ACOUSTICAL AND CONCEPTUAL INFORMATION FOR THE PERCEPTION OF ANIMATE AND INANIMATE SOUND SOURCES

Bruno L. Giordano

Stephen McAdams

CIRMMT – Schulich School of Music
McGill University
Montréal, Québec, Canada
bruno.giordano@music.mcgill.ca
smc@music.mcgill.ca

John McDonnell

Cognitive Science
McGill University
Montréal, Québec, Canada
jmcdon10@po-box.mcgill.ca

ABSTRACT

Is the sound of a train whistling more similar to the sound of the wheels of a train or to the sound of your whistle? This question addresses the comparative relevance of acoustical and conceptual information to the perceived similarity of sound events. The answer to this question has theoretical and methodological consequences for the field of sound source perception, and for the behaviorally informed synthesis of environmental sounds. Hierarchical sorting was used to collect measures of the similarity of large sets of animate or inanimate sounds in naive listeners. Results were compared with those from two other conditions based on the same data–collection technique. Conceptual similarity was measured by presenting the sound source identification labels (written words) collected during a free–identification experiment. Acoustical similarity was measured on heard sounds, after participants received a training meant to minimize the effects of conceptual information on sorting. Acoustical similarity was only weakly correlated with conceptual similarity, proving the effectiveness of the training methodology in the acoustical condition. Also, naive listeners focused on conceptual and acoustical information when judging the similarity of animate and inanimate sound events, respectively. Theoretical and methodological consequences of these results are discussed.

[Keywords: Sound source perception, Environmental sounds, Hierarchical sorting]

1. INTRODUCTION

The field of sound source perception investigates the ability of a listener to recognize sound–generating events populating the everyday environment, and the information relevant to this process. Research on the perception of a sound source is often linked with the ecological approach to perception, whose main assumption is that the primary object of perception is not the sound but the sound source (cf. [1] [2]). Accordingly, a shattered glass is identified because the perceptual system "resonates" to the acoustical pattern specifying the physical event of a breaking (cf. [3]). From a computational standpoint [4], the sound is transduced into patterns of neural activity within the peripheral auditory system; subjected to grouping processes that merge together pieces of auditory information likely originating from the same event; analyzed in terms of auditory features; associated with the mental representation defined by features matching those of the incoming auditory object [5]. It should be noted that although the matching process operates on auditory features, it is not completely independent of top–down influences (e.g., relevant features might be selected by means of top–down attentional processes [6]). Finally, in virtue of the matching process the sound source can be named, and an appropriate motor program eventually selected and executed [7].

A knowledge of the features of environmental sounds weighted by the matching process has both a theoretical and an applied impact. From the theoretical point of view, such a knowledge provides an ecologically meaningful understanding of the architecture of the auditory cognition system. Specifically, it was used in past studies to address specific assumptions of the ecological approach (e.g., does perception focus on acoustical properties that accurately discriminate between source categories? [8]). From the applied point of view, knowledge of the acoustical features for the perception of environmental sounds could inform the design of hearing aids and of sonification systems. More specifically, physically–informed sound synthesis could be guided so as to concentrate modeling efforts onto parameters characterized by the highest perceptual effectiveness.

Within the field of source perception, perceptually relevant features are inferred from models of human judgment based on the acoustical properties of a signal. Most often, participants were asked to judge a property of the sound source specified by the experimenter (e.g., identify the material of the struck object [9]). Generalization of these results to a non–directed listening situation (e.g., hearing the clinking of glasses while at a restaurant) requires assuming that such a judgment is indeed made in absence of the instructions of the experimenter. This assumption can be mistaken, and testing it empirically might prove far from trivial. Alternatively, participants can be asked to estimate the similarity of sound events, thus avoiding the need to constrain judgment along a pre–specified property of the sound source [10]. Thus, acoustical features for perception would be derived from the measurement of the basic cognitive operation of similarity estimation, an operation frequently performed outside of the laboratory to categorize [11] and ultimately identify [7] events and objects in the environment.

Similarity estimation has been used to investigate sets of homogeneous or highly heterogeneous sound sources (e.g., impacting objects varying in size and material in [10]; from vocalizations to splashes in [12]). The study of homogeneous sound sets reveals the acoustical basis for the perceptual differentiations within a class of environmental sounds, but does not uncover the acoustics for the perception of the class itself [13]. The acoustical basis for the identification of large classes of sound events can instead be investigated with sets of highly heterogeneous events. However, similarity estimation of heterogeneous sound sets is not free of methodological drawbacks. With both homogeneous and heterogeneous sound sets, similarity is highly likely influenced by the acoustical structure of the signals. However, particularly for heterogeneous sets, similarity is likely to weight also the links between the mental structures activated by the recognized sound–generating event (e.g., sensory–related structures as visual, haptic and motor memories associated with the sound source; conceptual structures as knowledge of the sound–generation mechanics and of the most frequent context, or folk taxonomies of sound–generating events).

A test for the relevance of non–acoustical information to the estimation of the similarity of environmental sounds is then of both methodological and theoretical relevance. From a methodological point of view, it tests the extent to which similarity estimates can yield a reliable knowledge of those acoustical features relevant to the perception of environmental sounds. Indeed, if the contribution of non–acoustical factors to judgment was ignored, acoustical measures would be used to explain human judgment of non–acoustical information. As a consequence, acoustical models of human judgment would be biased at best. From a theoretical point of view, characterization of the relative relevance of acoustical and non–acoustical information to similarity estimation would yield a deeper understanding of the nature of the representations and processes upon which source perception is based. Notably, the ecological approach to perception assumes a perceptual primacy of source properties (cf. [1] [2]). This assumption would be confirmed by a focus of judgment on auditory information, but would be disconfirmed by a focus on non–auditory information. For example, source properties wouldn't clearly be the primary object of perception if listeners spontaneously evaluated sounds' similarity on the basis of context commonalities. On the other hand, if similarities were spontaneously evaluated focusing on acoustical properties, the same assumption would be grossly validated, for the simple fact that similar physical system tend to produce signals with common acoustical properties (e.g., liquid sounds in general are likely more similar to each other than to solid sounds). As a consequence, a test of the above–mentioned assumption of the ecological approach would require quantification of the relevance of non–auditory information to similarity estimation in naive listeners, i.e., in absence of instructions biasing judgment toward the selective use of a specific type of information.

Among the previous studies focusing on sets of heterogeneous environmental sounds, the most relevant to the current investigation are those illuminating effects of stimulus type and of instructions on similarity estimation (see [12] for a comprehensive review of the literature). In [14], a set of inanimate sounds (i.e., generated by the vibration of objects not part of a living being; cf. Section 2.1.1) was evaluated in two different conditions. Participants arranged sound stimuli on a two–dimensional display, placing similar sounds closer to each other. They focused on the similarity of either the timbre of the sounds (acoustical properties) or of the visual image activated by the sounds (likely the visual memories of the sound–generating event). Similarity estimates differed among conditions, supporting the ability of listeners to selectively focus on the acoustical properties of sound events. In [12], a mixed set of animate and inanimate sounds was evaluated in four different conditions. When presented sound identification labels, participants rated the similarity of the mental representation of either the sound (memory trace of the sound event) or of the sound–generating event (knowledge of the sound generation mechanics and/or visual memory of the sound source). When presented sound events, participants were instructed to focus on their similarity, without further specifications (unbiased conditions). They estimated similarity either in a rating or in a free–sorting task. Two main effects emerged. Firstly, data in the rating conditions strongly resembled each other. In contrast with what observed in [14], this effect was interpreted as revealing an inability to ignore knowledge of the sound generating event when estimating similarities. Secondly, rating data from all conditions were, at best, weakly correlated with similarity estimates from the free–sorting task. This was interpreted as reflecting a difference in the mental processes operating while the task was carried out. An alternative interpretation for the results in [12] can be advanced. In particular, the high resemblance of data from the rating conditions was likely caused by a focus of all participants on a highly salient distinction, that between animate and inanimate sound events. Consistently, multidimensional scaling (MDS) models for these data sets invariably showed the first MDS dimension to almost perfectly separate animate sounds (mainly vocalizations) from inanimate sounds. As a reference, the introduction of the categorical distinction between impulsive and continuant timbres likely lead [15] to overestimate the invariance of timbre across a variety of signal manipulations (see [16] for further comments about this study). Interestingly, also the first MDS dimension for the sorting task in [12] afforded a categorical distinction between animate and inanimate sounds, and participants frequently created groups of animals/human sounds.

To summarize, previous studies disagree on the extent to which similarity can be estimated focusing on acoustical properties alone [14], or, almost equivalently, ignoring higher–level knowledge of the sound–generation mechanics [12]. In particular, selective judgment capabilities emerge when listeners do not focus on the animate/inanimate source distinction, which seems to predominate on judgment independently of instructions. A logical next step would be to characterize acoustical and unbiased judgment of sound events for animate and inanimate sources independently. This methodological choice would allow a rigorous comparison of the cognitive processes involved in the perception of these classes of sound events. Interestingly, previous studies of audition pointed out temporal and spatial differences in the neural processing of the sounds generated by living and man-made objects (animate and inanimate sources, respectively) [17] [18]. Most importantly, [19] investigated semantic priming in the identification of animate and inanimate sounds. While a semantically related prime significantly reduced identification time for animate sounds, a facilitation was not observed for inanimate sounds. As a consequence it could be concluded that conceptual components represent a relevant part of the cognitive processing of animate, but not of inanimate sound events.

The rest of this paper is structured as follows. In Section 2, a free–identification study is presented, conducted on a large set of animate and inanimate sound events. Identifiability measures thus derived guided the selection of stimuli for a second experiment. In Section 3, a similarity estimation study based on the hierarchical sorting technique is presented. The similarity of animate and inanimate sound events was estimated in separate experimental sessions, and in three different conditions. Separate groups of participants estimated either the similarity of the acoustical properties of the sound events or of the meaning of the sound identification labels collected during Experiment 1. A third group of participants estimated the similarity of sound events in unbiased conditions, i.e., without specification of the similarity estimation criteria. Comparison of the results from the different conditions revealed an ability to estimate the similarity of acoustical properties independently of the conceptual knowledge activated by the sound events. Also, estimation of similarity in unbiased conditions relied on conceptual information for animate sources and on acoustical information for inanimate sources. The implications of these results are discussed in Section 3.3.

2. EXPERIMENT 1: FREE–IDENTIFICATION

Free–identification of a large set of environmental sounds was investigated. The identification labels and the measures of identification accuracy thus derived were used as stimuli and as stimuli selection guidelines in Experiment 2, respectively. Identification times were correlated with measures of identification accuracy. Previous studies found this correlation to be negative [20]. A replication of this result was assumed to validate the adopted measure of identification accuracy.

2.1. Methods

2.1.1. Stimuli

Stimuli were selected from a royalty–free database of sound effects (The General 6000 from Sound Ideas), complemented by additional online and published resources [21], and by a database of musical instrument tones [22]. The stimulus set did not include speech samples, synthetic sounds, Foley sounds, and *complex* and *hybrid* sounds, generated by multiple interaction types (e.g., rolling and impact in bowling sounds) and by vibrating matters of multiple states (e.g., liquid and solid as in coffee stirring), respectively [1]. These selection guidelines were however violated for a limited number of sound events (e.g., the hybrid sound of crackling fire was included).

Signals were classified in terms of the properties of the sound–generating objects and events, and in terms of higher–level source–related properties (e.g., context). The classification system guided stimuli selection and was meant to maximize the acoustical diversity and conceptual connectedness of the selected events, rather than to provide a comprehensive taxonomy of environmental sounds.

All sounds were classified on the basis of the following three criteria: **1. Source animacy** (the sound–generating object is/is not part of the body of a living being); **2. Agent animacy** (the sound is/is not the result of the motor activity of a living being); **3. Musicality** (the sound source is commonly referred to as a musical instrument).

Additional distinctions were carried within each of four different classes. **1. Animate sources**: ***1a. Taxonomical class*** (amphibians birds, insects, humans, non–human mammals); ***1b. Vocalization*** (the sound is/is not produced by a phonatory apparatus); ***1c. Communication*** (the sound has/has not a communicative function). **2. Inanimate–musical sources**: ***2a. Musical instrument family*** [23] (aerophone, chordophone, idiophone, membranophone); ***2b. Excitation type*** [16] (impulsive, continuant, multiple impacts). **3. Inanimate–non musical sources**: ***3a. Material class*** [1] (aerodynamic, combustion, electric, liquid, solid); ***3b. Interaction type*** [1] (*Aerodynamic* – continuous, steam, whoosh, wind; *Combustion* – simple, crackling; *Electric* – explosive, continuous; *Liquid* – bubbling, dripping, flowing, pouring, sloshing, splashing; *Solid* – deformation, impact, rolling, scraping). **4. Animate agent**: ***4a. Locomotion*** (the sound is generated by the locomotion of the agent); ***4b. Alimentation*** (the sound is generated during the alimentation of the agent).

Finally, signals were classified in terms of their context, i.e., the location where sounds were generated (and not experienced). In absence of such information a guess was made about the most frequent location of a sound source. The following classification was adopted: **1. Animate sources** anywhere, indoors–generic, toilet, farm, sea, wild; **2. Inanimate–non musical sources**: anywhere, casino, party, indoors–generic, kitchen, toilet, construction, military, office, outdoors–generic, sea, wild, sport, store, travel–generic, bicycle travel, marine travel, railways travel. Musical sources were assumed as potentially generated anywhere.

The intersection of the above–defined classes defined categories of interest considered for the sound selection (e.g., a vacuuming sound belonged to the “inanimate source – animate agent – non musical – aerodynamic – continuous – indoors–generic” category). At least one sound per category was selected randomly. Signals judged non–typical or unidentifiable were replaced with a more suitable category member. A set of 70 stimuli was selected from the animate–source, alimentation and locomotion classes; another set of 70 stimuli were selected from the inanimate–source class. From now on these stimulus sets will be referred to as animate and inanimate sounds, respectively.

Signals were edited to a minimal duration, following the requirement that the stimulus allowed the event to “unfold naturally” (cf. [24]). More precisely, sounds were edited to the minimal duration required to keep a signal representative of the generating sound source (in this sense, a single shoe impact is not a representative footsteps sound). Signals’ level was left unmodified from the recordings.

2.1.2. Procedure

On each trial, participants were presented a stimulus and asked to identify the sound generating event using at least one verb and one noun. Blank responses were not allowed. They were free to use a second noun if necessary. Responses were typed in on-screen blank areas labeled “Verb”, “Noun1” and “Noun2”. Participants were asked to maximize identification accuracy, avoiding generic responses (e.g., thing). They could play each of the stimuli as many times as needed, but were instructed to maximize identification speed. When they were satisfied with their identification they clicked on an on–screen button to begin the next trial. Identification time measured the temporal distance between the beginning and the end of a trial. Each of the stimuli was identified once by each of the participants. Stimuli were presented in random order. At the beginning of the experiment participants were presented with all the stimuli in random order, separated by a silence interval of 100 ms. The experiment lasted approximately 2 hours.

Stimuli were stored on the hard disk of a Mac G5 Worksation, equipped with a M–Audio Audiophile 192 S/PDIF interface. Audio signals were amplified with a Grace Design m904 monitor system and presented through Sennheiser HD280 headphones. Participants sat inside a IAC double–wall soundproof booth. Signal peak level ranged from 10 to 53 dB SPL.

2.1.3. Participants

Twenty–one native English speakers took part in the experiment (10 females, 11 males; age: 18–25; mean age: 21.14; 6 amateur musicians, 15 professional musicians). Normal hearing was assessed, measuring the hearing threshold in both ears on octave-spaced frequencies (125–8000 Hz). A standard audiometric procedure was used to this purpose. Hearing thresholds never exceeded normative values of more than 15 dB [25] [26].

2.2. Results

Data from one participant who consistently confused the "Noun1" and "Verb" fields, and never used the "Noun2" field, were not considered.

Three subsequent analyses were carried out on the verbal data: naming agreement; conceptual agreement; identification performance.

Occasional confusions between the verb and noun fields and spelling mistakes were corrected. Non–words and adjectives were discarded. Complex nouns (two nouns or one noun and one adverb) were kept if commonly used to reference specific exemplars of an object. Verbs and nouns were reduced to the gerund and singular form, respectively. Only the first of multiple alternatives for each of the response categories (i.e., Verb, Noun1 and Noun2) was considered. For each of the stimuli, modal (i.e., most frequent) responses were extracted for each of the response categories. Blank responses were not considered to this purpose. Naming agreement for each of the response categories was given by the proportion of participants who gave the modal response. Naming agreement was also computed for the Noun1 and Noun2 responses considered together. A summary of the naming agreement analysis is given in Figure 1

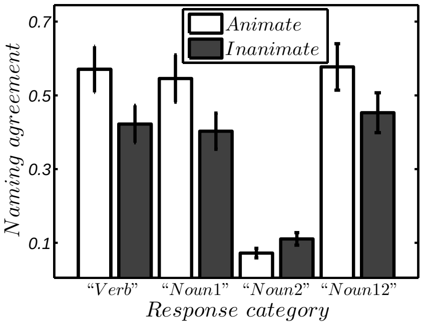

Figure 1: *Experiment 1. Average naming agreement for the different response categories for the animate and inanimate sets. A score of 1.0 indicates perfect naming agreement. Error bars = 95% confidence interval for the average naming agreement.*

A particularly low naming agreement for Noun2 reflects the fact that most participants left this field blank (75% of Noun2 responses; blanks were not permitted in the other fields). For this reason, and because participants were not instructed to treat them differently, the Noun1 and Noun2 responses were considered together in the following stages of the analysis.

Conceptual agreement for the verb and noun categories was quantified with reference to the modal responses for each of the stimuli. If necessary, responses were disambiguated considering together the noun and verb fields (e.g., a noun key associated with the verb typing agreed conceptually with the modal noun keyboard). A non–modal response was scored as in agreement with the modal response if: a synonym of the modal response; a specification of the modal response (e.g., coffee for liquid); a part of the modal response (e.g., open string for guitar); an acoustically plausible coordinate of the modal response (e.g., splashing for lapping); an implication of the modal response (e.g., toothbrush is implied by brushing teeth). A superordinate of the modal response (e.g., metal for keys) was not scored as in agreement with the modal response. It should be noted that these criteria represent a stricter assessment of conceptual agreement than that adopted in previous studies ([24] scored as correct a superordinate and, in general, any "acoustically plausible alternative" of the modal response). The conceptual agreement score was defined as the proportion of participants who gave a response in conceptual agreement with the modal response. A summary of this analysis is presented in Figure 2.

For each of the stimuli, a verbal description was extracted, given by the modal verb and modal noun characterized by the highest conceptual agreement. These verbal descriptions were used as stimuli in Experiment 2, and established the specificity level against which identification performance was assessed. Consequently, a correct identification agreed conceptually, and was at least as specific as the verbal description thus extracted. Also acoustically plausible coordinates were scored as correct, while responses more generic than the reference verbal description were not. The identification performance score was given by the proportion of correct responses for a given stimulus. Average identification performance scores are shown in Figure 2.

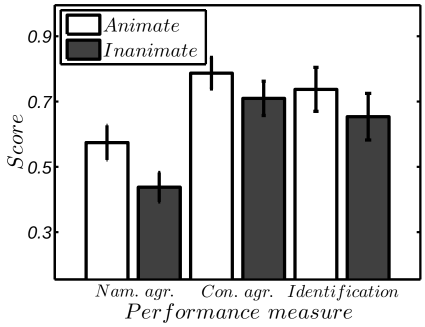

Figure 2: *Experiment 1. Performance measures averaged across stimuli, for the animate and inanimate sound sets. Nam. = naming; Con. = conceptual; agr. = agreement. Error bars: 95% confidence intervals for the average value.*

Associations between the above–defined performance measures were quantified by means of a robust version of the Spearman rank correlation coefficient [27]. Not surprisingly, all measures were strongly and positively correlated (minimum correlation: 0.883; average correlation: 0.897). Also, in agreement with data from [20], identification performance decreased with increasing identification time (Spearman correlation: -0.777).

Significant differences between the animate and inanimate sets in naming and conceptual agreement, and in identification performance were assessed by means of a Wilcoxon rank sum test for equal medians. Naming and conceptual agreement scores were averaged across verb and noun responses. On the one hand, naming and conceptual agreement were significantly higher for the animate than for the inanimate sounds ($p < 0.001$ and $p = 0.032$, respectively). On the other hand, identification performance did not differ between the two sets ($p = 0.069$).

2.3. Discussion

The results of a first free–identification experiment provided data useful for the design of the main experiment of this study. Identifiability of sound events was assessed modifying criteria followed in previous studies (e.g., [24]). In the current experiment, the shorter the identification time the more likely it was that the event was correctly identified. Since this result was highly consistent with those from a previous identification study [20], our measure of identifiability was assumed to be valid.

Animate and inanimate sounds did not significantly differ in either identification performance or in the between–participants agreement for the conceptual content of the verbal identifications. However, naming agreement was significantly higher for the identification of animate than for that of inanimate sounds. This result suggested that a larger vocabulary is available for the verbal identification of inanimate than animate sound events.

3. EXPERIMENT 2: HIERARCHICAL SORTING

Estimation of the similarity of animate and inanimate sound events was investigated in three experimental conditions. In the acoustical and conceptual conditions participants focused on the acoustical properties of the sound stimuli and on the meaning of the verbal descriptions derived from Experiment 1, respectively. In the unbiased condition, participants were instructed to estimate the similarity of sound events, without further specification of the response criteria.

Comparison of condition–specific data allowed addressing two questions. Firstly, whether similarity estimation could selectively focus on the acoustical properties of the sound stimuli, independently of the conceptual correlates of identified sound–generating events. Previous studies were not conclusive on this issue (see Section 1). A novel training technique was therefore designed to aid the adoption of a purely–acoustical similarity estimation criterion, and its effects on the behavioral relevance of conceptual correlates quantified. Secondly, biases in naive listeners for the estimation of the similarity of sound events were characterized. In particular, the comparative relevance of acoustical and conceptual information was quantified. On the basis of previous studies [19], we expected conceptual and acoustical information to predominate for the unbiased estimation of the similarity of animate and inanimate sound events, respectively.

3.1. Methods

3.1.1. Stimuli

Forty animate and forty inanimate sounds were selected from those investigated in Experiment 1. All of them were identified correctly by at least 50% of the participants. Less identifiable sounds were not considered. Indeed, since these sounds were characterized by a lower conceptual agreement (see Section 2.2), a larger misalignment was expected between the conceptual information evoked by verbal descriptions and sound stimuli. As such, estimates of the similarity of verbal descriptions were not granted to measure the conceptual information considered when estimating sounds' similarity. Stimuli selection aimed at equalizing the identifiability distribution in the two sets, and at maximizing the diversity of the identification labels. Identifiability distributions for the animate and inanimate sets were not significantly different, as assessed with an unpaired samples Kolmogorov–Smirnov test ($p = 0.893$, see Figure 3).

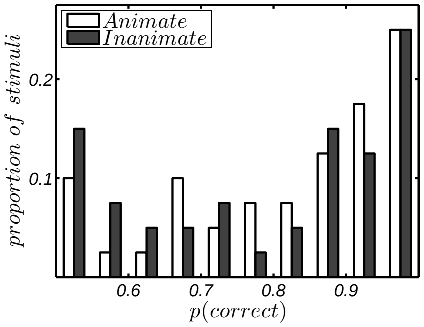

Figure 3: *Experiment 2. Distribution of the identifiability scores for the animate and inanimate sound sets.*

Verbal descriptions for the selected sounds were investigated in the conceptual condition, and were derived from data of Experiment 1. The verb always preceded the noun. The noun was either in the singular or plural form, depending on the most frequent response in Experiment 1.

Ten trial stimuli were (5 animate, 5 inanimate) were finally chosen among the highly identifiable stimuli not selected for the experimental phase. In particular, the sounds, the corresponding verbal descriptions and the manipulated sounds were used during the trial phase of the unbiased, conceptual and acoustical conditions, respectively.

The trial phase for the acoustical condition was meant to train participants in estimating sounds similarity independently of conceptual information. Therefore, signals manipulation aimed at rendering sounds unidentifiable while preserving gross acoustical properties. A method similar to the Event Noise Modulation technique by [28] was used to this purpose. The amplitude envelope $E(t)$ of the original signal $x(t)$ was defined as:

$$E(t) = |x(t) + i\mathrm{H}[x(t)]| \qquad (1)$$

where H is the Hilbert transform [29]. Hearing–range amplitude fluctuations were attenuated by forward–reverse filtering $E(t)$ using a third–order Butterworth filter with a low–pass cutoff frequency of 50 Hz [10]. Further acoustical properties were estimated from the fast Fourier transform (FFT) of the entire signal (Hanning window): the spectral center of gravity SCG, the linear–amplitude weighted average of frequencies from 16 to 16000 Hz.; the spectral mode SM, the frequency of the highest amplitude FFT bin; the lower LL and upper spectral slope UL, given by the slope of the least squares line of the dB spectrum, from the lowest frequency to SM and from SM to the Nyquist frequency, respectively.

A random–phase signal was synthesized, with the same duration as the original signal. Its SM corresponded to that of the original signal, with spectral level decaying linearly in dB, as a function of the distance from the SM. LL_{synth} and UL_{synth} were recursively adjusted, starting from LL_{orig} and UL_{orig}, respectively. At each step of the iterative procedure the SCG of the synthetic signal modulated with the original amplitude envelope was calculated. If SCG_{synth} differed from SCG_{orig} by more than 0.1%, the current LL_{synth} and UL_{synth} values were multiplied by the SCG_{orig}/SCG_{synth} ratio, and by the inverse of this quantity, respectively. If the SCG_{synth} differed from SCG_{orig} by less than 0.1%, the procedure was terminated. Convergence was always reached in less than 28 recursive steps.

3.1.2. Procedure

Participants estimated the similarity of the sounds, or of their acoustical properties, or of the meaning of the corresponding verbal descriptions (unbiased, acoustical and conceptual condition, respectively).

The agglomerative hierarchical sorting technique was used [30]. Participants sorted each of three stimulus sets, one at a time. The sorting involved three sequential phases. Firstly, all the stimuli were initially presented in sequential random order, separated by a pause of 100 ms. Secondly, participants were asked to create a given number of groups out of the available stimuli. They did so dragging randomly numbered on–screen buttons onto one of different rectangles, each representing a single group. They were not allowed to leave empty groups, and could examine each of the stimuli as many times as needed by clicking on the appropriate button. Finally, participants were presented with as many numbered buttons as the groups they created. They were asked to merge the two most similar groups. They could inspect the content of each of the groups as many times as necessary, clicking on the corresponding numbered button. The merging was iterated until only two groups remained to be joined. The sorting task was carried out on the trial set first. Then, half of the participants sorted the animate or the inanimate set, and the remaining set last. During the trial and experimental phases the starting number of groups was 4 and 15, respectively. The entire experiment lasted a maximum of 2 hours and a half.

Stimuli were presented using the same apparatus as for Experiment 1. Signal peak level ranged from 10 to 53 dB SPL. Verbal descriptions were presented at the screen center for a duration of 5 s, approximately the average duration of the sound stimuli.

3.1.3. Participants

Sixty native English speakers took part in the experiment (42 females, 18 males; age: 17–37; mean age: 21.23; 16 non–musicians, 25 amateur musicians, 19 professional musicians). None of them had participated to Experiment 1. An equal number of participants was assigned to each of the experimental conditions. Half of the participants evaluated the animate set first, the inanimate set second and vice versa. Normal hearing for participants in the acoustical and conceptual conditions was assessed using the same methodology as in Experiment 1. All participants reported having normal hearing and normal or corrected–to–normal vision.

3.2. Results

Agglomerative hierarchical sorting data can be assumed to yield an ordinal estimate of between–stimuli dissimilarity, given by the step of the merging procedure when two stimuli are grouped first. Plausibly, similar stimuli are grouped earlier than dissimilar stimuli. A full pairwise dissimilarity matrix was therefore collected from each of the participants for each of the stimulus sets.

Data modeling focused on the agreement among condition–specific sortings, measured by means of the correlation between the median dissimilarity matrices from the different conditions. Agreement was quantified using the robust Spearman rank correlation [27]. Analysis focused on group rather than individual data since the former were much more reliable than the latter. For example, while the highest between–groups correlation was 0.85, the average correlation between data of participants in the same two groups was 0.31 (range: −0.03 – 0.99). This discrepancy was probably caused by participant–specific sorting criteria at times independent of perceived similarity. Nonetheless, these idiosyncratic influences (e.g., influence of the buttons numberings on the grouping decisions) were likely attenuated in the group data.

Statistical modeling aimed at testing for an influence of the experimental factors, and of their interactions, on the level of agreement between condition–specific data. An *ad hoc* data manipulation was carried so as to introduce a font of between–participants variability, thus allowing the adoption of the repeated measures ANOVA framework. The same between–groups correlation was then computed leaving out the data of one of the participants at a time. Thus, one vector of between–conditions correlations was computed for each of the participants. Figure 4 summarizes the data relevant to the following of this paper.

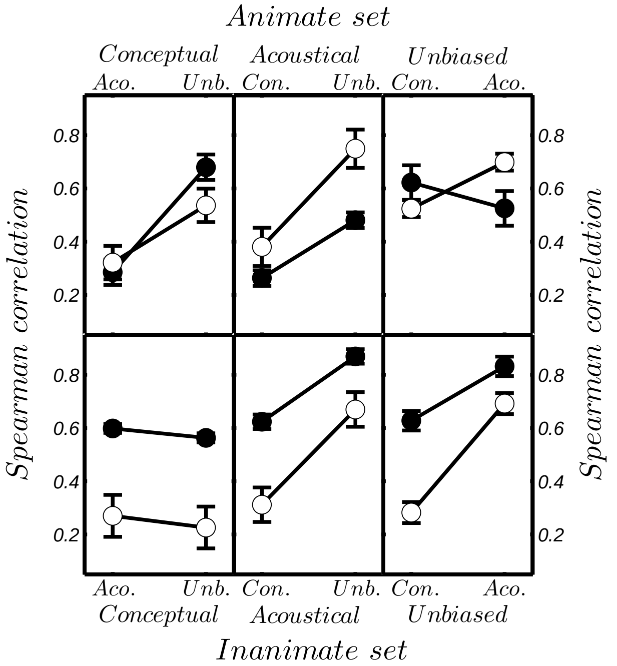

Figure 4: *Experiment 2. Spearman rank correlation between median condition–specific sortings. Data points show the average of the correlation coefficients computed leaving out the data of one participant at a time (error bars = ±1 SD). Upper panels: animate set; lower panels; inanimate set. The abscissa label shows the condition factor constant within each of the panels (Con. = conceptual; Aco. = acoustical; Unb. = unbiased). Order factor: white and black (animate before inanimate and vice versa).*

A resampling–based variant of the repeated measures ANOVA model was used, the bootstrap–F technique, resistant to violations of the data–normality and sphericity assumptions (e.g., [31]).

Three repeated measures ANOVA model were computed, considering either the conceptual, the unbiased, or the acoustical condition as a reference (left, middle and right panels of Figure 4, respectively). The dependent variables were the above–defined participant–specific correlations. The model for the reference–acoustical condition (middle panels of Figure 4) did not significantly add to the overall conclusions and is only briefly mentioned here. Each of the models had two within–subjects factors: comparison condition (e.g., acoustical and unbiased in the model for the reference–conceptual condition) and set type (animate vs. inanimate). The order of presentation of the sets (first animate or inanimate) was a between–subjects factor. Significance tests were computed from 10000 bootstrap replicates, drawn independently within different levels of the between–subjects factor.

All the effects in both the conceptual and unbiased model were significant ($p \leq 0.039$). Further contrasts were investigated computing a separate ANOVA model for the animate and inanimate sets (top and bottom panels of Figure 4), or using two–samples bootstrap hypothesis tests (bootstrap replicates = 10000) [32].

The conceptual model compared the extent to which participants in the acoustical and unbiased conditions made use of conceptual information. With animate sounds (left–top panel in Figure 4), conceptual information was less relevant in the acoustical than in the unbiased condition, independently of whether they were evaluated before or after the inanimate sounds ($p < 0.001$). Also, when the animate sounds were evaluated second, the relevance of the conceptual information decreased for the unbiased, but not for the acoustical condition ($p < 0.001$ and $p = 0.171$, respectively). When inanimate sounds (left–bottom panel in Figure 4) were presented first, conceptual information was equally relevant in the acoustical and unbiased conditions ($p = 0.132$); when they were evaluated after the animate sounds, conceptual information was slightly more relevant in the acoustical than in the unbiased condition ($p = 0.001$). Most notably, the relevance of conceptual information was higher when the inanimate set was evaluated after the animate set, independently of the condition ($p < 0.001$). Finally, when animate or inanimate sounds were evaluated first, the weight of the conceptual information in the acoustical condition was constant and extremely low ($p = 0.596$, average Spearman correlation ≤ 0.29). However, even when each of the sets was evaluated first, the weight of the conceptual information to similarity estimation in the acoustical condition was always significantly higher than zero, as measured by the p–value for the correlation among median group data ($p < 0.001$ for both the animate and inanimate sets).

The acoustical model (middle panels in Figure 4) revealed that, independently of the set type and order, the relevance of acoustical information was invariably higher for participants in the unbiased condition than for those in the conceptual condition ($p < 0.001$).

The unbiased model compared the relevance of conceptual and acoustical information for participants in the unbiased condition. For animate sounds (right–top panel of Figure 4), conceptual information was more relevant than acoustical information when they were evaluated first, while the opposite was true when they were evaluated after the inanimate set ($p \leq 0.002$). For inanimate sounds (right–bottom panel in Figure 4), acoustical information was always more relevant than conceptual information, independently of the order in which the sound sets were evaluated ($p < 0.001$). Interestingly, when inanimate sounds were evaluated second, both the acoustical and the conceptual information increased in relevance ($p < 0.001$).

3.3. Discussion and conclusions

The agglomerative hierarchical sorting technique was used to measure the estimation of the similarity of large sets of animate and inanimate sound events, and of their identification labels, in different conditions.

Clear condition effects were observed when the animate or inanimate sets were evaluated first. Firstly, the relevance of conceptual information to similarity estimation in the acoustical condition was extremely low, for both animate and inanimate sounds. Therefore, consistently with results by [14], a selective focus on acoustical information was possible, independent of the knowledge structures activated by the recognition of the sound–generating event. Furthermore, the weight of conceptual information in the acoustical conditions never approached zero. Plausibly, a perfect independence of the judgment of conceptual and acoustical correlates of sound events can never be observed, since events activating similar knowledge structures likely share some acoustical structure (e.g., most of the sounds recognized as insects sounds likely share some acoustical properties). Interestingly, the training and instructions in the acoustical condition significantly decreased the relevance of conceptual information only for animate sounds, while with inanimate sounds the relevance of conceptual information was already extremely low in the unbiased condition.

A second result emerged from those conditions where sound sets were judged first: consistently with data by [19], unbiased similarity estimation of animate and inanimate sounds focused on conceptual and acoustical information, respectively. This result is also consistent with a preferential activation of the Brodmann's area 22 in response to animal sounds [17], where this cortical area is home to Wernicke's area, believed to underlie language comprehension. This result is however at odds with data by [12], where unbiased similarity estimation was strongly correlated with estimation of the similarity of the knowledge structures activated by the identification labels. As pointed out in Section 1, this likely resulted from a focus on the animate–inanimate distinction in all of the experimental conditions investigated by [12]. Finally, it should be noted that a strong perceptual relevance of conceptual information disagrees with the assumption of the ecological approach according to which the object of perception is the sound–generating event (cf. [3]). A focus on acoustical properties is instead consistent with this assumption, since it is not farfetched to state that the structure of a sound is determined by the mechanics of the sound–generating event [3]. In light of the above–mentioned results it can be concluded that the ecological approach accurately explains perception of inanimate sounds, but not of animate sounds.

Results differed when the sound sets were evaluated second. Firstly, the relevance of conceptual information to similarity estimation in the acoustical condition remained constant and increased for animate and inanimate sounds, respectively. As a consequence, an unbiased characterization of the acoustical correlates for the perception of animate sounds is likely obtained independently of previously judged sounds. The same goal can be reached with inanimate sounds if they are judged before animate sounds. In the unbiased condition, a general tendency was observed for judgment criteria for the second set to resemble those adopted for the first set. However, this tendency characterized the conceptual but not the acoustical information. Indeed, the weight of conceptual information increased and decreased for the inanimate and animate sounds, respectively, while the weight of acoustical information increased for both sound sets. An explanation for this result is that while a focus on conceptual information can be controlled with relative ease by a listener, the same is not possible for acoustical information. This result would support the hypothesis of a perceptual primacy of a listening mode based on acoustics rather than on higher–level knowledge, for both animate and inanimate sounds. Given the likely lawful relationship between sound source mechanics and acoustical structure, this hypothesis would be highly consistent with the assumptions of the ecological approach.

In conclusion, two main points emerged from this study. Firstly, accurate characterization of the acoustical correlates of the perception of environmental sound is possible, following a few simple methodological guidelines. Secondly, naive listeners estimate the similarity of animate and inanimate sounds focusing on the higher–level knowledge activated by the recognition of the sound–generating event and on the acoustical structure, respectively. In other words, everyday we hear animate concepts and inanimate sound sources.

4. REFERENCES

[1] W. W. Gaver, "What in the world do we hear? An ecological approach to auditory event perception," *Ecological Psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[2] C. Carello, J. B. Wagman, and M. T. Turvey, "Acoustical specification of object properties," in *Moving image theory: Ecological considerations*, J. Anderson and B. Anderson, Eds. Carbondale, IL: Southern Illinois University Press, 2003.

[3] C. F. Michaels and C. Carello, *Direct perception*, Prentice-Hall, Englewood Cliffs, New Jersey, 1981.

[4] S. McAdams, "Recognition of sound sources and events," in *Thinking in sound: the cognitive psychology of human audition*, S. McAdams and E. Bigand, Eds., pp. 146–198. Oxford University Press, 1993.

[5] R. Smits, J. Sereno, and A. Jongman, "Categorization of sounds," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 32, no. 3, pp. 733–754, 2006.

[6] T. J. Palmeri, Alan C-N. Wond, and Isabel Gauthier, "Computational approaches to the development of perceptual expertise," *Trends in Cognitive Sciences*, vol. 8, no. 8, pp. 378–386, 2004.

[7] H. Cohen and C. Lefebvre, Eds., *Handbook of Categorization in Cognitive Science*, Elsevier, Oxford, UK, 2005.

[8] B. L. Giordano and S. McAdams, "Material identification of real impact sounds: effects of size variation in steel, glass, wood and plexiglass plates," *J. Acoust. Soc. Am.*, vol. 119, no. 2, pp. 1171–1181, 2006.

[9] A. J. Kunkler-Peck and M. T. Turvey, "Hearing shape," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 26, no. 1, pp. 279–294, 2000.

[10] S. McAdams, A. Chaigne, and V. Roussarie, "The psychomecanics of simulated sound sources: Material properties of impacted bars," *J. Acoust. Soc. Am.*, vol. 115, no. 3, pp. 1306–1320, 2004.

[11] R. L. Goldstone, "The role of similarity in categorization: providing a framework," *Cognition*, vol. 52, pp. 125–157, 1994.

[12] B. Gygi, G. R. Kidd, and C. S. Watson, "Similarity and categorization of environmental sounds," *Perception & Psychophysics*, in press.

[13] B. L. Giordano, *Sound source perception in impact sounds*, Ph.D. thesis, University of Padova, Italy, 2005.

[14] G. P. Scavone, S. Lakatos, P. Cook, and C. R. Harbke, "Perceptual spaces for sound effects obtained with an interactive similarity rating program," in *Proceedings of the International Symposium on Musical Acoustics*, September 2001.

[15] P. Iverson and C. L. Krumhansl, "Isolating the dynamic attributes of musical timbre," *J. Acoust. Soc. Am.*, vol. 94, no. 5, pp. 2595–2603, 1993.

[16] J. M. Hajda, R. A. Kendall, E. C. Carterette, and M. L. Harshberger, "Methodological issues in timbre research," in *The Perception and Cognition of Music*, I. Deliege and J. Sloboda, Eds., pp. 253–306. L. Erlbaum, London, 1997.

[17] J. W. Lewis, J. A. Brefczynski, R. E. Phinney, J. J. Jannik, and E. D. DeYoe, "Distinct cortical pathways for processing tool versus animal sounds," *The Journal of Neuroscience*, vol. 25, no. 21, pp. 5148–5158, 2005.

[18] M. M. Murray, C. Camen, S. L. Gonzalez Andino, P. Bovet, and S. Clarke, "Rapid brain discrimination of sounds of objects," *The Journal of Neuroscience*, vol. 26, no. 4, pp. 1293–1302, 2006.

[19] Y. Gèrard, *Mémoire sémantique et sons de l'environnement*, Ph.D. thesis, Université Paris 6, France, 2005.

[20] J. A. Ballas, "Common factors in the identification of an assortment of brief everyday sounds," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 19, no. 2, pp. 250–267, 1993.

[21] L. Elliott, *A guide to wildlife sounds*, Stackpole Books, Mechanisburg, PA, 2005.

[22] F. Opolko and J. Wapnick, *McGill University Master Samples [Compact Disc]*, McGill University, Montréal, Québec, 1987.

[23] E. M. von Hornbostel and C. Sachs, "Systematik der musikinstrumente. Ein versuch," *Zeitschrift für Ethnologie*, vol. 4-5, pp. 553–590, 1914.

[24] M. E. Marcell, D. Borella, M. Greene, E. Kerr, and S. Rogers, "Confrontation naming of environmental sounds," *Journal of Clinical and Experimental Neuropsychology*, vol. 22, no. 6, pp. 830–864, 2000.

[25] ISO 389-8, "Acoustics – reference zero for the calibration of audiometric equipment – Part 8: Reference equivalent threshold sound pressure levels for pure tones and circumaural earphones," Tech. Rep., International Organization for Standardization, Geneva, 2004.

[26] F. N. Martin and C. A. Champlin, "Reconsidering the limits of normal hearing," *Journal of the American Academy of Audiology*, vol. 11, no. 2, pp. 64–66, 2000.

[27] S. Verboven and M. Hubert, "LIBRA: a Matlab library for robust analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 75, pp. 127–136, 2005.

[28] B. Gygi, G. R. Kidd, and C. S. Watson, "Spectral-temporal factors in the identification of environmental sounds," *J. Acoust. Soc. Am.*, vol. 115, no. 3, pp. 1252–1265, March 2004.

[29] W. M. Hartmann, *Signals, Sound and Sensation*, AIP Press, Woodbury, NY, 1997.

[30] A. P. M. Coxon, *Sorting data: Collection and analysis*, Sage University Papers on Quantitative Applications in the Social Sciences. Sage Publications, Thousand Oaks, CA, 1999.

[31] I. Berkovits, G.R. Hancock, and J. Nevitt, "Bootstrap Resampling Approaches for Repeated Measure Designs: Relative Robustness to Sphericity and Normality Violations," *Educational and Psychological Measurement*, vol. 60, no. 6, pp. 877–892, 2000.

[32] B. Efron and R. J. Tibishirani, *An introduction to the bootstrap*, Chapman & Hall, New York, NY, 1993.