resulting scores Co^{action} and Co^{object} (respectively confidence score for identifying the action, and for identifying the object) were considered as a numerical scale ranging from 0 (no confidence in the identification) to 4 (perfect confidence in the identification).

Object and action ratings. The correlation between the confidence ratings averaged over the participants in the group identifying the action, and the confidence ratings averaged over the participants in the group identifying the object was $r(58) = .94$ ($p < .01$). Following the same reasoning as in Study 1, both confidence scores were combined: for each sound, the confidence scores in each group were averaged within the group, and the scores for the two groups were added, resulting in a global confidence score $Co = Co^{action} + Co^{object}$, theoretically ranging from 0 to 8 for each sound (and practically ranging from 1.95 to 7.95 for the whole set of sounds).

Confidence scores and uncertainty measures. The correlation coefficient between the confidence scores *Co* measured in this study, and the causal uncertainty *Hcu* measured in Study 1 was not strong but significant: $r(58) = -0.58$ ($p < .01$). As noted in Ballas (1993), both measures were related. Figure 5 reveals, however, that these measures were different. This figure represents the correlation between the Hcu values measured in Study 1, and the confidence scores Co measured in this Study 3, as well as the distributions of these two measures.

The discrepancies between the two measures occurred because of sounds with high Hcu values and high Co values (but not the opposite). This difference was coherent with the definition of both measures. Indeed, the causal uncertainty actually measured the agreement on the identification of the cause of a sound *among* the participants: a sound had a low uncertainty value when all the participants had identified the same cause. The confidence score was the average of the *individual* confidence values in identifying the cause of the sound. A sound had a high confidence score when all the participants were certain that they were able to identify the cause of the sound. The discrepancy occurred because there were sounds for which most of the participants in Study 3 were confident in their own identification, but the identified causes of which, described in Study 1, may have been different (according to our analysis rules). In other words, participants were confident in their identification (high Co value), but what they identified may have been different (high Hcu value).

Confidence and categorization strategies. To study whether the confidence scores would allow us to explain the strategy of categorization, the logistic regression analyses of the results of Study 2 were reproduced, exchanging the Hcu values for confi-

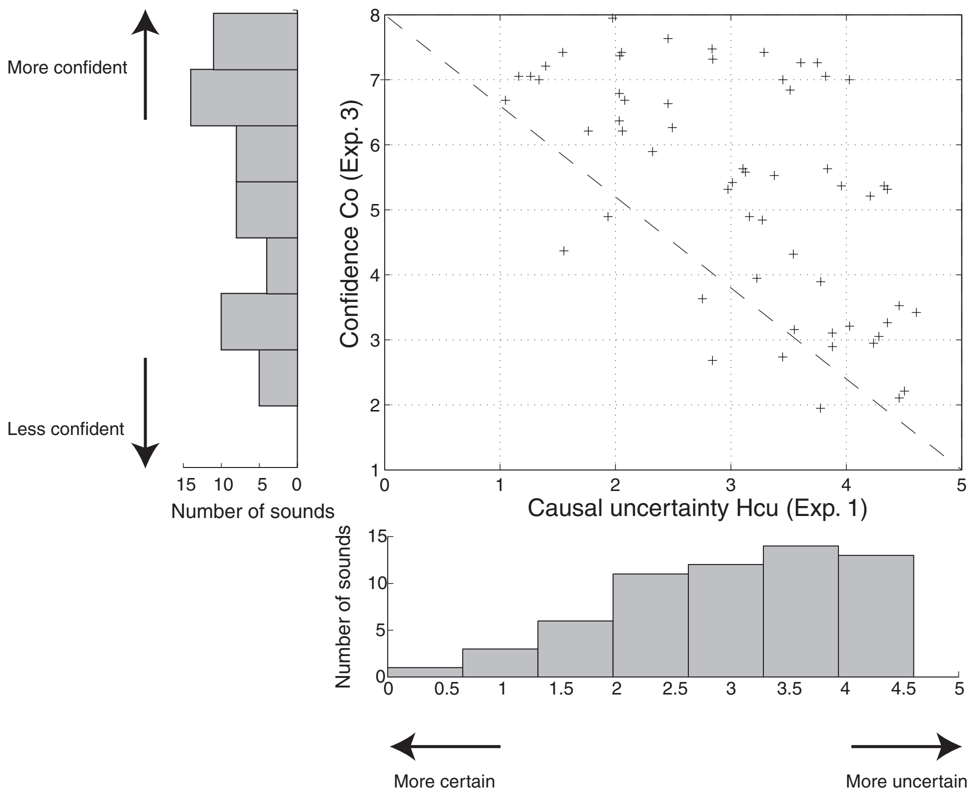

Figure 5. Correlation and distribution of the causal uncertainty values measured in Study 1 (Hcu), and the confidence scores measured in Study 3 (Co). This figure shows that some sounds with high Hcu values might also have high Co values (points spread above the dashed line), whereas the opposite is not true (there are only a few points below the diagonal).

dence scores Co. Comparing the QICC indicated that the most appropriate structure for the working correlation matrices was the independent structure for both models. The significance of the coefficients of the model is summarized in Table 4.

Table 4
Logistic Regression Coefficients Predicting the Odds Ratio of the Probability of "Acoustical" Over the Probability of the Other Categories, and the Probability of "Causal" Over the Probability of the Other Categories

Variable	b	*SD*	Wald	*p*
	log ($p(acoustical)/1-p(acoustical)$)			
Constant	1.590	0.5286	9.049	.003**
Nonexpert	−1.838	0.6482	8.039	.005**
Co	−0.225	0.0637	12.510	.000**
Co (Nonexpert)	−0.024	0.0792	0.094	.759
	log ($p(causal)/1-p(causal)$)			
Constant	−2.186	0.3551	37.908	.000**
Nonexpert	1.423	0.4159	11.704	.001**
Co	0.229	0.0435	27.784	.000**

Note. Co = confidence score; b = value of the model coefficient; p = probability of the null hypothesis (b = 0).
* $p < .05$. ** $p < .01$.

The influence of expertise and confidence scores on both odd ratios were significant. The influence of the confidence score was slightly different for expert and nonexpert participants. The interaction between confidence score and expertise was not significant. These results are summarized in Figure 6, which displays the logarithms of the experimental odd ratios as a function of the confidence score and of the expertise, as well as the logarithms of the odd ratio predicted by the regression model.

Qualitatively, the results of the regression obtained with the confidence scores were the same as the results obtained with the Hcu values. Comparing the QICC values for the regression model based on Hcu and the regression model based on Co showed, however, that the model using the confidence scores best fitted the experimental data (QICC = 2274 for the model based on causal uncertainty values, and QICC = 2225 for the model based on confidence scores).

Discussion. The confidence score provides an interesting alternative to the causal uncertainty measure. Overall, when predicting the participants' categorization strategies, it lead to the same conclusions as the causal uncertainty values: sounds with a low causal uncertainty or with a high confidence score were likely to be categorized on the basis of the identified cause, whereas sounds with a high causal uncertainty or with a low confidence score were likely to be categorized on the basis of their acoustical properties.

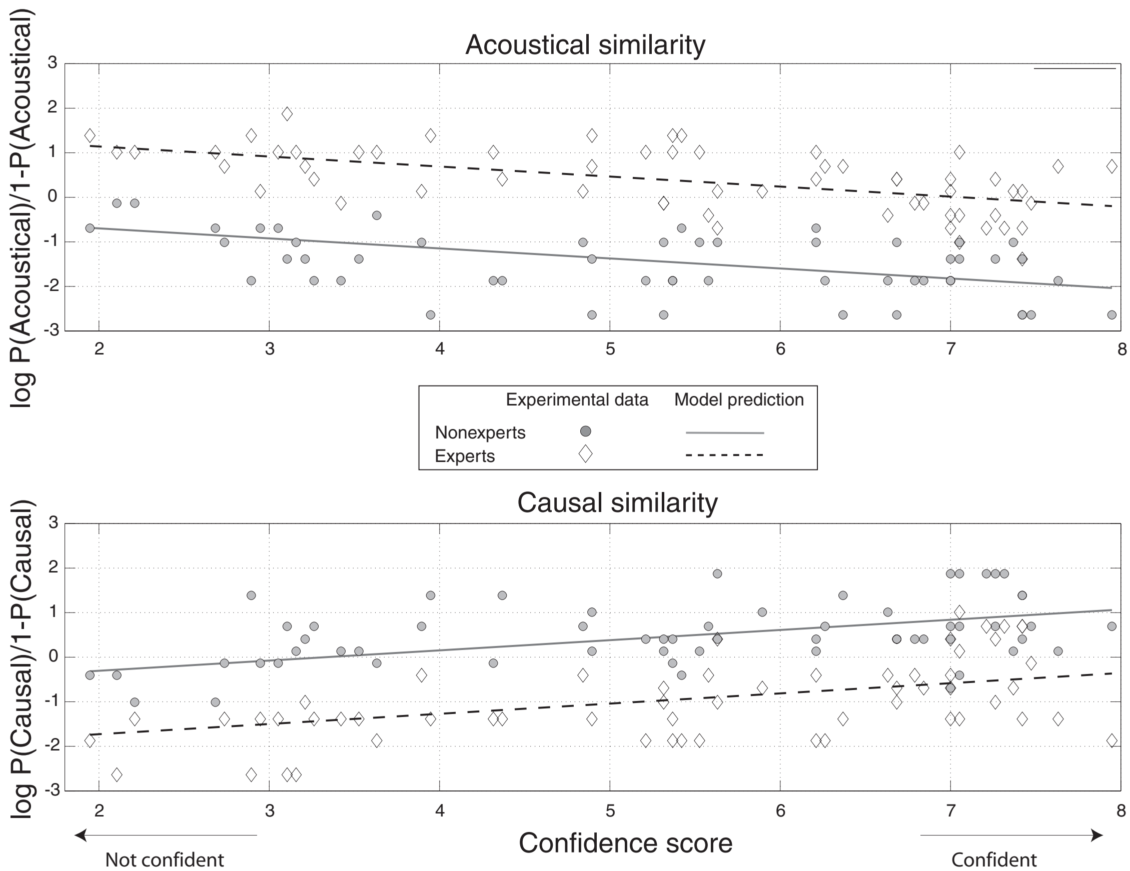

Figure 6. Logarithms of the experimental (data points) and predicted (straight lines) odd ratios, as a function of confidence score and expertise. The odd ratios are the ratios of the frequency with which the participants in Study 2 have reported "acoustical" similarity over the frequency of the other categories of answer (upper panel), or the ratios of the frequency with which they have reported "causal" similarity over the frequency of the other categories of answer (lower panel). Negative ratios indicate that more participants have reported "acoustical" (resp. "causal") similarity for the other categories of answer, and positive ratios indicate the opposite.

The confidence score, however, measured something different from what the causal uncertainty value measured. Causal uncertainty measured the agreement among listeners identifying the cause of a sound, while the confidence score measured how confident the listeners were. A typical discrepancy occurred when different individuals identified different causes but were confident in their identification. Comparing the different models, however, showed that the models using the confidence score predicted the experimental data better. This shows that the propensity to use acoustical or causal similarities was more related to the confidence scores than to the causal uncertainty. Indeed, if the participants had identify a sound with confidence (high confidence score), they would be likely to use causal similarity to group it together with other sounds, no matter what they have identified. Conversely, if they all reported similar causes (low causal uncertainty), but were not very confident in this identification, the similarities used to group this sound together with certain others might be more variable. This interpretation may explain why the confidence score is a better predictor of the categorization strategy than causal uncertainty.

General Discussion

The literature on sound perception reports theoretical considerations and experimental results suggesting that listeners may focus on different aspects of the sounds: they can focus on the sound signal itself, on what they identify as the cause of the sound signal, or on whatever meaning or memories they associate with the sound or the cause. These different "modes of listening" (Gaver, 1993a, 1993b) have practical implications in the case of sorting tasks, which have been commonly used in studies of sound perception: the main consequence is that listeners, when required to make groups of similar sounds, can use and mix the different similarities. Therefore, in the present study, we have replaced the notion of modes of listening by introducing different "types of similarity" (acoustical, causal, semantic). In many cases, however, the research goals underlying the use of a sorting task require that the listeners focus on a specific aspect of the sounds. Researchers have few methodological tools to control for this variability, because little is known about the psychological mechanisms that might influence the different modes of listening.

To address these issues, this paper has reported the results of three experimental studies using French listeners. These studies showed that both the identifiability of the sounds and the expertise of the listeners influenced the similarity used: in Study 1, the causal uncertainty of a set of sounds was measured, following a procedure developed by Ballas (1993). In Study 2, participants were required to categorize a subset of the sounds, and to indicate which type of similarity they had used. In Study 3, the confidence in identifying the sounds was measured, following a procedure also developed by Ballas, and compared with the results of Study 2. Comparison of the results of the three studies showed that the similarity used to categorize the sounds could be fairly well predicted by a combination of the expertise of the listeners and of the confidence scores. Expert participants tended to categorize sounds on the basis of their acoustical similarities, whereas nonexperts tended to base the categorization on causal similarities. Sounds confidently identified were grouped together more often on the basis of causal similarities than sounds identified without confidence. These latter sounds were grouped together more often on the basis of acoustical similarities.

The results also suggest some indication of the psychological mechanisms that support the reported difference between expert and nonexpert listeners. Two assumptions can be formulated: first, experts are different from nonexpert listeners because they have developed specific listening techniques allowing them to focus on specific aspects of a sound, useful for the craft of a sound engineer or a musician. This knowledge is also organized differently in experts and nonexperts (for instance, as suggested in Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Rosch, 1978, the base level in a taxonomy might be different for experts and nonexperts). The second assumption is that experts are capable of using a technical vocabulary that is not available to nonexperts, and can understand better the meaning of "acoustical similarity." The results of the corroboration by independent judges of the participants' description and labeling of their categories in Study 2 provide an interesting clue. Judges and participants agreed most of the time, except in one case where they strongly disagreed: when the participants reported that they had used acoustical similarities, the judges found that the categories were based on causal similarities. This systematic disagreement occurred equivalently for experts and nonexperts. Because expert listeners could not be suspected of having misunderstood the instructions (by definition, they know how to compare sounds on acoustical bases), the only possibility was that they have reported descriptions of the groups that are so confusing that the independent judges (experts as well) could not understand how they made the groups, when listening to the sounds in the group. Because this also happened with nonexperts, it must be concluded that both groups of participants used the same confusing kinds of descriptions. Therefore, the reported difference between experts and nonexperts cannot be attributed to a different comprehension of the instructions, but to different techniques of listening and to a different cognitive organization of knowledge about the sounds.

The result showing that the identifiability of the sounds had an influence on the similarity used to sort them is also important. We showed here that not all the sounds could be considered as equivalent. "Everyday listening" (Gaver, 1993a, 1993b) is only possible if the listener can identify what has caused the sound. Furthermore, the results showed that, even for expert listeners, some sounds were so easily and immediately identified that they could only be grouped together with sounds that had the same cause. The effects of sound identification (measured with Hcu or Co) and listener expertise did not interact. The effect of sound identification was rather added to the effect of expertise: the identification of the cause of the sounds had the same effect on both groups of participants. When the sounds became less identifiable, all the participants tended to categorize the sounds more often on an acoustical basis. Similarly, the more the sounds became identifiable, the more both groups of participants tended to categorize the sounds on the basis of causal similarities.

Taken altogether, these results indicate that the results of a free sorting task might collect very different results. A lot of precautions are therefore required in the interpretation of such results. It can, for instance, be expected that experimental classifications of sounds "averaged" over a set of unselected participants also average different strategies of categorization. Without any specific instruction, a participant is more likely to do what he or she is used to doing: listening to the cause for listeners without any specific sound expertise, and analyzing the sound for listeners used to analyzing the sounds. It is likely, however, that specific instructions could also direct the listener's attention toward specific properties of the sounds or toward the cause of sounds. Carefully choosing participants and designing experimental instructions are therefore a first solution to obtaining an experimental classification of environmental sounds based, for instance, solely on causal similarities (if, for instance, the purpose is to study the performances of a user required to recover the properties of the source of the sound). Moreover, not all the sounds can be considered as equivalent as regards identification, and are liable to influence the mode of listening. The measure of the confidence score of the sounds used in Study 3 offers an interesting tool for selecting sounds that trigger one or other modes of listening.

From another methodological perspective, the results reported in this article suggest that using assessments directly made by the participants is more reliable than interpretations provided a posteriori by independent judges. First, the causal uncertainty measured in Study 1 required a huge effort from the independent judges, and was suspected of being very dependent on each judge's subjective interpretation. The confidence scores gathered in Study 3 required a lighter procedure, and proved to be correlated with the causal uncertainty, and could explain many of the data in Study 2 as well as does the causal uncertainty. Second, the analysis of the corroboration of the categorization strategies by independent judges in Study 2 suggested that they could not interpret the verbalizations of the participants properly. Altogether, these results suggest that the participants in an experimental study are probably the best experts to judge and analyze their behaviors.

The results of these studies also have practical implications for audio applications. Because of the increasing quantities of sounds currently available to sound practitioners, there is a growing demand for systems capable of automatically classifying enormous sound databases. Some of these systems (called supervised classifiers) learn how to categorize a sound in a set of labeled categories (i.e., to place it in a category), by being trained on datasets consisting of a great number of labeled sounds. Users of such automatic tools (e.g., a sound engineer trying to sort a sound from a huge collection of samples) usually complain that such systems do not provide the proper categories. The results reported here suggest why: there are many possible categorization processes, and the most appropriate categorization depends on the sound, the user, and the context of use. The future of such systems lies therefore in accounting for different possible similarities between the sounds, thus being capable of providing different concurrent categorization strategies.

References

Aldrich, K. M., Hellier, E. J., & Edworthy, J. (2009). What determines auditory similarity? the effect of stimulus group and methodology. *Quarterly Journal of Experimental Psychology, 62,* 63–83.

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

Ballas, J. A., & Howard, J. H. (1987). Interpreting the language of environmental sounds. *Environment and Behavior, 19,* 91–114.

Ballas, J. A., & Mullins, T. (1991). Effect of context on the identification of everyday sounds. *Human Performance, 4,* 199–219.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *TRENDS in Cognitive Sciences, 7,* 84–91.

Bigand, E., & Poulin-Charronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. *Cognition, 100,* 100–130.

Bonebright, T. L., Miner, N. E., Goldsmith, T. E., & Caudell, T. P. (2005). Data collection and analysis techniques for evaluating the perceptual qualities of auditory stimuli. *ACM Transactions on Applied Perception, 2,* 505–516.

Cabe, P. A., & Pittenger, J. B. (2000). Human sensitivity to acoustic information from vessel filling. *Journal of experimental psychology: Human perception and performance, 26,* 313–324.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America, 118,* 471–482.

Carello, C., Anderson, K. L., & Kunkler-Peck, A. J. (1998). Perception of object length by sound. *Psychological Science, 9,* 211–214.

Faure, A., McAdams, S., & Nosulenko, V. (1996, August). Verbal correlates of perceptual dimensions of timbre. In *Proceedings of the fourth international conference on music perception and cognition* (pp. 79–84). Montréal, Canada: McGill University.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin, 76,* 378–382.

Freed, D. J. (1990). Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events. *Journal of the Acoustical Society of America, 87,* 311–322.

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

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

Gérard, Y. (2004). *Mémoire sémantique et sons de l'environnement.* [*Semantic memory and environmental sounds.*] Unpublished doctoral dissertation, Université de Bourgogne.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin.

Giordano, B. L., & McAdams, S. (2006). Material identification of real impact sounds: Effect of size variation in steel, glass, wood and plexiglass plates. *Journal of the Acoustical Society of America, 119,* 1171–1881.

Goldstone, R. L., & Kersten, A. (2003). Concepts and categorization. In A. F. Healy & R. W. Proctor (Eds.), *Comprehensive handbook of psychology, Vol. 4: Experimental psychology* (p. 599–621). Hoboken, NJ: Wiley.

Grassi, M. (2005). Do we hear size or sound? Balls dropped on plates. *Perception and Psychophysics, 67,* 274–284.

Guastavino, C. (2007). Categorization of environmental sounds. *Canadian Journal of Experimental Psychology, 61,* 54–63.

Guyot, F. (1996). *Éétude de la perception sonore en terme de reconnaissance et d'appréciation qualitative: Une approche par la catégorisation.* [*A study of sound perception, in terms of recognition and qualitative assessment: A categorization approach.*] Unpublished doctoral dissertation, Université du Maine.

Hajda, J. M., Kendall, R. A., Carterette, E. C., & Harshberger, M. L. (1997). Methodological issues in timbre research. In I. Deliège & J. Sloboda (Eds.), *Perception and cognition of music* (p. 253–306). Hove, East Sussex, UK: Psychology Press.

Hardin, J. W., & Hilbe, J. (2002). *Generalized estimating equations.* Bora Raton, FL: CRC Press.

Houben, M. M. J., Kohlrausch, A., & Hermes, D. J. (2004). Perception of the size and speed of rolling balls by sound. *Speech Communication, 43,* 331–345.

Howard, J. H., & Ballas, J. A. (1980). Syntactic and semantic factors in the classification of nonspeech transient patterns. *Perception and Psychophysics, 28,* 431–439.

Kendall, R. A., & Carterette, E. C. (1991). Perceptual scaling of simultaneous wind instrument timbres. *Music Perception, 8,* 369–404.

Klatzky, R. L., Pai, D. K., & Krotkov, E. P. (2000). Perception of material from contact sounds. *Presence, 9,* 399–410.

Köster, O., Jessen, M., Khairi, F., & Eckert, H. (2007, August). Auditory-perceptual identification of voice quality by expert non-expert listeners. In *Proceedings of the 16th international congress of phonetic sciences (ICPhS XVI),* Sarrebrücken, Germany: Universität des Saarkandes.

Kunkler-Peck, A. J., & Turvey, M. T. (2000). Hearing shape. *Journal of Experimental psychology: Human perception and performance, 26,* 279–294.

Lakatos, S., McAdams, S., & Caussé, R. (1997). The representation of auditory source characteristics: Simple geometric sources. *Perception and Psychophysics, 59,* 1180–1190.

Legendre, P., & Legendre, L. (1998). *Numerical ecology* (2nd English ed.). Amsterdam: Elsevier.

Lemaitre, G., Susini, P., Winsberg, S., Letinturier, B., & McAdams, S. (2007). The sound quality of car horns: A psychoacoustical study of timbre. *Acta Acustica United With AcusticFia, 93,* 457–468.

Li, X., Logan, R. J., & Pastore, R. E. (1991). Perception of acoustic source characteristics: Walking sounds. *Journal of the Acoustical Society of America, 90,* 3036–3049.

Lutfi, R. A. (1995). Correlation coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks. [Letter to the Editor] *Journal of the Acoustical Society of America, 97,* 1333–1334.

Lutfi, R. A. (2001). Auditory detection of hollowness. *Journal of the Acoustical Society of America, 110,* 1010–1019.

Lutfi, R. A., & Oh, E. L. (1997). Auditory discrimination of material changes in a struck-clamped bar. *Journal of the Acoustical Society of America, 102,* 3647–3656.

Lutfi, R. A., Oh, E., Storm, E., & Alexander, J. M. (2005). Classification and identification of recorded and synthesized impact sounds by practiced listeners, musicians and non musicians. *Journal of the Acoustical Society of America, 118,* 393–404.

Marcell, M. M., Borella, D., Greene, M., Kerr, E., & Rogers, S. (2000). Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology, 22,* 830–864.

Marozeau, J., Cheveigné, A. de, McAdams, S., & Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *Journal of the Acoustical Society of America, 114,* 2946–2957.

McAdams, S. (1993). Recognition of auditory sound sources and events. In S. McAdams & E. Bigand (Eds.), *Thinking in sound: The cognitive psychology of human audition.* Oxford: Oxford University Press.

McAdams, S., Chaigne, A., & Roussarie, V. (2004). The psychomechanics of simulated sound sources: Material properties of impacted bars. *Journal of the Acoustical Society of America, 115,* 1306–1320.

McAdams, S., Winsberg, S., Donnadieu, S., Soete, G. D., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities and latent subject classes. *Psychological Research, 58,* 177–192.

Peretz, I., & Zatorre, R. J. (2005). Brain organization for music processing. *Annual Review of Psychology,* 56, 89–114.

Preis, A., & Chuzdzicka, M. (2004, April). Expert and non-expert judgements of musical instruments: Subjective evaluation vs. acoustical characteristics of musical sound. In *Proceedings of the international conference on acoustics ICA 04.* Kyoto, Japan: Acoustical Society of Japan.

Repp, B. H. (1987). The sound of two hands clapping: An exploratory study. *Journal of the Acoustical Society of America, 81,* 1100–1109.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (p. 27–48). Philadelphia: Erlbaum.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology, 8,* 382–343.

Smith, B. K. (1995). PsiExp: An environment for psychoacoustic experimentation using the IRCAM musical workstation. In *Society for music perception and cognition conference '95.* Berkeley, CA: University of Berkeley.

Smits, R., Sereno, J., & Jongman, A. (2006). Categorization of sounds. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 733–754.

Solomon, L. N. (1958). Semantic approach to the perception of complex sounds. *Journal of the Acoustical Society of America, 30,* 421–425.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology, 23,* 457–482.

Vanderveer, N. J. (1979). *Ecological acoustics: Human perception of environmental sounds.* Unpublished doctoral dissertation, Cornell University, Ithaca, New York.

von Bismarck, G. (1974). Timbre of steady sounds: A factorial investigation of its verbal attributes. *Acustica, 30,* 146–159.

Warren, W. H., & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events: A case study in ecological acoustics. *Journal of Experimental Psychology: Human perception and performance, 10,* 704–712.

Appendix A

List of the Sounds Used in Study 1

#	Description (translated from the French)	Maximum level (dB)	Duration (s)
1	Ice cubes in an empty glass	67	1.0
2	Air conditioning	55	5.1
3	Water drops	61	4.4
4	Boiling water	53	3.2
5	Closing a dishwasher door	71	3.2
6	Dishwasher on	66	3.0
7	Coffee maker with filter on	62	6.9
8	Water boiling in a pan	56	3.9
9	Gas open and furnace on	66	4.1
10	Champagne cup shocked	59	1.0
11	Furnace on. Hot thermostat	59	1.9
12	Striking and igniting a match	62	1.1
13	Opening and closing a furnace	72	2.0
14	Lowering the toaster compartment	63	2.9
15	Ejection of the toaster compartment	68	1.9
16	Bip-bip of a microwave	55	3.0
17	Agitating hands in water	66	2.4
18	Microwave on	55	3.8
19	Food processor	72	3.5
20	Mixer on	72	2.8
21	Electric lemon squeezer	71	3.1
22	Knife removed from his case	66	1
23	Cutting foods with a knife	65	2.9
24	Scraping a metal pan	67	0.56
25	Closing a refrigerator door	65	1.2
26	Pop-up from a toaster	56	5.0
27	Closing a refrigerator door	62	0.71
28	Compressor noise of a refrigerator	44	3.5
29	Gas open of a furnace	42	3.6
30	Putting a bowl on a table	59	1.4
31	Putting a bowl on a table	67	1.1
32	Closing a cupboard door	71	2.4
33	Closing a cupboard door	67	1.1
34	Turning on a faucet	59	2.4
35	Emptying a sink	68	5.0
36	Pouring wine into a glass	67	7.5
37	Bottle shocked	72	0.79
38	Bottle top	58	0.72
39	Putting a bowl on a saucer	70	2.7

#	Description (translated from the French)	Maximum level (dB)	Duration (s)
40	Cutting bread	61	1.2
41	Coffee maker is whistling	67	10
42	Coffee maker with filter on	63	2.7
43	Removing a cork stopper	63	1.5
44	Refrigerator	64	3.4
45	Dishes	73	0.47
46	Removing a metal top from a kettle	69	0.94
47	Pouring water into a metal kettle	67	3.8
48	Closing a kettle	70	0.97
49	Cooking with fat	64	2.4
50	Microwave alarm	69	3.5
51	Crushing a paper bag	68	2.5
52	Crumpling a plastic bag	67	4.4
53	A big bubble inside a metal kettle	62	4.1
54	Shutting a wooden chair	75	1.7
55	Shaking water in a basin	66	3.8
56	Unscrewing a stopper	58	3.2
57	Closing a cupboard door	72	1.4
58	Closing a door	73	1.0
59	Opening a drawer with castors	71	1.7
60	Unrolling a blind	68	2.7
61	Pouring a drink into a glass	61	6.6
62	Screwing the bottle top	67	1.0
63	Taking off the bottle top	65	2
64	Screwing a bottle stopper	60	2.5
65	Several sprays from an atomizer	59	1.9
66	Evacuating air from a crushed bottle	62	1.4
67	Opening a metallic can	70	1.3
68	Crushing a metallic can	65	1.2
69	Closing the top of an aerosol bomb	69	1.4
70	Spray from an aerosol	63	1.3
71	Irregular spray from an aerosol	66	2.2
72	Spray from an aerosol	68	3.7
73	Putting a porcelain lid on a pan	70	0.55
74	Removing the top of a plastic container	59	0.52
75	Closing the top of a plastic container	67	1.5
76	Removing the metallic lid of a pan	66	3.2
77	Turning a spoon inside an empty cup	70	2.8
78	Hand washing-up	61	2.5
79	Beating eggs inside a container	71	6.1
80	Pouring cereals into a bowl	68	5.1
81	Pouring milk on cereal in a bowl	57	5.5
82	Egg opening	62	2.2
83	Grating carrots	56	4.4
84	Cutting vegetable with a knife	67	6.2
85	Pulling out vegetable sprays	62	1.8
86	Cutting salad	59	2.6
87	Gas noise of a furnace	32	2.1
88	Garbage top falling	72	1
89	Grinding salt mechanically	62	0.51
90	Putting a top on a container	65	0.87
91	Knife sharpening	66	0.47
92	Closing the micro-wave door	72	1.5
93	Mixer on	70	1.3
94	Unrolling absorbing paper. Detaching a sheet	67	2.3
95	Lamp switch	66	0.59
96	Drops in a container	60	5.8
97	Drops in a container	64	3.6.0
98	Water running in a sink	67	5.0
99	Filling a sink with water	70	6.0
100	Emptying a sink	63	3.4
101	Flow of water and stop	69	4.1

(*Appendices continue*)

Appendix B

Rules Used to Sort the Descriptions in Study 1 (translated from the French)

- If an object described is a part of another object described, both descriptions have to be grouped together (e.g., “blade” and “knife”).
- If two objects belong to the same category, and if this category is defined by shared properties of form and material, the objects have to be considered as equivalent (e.g.. “knife” and “spoon”, “plate” and “dish”).
- If the category is too broad, and if the objects belonging to the category have different geometries, materials, etc., the objects have to be considered as different (e.g., “microwave” and “refrigerator” have not to be considered as equivalent, yet they are both electrical appliances; the same applies to “soup” and “ham,” yet both are foods).
- The name of a material has to be considered as equivalent to an object made out of this material (e.g., “tin” and “knife”).
- If an action is a particular case of another one, both have to be considered as equivalent (e.g., “taking a knife” and “grabbing a knife”).
- If an action is needed for doing another one, they have to be considered as equivalent (e.g., “sharpening a knife” and “rubbing the blade of a knife”).
- If an action is an element of a sequence composing another one, they have to be considered as equivalent (e.g., “touching,” “grabbing,” “lifting” are part of “taking a knife”).
- “Using” and a verb corresponding to the regular use of an object have to be considered as equivalent (e.g., “using a knife” and “cutting with a knife”).

Received October 25, 2008
Revision received October 22, 2009
Accepted November 20, 2009 ■