ARL-TR-8672 ● MAR 2019

DEVCOM
ARMY RESEARCH
LABORATORY

Characterizing Asymmetrical Ratings of Similarity for Real-World Complex Environmental Sounds

by Brandon S Perelman, Kelly Dickerson, and Jeremy Gaston

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer’s or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-8672 ● Mar 2019

Characterizing Asymmetrical Ratings of Similarity for Real-World Complex Environmental Sounds

by Brandon S Perelman
Oak Ridge Associate Universities

Kelly Dickerson, and Jeremy Gaston
Human Research and Engineering Directorate,
CCDC Army Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* March 2019	**2. REPORT TYPE** Technical Report	**3. DATES COVERED (From - To)** 1 October 2015–28 September 2016

4. TITLE AND SUBTITLE
Characterizing Asymmetrical Ratings of Similarity for Real-World Complex Environmental Sounds

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Brandon Perelman, Kelly Dickerson, and Jeremy Gaston

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Combat Capabilities Development Command
Army Research Laboratory*
ATTN: FCDD-RLH-FD
Aberdeen Proving Ground, MD 21005

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-8672

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
*The work outlined in this report was performed while the US Army Research Laboratory (ARL) was part of the US Army Research, Development, and Engineering Command (RDECOM). As of 1 February 2019, the organization is part of the US Army Combat Capabilities Development Command (formerly RDECOM) and is now called CCDC Army Research Laboratory.

14. ABSTRACT

Understanding complex environmental sound perception is critical for understanding human behavior in real-world settings. Whereas simple stimuli can be processed on the basis of physical characteristics alone, environmental sounds contain semantic and contextual information, which can lead to asymmetrical similarity ratings. The goals of this study were to 1) quantify asymmetries in pairwise similarity ratings of 25 environmental sounds, 2) test the hypothesis that these asymmetries are systematic order effects, and 3) characterize the impact of asymmetries on similarity spaces constructed using multidimensional scaling (MDS). First, 26 participants rated the similarity of every pair of the sounds (from 1 = most similar, to 7 = as different as possible) in both orders. In the second experiment, participants were asked whether they could identify the source of each sound. The relative identifiability of each sound in the pair influenced rated similarity; presenting the more identifiable sound in the base position (second) produced higher-rated similarity. MDS spaces constructed using randomly assigned presentation orders were highly intercorrelated, whereas MDS spaces generated from pairs ordered by identifiability (i.e., higher to lower vs. lower to higher identifiability) shared roughly 10% less variance, indicating that randomization is effective for controlling for order effects in pairwise similarity ratings.

15. SUBJECT TERMS
auditory perception, similarity, real-world sounds, multidimensional scaling, identifiability

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Brandon Perelman
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	31	**19b. TELEPHONE NUMBER (Include area code)** (410) 278-3587

Contents

List of Figures

List of Tables

1. Introduction

This report describes a subset of techniques developed to describe the role of semantic information in understanding the auditory environment. Auditory situation awareness is important in the dismounted Soldier operational environment, which is a complex and dynamic soundscape that can change at a moment's notice. These changes could be simple distractors or hold some operational relevance. To make a determination of relevance on the fly, Soldiers must localize and identify the source of the sound-producing event. This cannot be achieved using acoustic information alone, but rather requires the integration of bottom-up (acoustic) and top-down (semantic, e.g., context and experience) information.

Acoustic information, defined as the physical stimulus features, can be objectively quantified relatively easily; however, quantifying semantic content and determining the stability and thus applicability of that information across multiple individuals has been more difficult. To address this gap in the available literature, we have conducted several studies to establish the semantic value of everyday sounds for improving performance on auditory situation-awareness tasks. This report represents just one of those characterization efforts (Dickerson et al. 2015a, 2016; McArdle et al. 2017; Dickerson et al. 2018) that document the other techniques that have been applied to this stimulus set. By pairing these types of subjective stimulus evaluations with objective measures of performance, we can quantify changes in auditory situation awareness that cannot be accounted for by acoustic measures alone.

2. Background

Environmental sounds are defined here as signals produced by humans, animals, and objects that make up the auditory ambiance of the everyday environment. Everyday listening, conceptualized by Truax (1996) is not focused listening but rather distracted listening, and unless a sound source suddenly changes or becomes relevant to the listener's current task, it is likely to go unnoticed. Sounds enter and exit awareness and are only considered thoughtfully if the listener has a reason to shift attentional resources to a particular aspect of the ongoing environmental scene. Salience and context (as well as other semantic attributes) provide information to the listener that can support situation awareness. Understanding the meaning of a particular sound heard during everyday listening is not a discrete event because of the complexity of the soundscape and the changing task demands of the listener. Sound is perceived in a rich semantic and temporal context. For example, a rhythmic crunching sound followed by a dull thump might indicate that someone has kicked a soccer ball, but if the temporal order was reversed, the listener may come to a different conclusion, perhaps inferring that an item had been dropped or broken. Thus, in the dynamic flow of sound-producing events, a particular signal's temporal relationship to other recently heard sounds may be critical for correctly identifying the sound source and its meaning in context. One way that this phenomenon manifests is change deafness, an inability to detect relevant changes in environmental sound scenes. Gregg and Samuel (2009) found a significant role of semantic similarity in producing change deafness, with participants exhibiting more errors when a sound was swapped with a semantically related distractor than when the distractor was acoustically similar but semantically unrelated.

Studies using temporal scrambling manipulations highlight the importance of acoustic information in signal recognition, particularly the importance of sound onset information. However, these types of studies examine sounds in isolation and do not address how the temporal ordering of sounds in a multi-sound context (as is the case in the everyday environment) affects perceptual performance. Pairwise comparison studies are one potential methodology for examining how temporal order influences the perceived link between two environmental sounds. Typically, pairwise rating studies present two signals for consideration (*A, B*), separated by a brief interstimulus interval, the listener is then asked to rate the two signals along a single dimension. The most common example of auditory pairwise ratings studies involve rating items for similarity (Klatzky et al. 2000; Bonebright 2001; Terasawa et al. 2005; Gygi et al. 2007; Lemaitre et al. 2007, 2010; Aldrichet et al. 2009; Misdariis et al. 2010; Gaston and Letowski 2012; Dickerson et al. 2015b). In pairwise ratings studies, every sound is paired with itself and every other signal in the set. This enables researchers to evaluate how the dimension of interest varies across all of the items in the set. For example, Gygi et al. (2007) examined pairwise similarity for a set of 50 sounds. They asked participants to rate similarity but did not specifically define similarity (i.e., differentiate between acoustic vs. semantic similarity) for the participants. The multi-dimensional scaling (MDS) space generated from these ratings demonstrated that in the absence of instructions to focus on either acoustic or semantic dimensions of the stimuli, participants group sounds together based on a combination of acoustic and semantic features.

To obtain this result, which has inspired substantial research in the realm of environmental sound perception, Gygi et al. collected 10,000 trials per participant and averaged across the diagonal of the similarity ratings matrix to obtain symmetrical similarity ratings. Subsequent studies with similar goals have collected only comparisons for a single order in an effort to reduce experiment duration (e.g., Klatzky et al. 2000; Lemaitre et al. 2007; Parizet and Koehl 2012) or have not specifically discussed this part of the procedure. Thus, these studies have generally assumed equivalency between stimulus presentation orders. This assumed equivalency potentially masks asymmetries in the similarity judgments that could be informative in addressing the question posed at the opening of this report: Do temporal order effects influence perceptual performance in multi-sound contexts? An asymmetry in pairwise ratings is very much like a traditional order effect. Order effects have been observed across a variety of different stimulus judgment domains; for example, color (Nosofsky 1991), line perception (Garner and Haun 1978) and analogical reasoning, where transfer is more successful when content is presented *A, B* instead of *B, A* (e.g., Tversky 1977; Gentner 1980, 1983, 1989; Bowdle and Gentner 1997).

The present study uses a pairwise similarity ratings task to probe for ratings asymmetries in a set of 25 environmental sounds to attempt to uncover potential patterns of related items in a seemingly disparate set of items. The benefit of this approach is that, much like the everyday environment, there are not always clear, a priori defined stimuli or conceptual links among stimuli. By looking at similarity and asymmetry, we can 1) observe the categorical structure of the sound set—related sounds should group together, and 2) examine if the concepts produced by the similarity grouping are meaningful in a way that is affected by temporal order, which will be apparent by the magnitude of the asymmetry for a pair of items.

3. Experiment 1

The goal of Experiment 1 was to present the full set of all possible pairwise comparison to participants in order to obtain all *AB* and *BA* ratings for a set of 25 common environmental sounds. These data were used to determine the presence and magnitude of asymmetries in the ratings. To compute asymmetries, we used the procedures described in Holman's (1979) *additive similarity and bias model*. In this model, asymmetries are quantified as bias, where an observer is more likely to rate one directional pairing as more similar (e.g., order *AB*) than when that pairing is presented in the reverse order (order *BA*). The measure of bias is obtained by subtracting *AB* and *BA* similarity ratings; a positive value indicates that the two stimuli are rated as more similar when presented in order *AB*, whereas a negative value indicates that the items are rated as more similar when presented in the reverse order. Importantly, and in contrast to prior work, we propose that the bias values can be interpreted in two ways: 1) as the raw bias values (both positive and negative), reflecting the *directional magnitude* of the bias toward one order over the other for each pair of stimuli, or 2) the absolute value of the bias measure, which would reflect the *absolute magnitude* of the overall asymmetry between ratings involving those two stimuli. Accordingly, the directional magnitude should be used to test directional hypotheses regarding order effects, while the absolute magnitude should be used when testing hypotheses designed to compare asymmetries among groups of stimuli. In this experiment, we expected that participants would produce asymmetrical similarity ratings across the stimuli; mathematically, this would manifest such that the bias directional magnitudes would be nonzero (consistent with Siedenburg et al. 2016), and the absolute magnitudes would be greater than 0. This experiment tested the hypothesis that order effects do in fact produce asymmetrical similarity ratings for environmental sounds, and an affirmative result would indicate the necessity of explaining these effects.

3.1 Method

3.1.1 Participants

Twenty-six participants were recruited from the Binghamton University undergraduate student population. Students enrolled in introductory psychology courses are given the option of volunteering in research studies to satisfy a course requirement. In every case, participants provided informed consent prior to completing the experimental tasks. Participants were each given a hearing screening during which they were seated in a sound-attenuated booth. The screening was performed using a Grason-Stadler Inc. Arrow portable audiometer and Telephonics TDH-50 headphones. Normal hearing is defined in this study as pure-tone air-conduction hearing thresholds better than 20-dB hearing level for octave audiometric frequencies from 250 to 8000 Hz (250, 500, 1000, 2000, 4000, 8000 Hz), inclusively (ANSI S3.6 2010), and as an absence of a self-reported history of otologic pathology (i.e., diagnosed hearing loss). All participants had normal hearing.

3.1.2 Stimuli and Apparatus

The stimulus set used consisted of 25 sounds from sources that generally represent those found in an urban environment, such as vehicles, animals, and machinery (Table 1). The majority of the sounds were downloaded from the website freesound.org, an online user-supported sound library. All sounds were truncated to 1000 ms with 5-ms linear on and offset ramps. Root-mean-square amplitude was used to normalize the sound amplitudes. All of the sound editing was done using Adobe Audition v.6. Each trial consisted of one pair of these sounds. Stimulus pairs were generated using all possible pairs of sounds presented in both orders (i.e., *AB* and *BA*) and with themselves (i.e., *AA*) to test participants' attention and motivation to the task (these pairs should be rated as similar as possible). The full stimulus set consisted of 25^2 = 625 comparisons. The stimuli were presented using custom software created and presented to participants in Psychology Software Inc. E-Prime 2.0 on a standard PC. All sounds were presented over Beyerdynamics T70 circumaural reference headphones at a comfortable listening level corresponding to 70 dB(C).

Table 1 Stimulus list and verbal descriptions of each auditory stimulus

No.	Stimulus name	Stimulus description
1	Tank1	Tank in motion, with both the engine and tread sounds audible
2	Shopvac1	Shop vacuum powering up and running
3	Shopvac2	Shop vacuum running continuously
4	Truck1	Truck engine idling; idling noise characterized by a distinct rattle
5	Truck2	Truck in motion, moving away from the listener
6	Truck3	Truck engine idling; a very smooth idle with an intermittent squeak
7	Plane1	High-pitched scorching sound of a jet engine passing by
8	Plane2	Droning sound of a jet at high altitude
9	Plane3	Low hum of a propeller plane traveling away from the listener
10	Motorcycle2	Ramping buzz of a sport motorcycle approaching the listener
11	Helicopter1	Ramping engine and propeller noise of an approaching helicopter
12	Helicopter2	Consistent engine and propeller noise of a hovering helicopter
13	Jackhammer1	Metallic chatter of a jackhammer against concrete
14	Bus1	Bus idling, accompanied by the high-pitched squeal of air brakes
15	Bus2	Bus moving away from the listener, accompanied by an intermittent squeak
16	Bike1	Rattling sound of a bike’s wheel turning
17	Bike2	Rattling sound of a bike’s wheels and pedals turning
18	Cell1	Cellular phone ring tone that sounds similar to a xylophone
19	Bell1	A hand bell being swung back and forth twice
20	Bell2	A very small bell jingling
21	Crickets1	Continuous chirping on a background of night insect noise
22	Crickets2	A pair of soft chirps on a background of night insect noise
23	Dog1	A small dog barking
24	Dog3	A small dog shaking off
25	Walking2	Two loud footsteps of someone wearing shoes

3.2 Procedure

In each trial of the experimental task, participants listened to the two sounds of the stimulus pair presented sequentially, separated by a 750-ms interstimulus interval, then rated their similarity on a scale (Fig. 1) of 1 (as similar as possible) to 7 (as different as possible using a keyboard). This rating scheme is used specifically because the subsequent analyses require a measure of distance, but the term *similarity* is retained to reduce confusion. If the participant had not responded after 5 s, the software automatically advanced to the next trial. All 625 trials took roughly 1 h to complete.

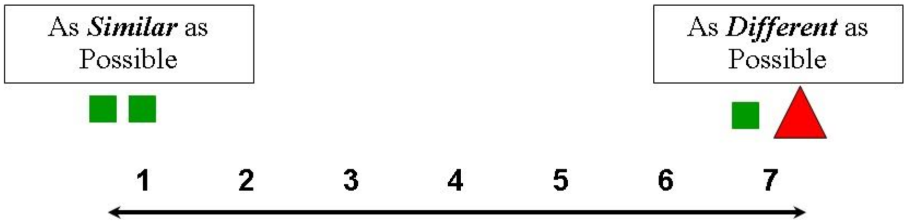


Fig. 1 Likert scale used for rating each stimulus pair. Participants were instructed to use the entire range, rating each stimulus from 1 = most similar, to 7 = most dissimilar.

4. Results and Discussion

The use of MDS to explicitly examine asymmetries in pairwise comparisons is fairly novel. Due to this novelty and the authors' goal to document the methods used in this study to encourage replication, a detailed description and justification of the analysis methodology is provided in this section. As described, this experiment requires calculating asymmetries (i.e., bias values) between pairwise similarity ratings for all possible stimulus pairs. To accomplish this, we first transformed the data to submit the raw similarity ratings to MDS analysis. MDS algorithms require *distance* values that satisfy the assumption of minimality (i.e., the distance between an item and itself should be zero; nonmetric MDS algorithms are robust to violations of triangle inequality, however). The following steps describe the quantitative approach taken to data transformation in detail.

First, similarity ratings for all possible *AB* and *BA* stimulus pairings were averaged across participants to populate each cell of an $n * n$ similarity matrix, where n is the number of stimuli and cell ij receives the mean similarity rating between stimuli A and B, whereas cell ji contains the mean similarity rating for those two stimuli presented in the opposite order. These values are transformed to distances by subtracting the 1 from their ratings (yielding values of 0 when the items are perceived as the same, and 6 when they are perceived as maximally different), then applying a max scale (i.e., dividing by 6) to produce values between 0 and 1.

Second, to compute the asymmetries, we applied Holman's (1979) additive similarity and bias model, whereby that similarity matrix can be decomposed into a symmetrical similarity matrix plus a bias matrix reflecting those asymmetries. We operationally defined asymmetry as $Asymmetry_{ij} = S_{ij} - S_{ji}$, where S indicates the similarity judgment for a particular participant on a particular stimulus. These asymmetry values are then used to populate the skew-symmetrical bias matrix. Perfect symmetry produces a value of 0, while the metric is sensitive to directionality. This particular operationalization of asymmetry is common in the literature (Nosofsky 1991; Johannesson 1997; Siedenburg et al. 2016). Note that the current dataset handles similarity data rather than dissimilarity data; thus, similarity will be calculated as $1 - d(ij)$ so that positive asymmetry values equate to a higher perceived similarity when stimuli are presented in the order *AB* than when presented in the order *BA*.

5. Similarity Space Construction

To characterize the underlying similarity relationships among the environmental sound stimuli used in this experiment (see Table 1), a 2-D MDS solution was constructed (Fig. 2) using Kruskal's nonmetric MDS algorithm (NMDS, via the isoMDS function in the MASS package) (Venables and Ripley 2002). Importantly, NMDS averages across the similarity matrix across the diagonal, which is one way to account for asymmetries in the similarity ratings. Flexible mixture modeling (FMM) (Leisch 2004) was used to cluster the resulting solution. In this case, a driver composed of a custom mixture of Gaussians was created to allow the models to vary in terms of shape, volume, and orientation. A stepwise procedure allowed us to test clustered solutions containing a variable number of Gaussian models ($k = 2{:}7$), and the best fit was provided by a two-cluster model as determined by the Bayesian information criterion (BIC = –5.82), which rewards goodness of fit while penalizing model complexity (Fig. 2).

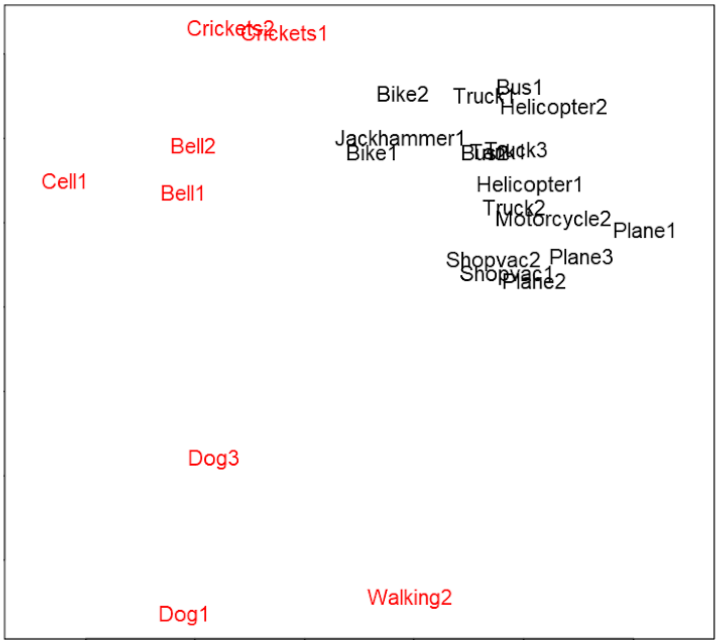


Fig. 2 Clustered MDS solution for the stimulus set, color-coded according to cluster membership. FMM identified two clusters of stimuli, one containing the mechanical sounds (*n* = 17) and another containing the remainder of the stimuli (nonmechanical sounds: *n* = 8).

5.1 Pairwise Similarity Ratings and Similarity Spaces

In general, participants' scaled ratings (refer to Section 4 for scaling details) were negatively skewed (i.e., toward dissimilar; skewness = –0.97, $M_{scaled\ rating}$ = 0.52, and $SD_{scaled\ ating}$ = 0.18) (Fig. 3). This type of skew has been reported previously for this sound set (Dickerson et al. 2015) but has also been observed by others for similar types of sounds (Gygi et al. 2007; though these authors report a positive skew, the Likert scales used in that study and our own are inverted). This likely reflects the fact that the majority of the sounds in the set used here could co-occur in a single environmental context (i.e., city sounds). Furthermore, this may be a general property of stimulus pools containing a broad variety of semantically loaded stimuli: What is more similar to the sound of a jackhammer, a bell ringing or a dog barking? The histogram in Fig. 3 illustrates this skew but also provides a qualitative look at the extent of asymmetry present across the entire sound set. An example to orient the reader to Fig. 3 can be found by careful inspection of the Bus2-Cricket1 pairing. The stimulus pair Bus2-Crickets1 was rated as more similar when the stimulus Crickets1 was presented first (scaled similarity rating = 0.65) than when it was presented second (scaled similarity rating = 0.43).

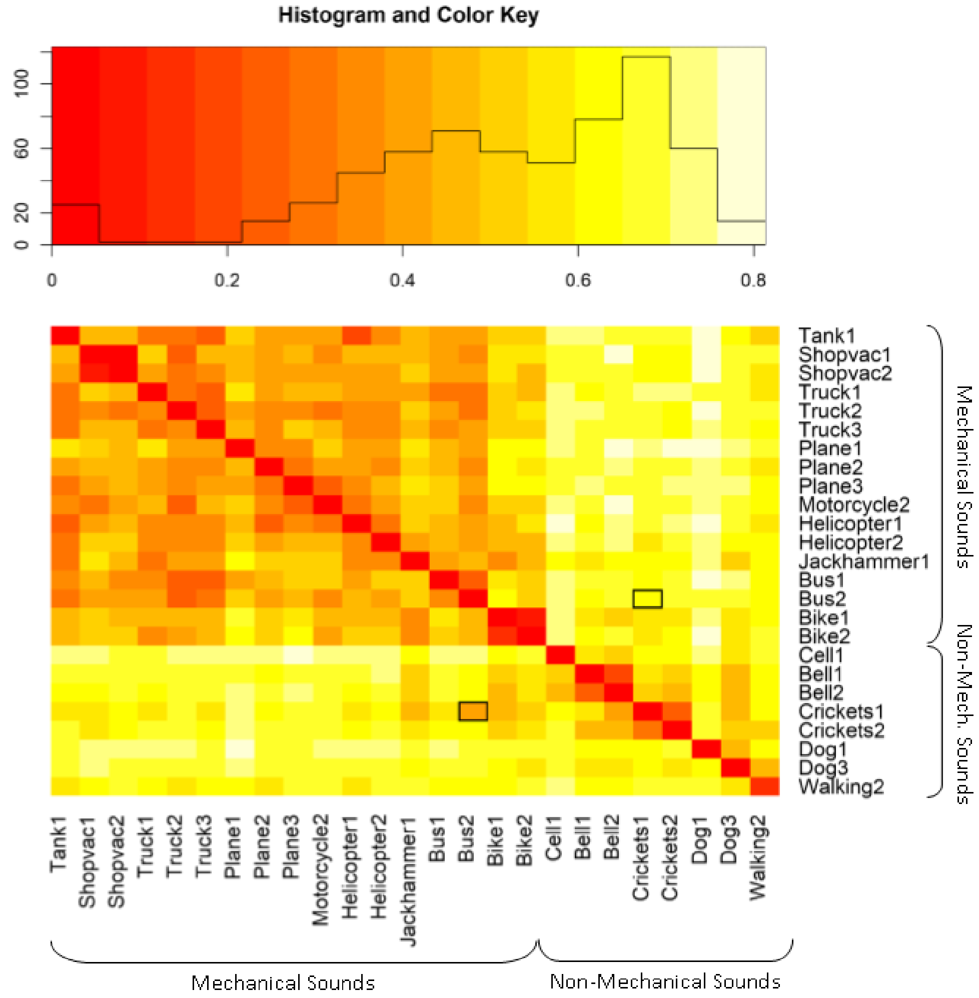


Fig. 3 Scaled similarity ratings represented in a color-coded histogram provide a qualitative look at the asymmetries present across the entire sound set. White indicates higher rated similarity. Columns (first sound) and rows (second sound) are ordered by stimulus cluster membership, but ordering within each cluster is arbitrary.

Using the procedure described in Section 4, a bias matrix was produced using the *ij* and *ji* ratings of the stimuli. In Fig. 4 the mean (scaled) absolute magnitude of the asymmetry values calculated for each stimulus is plotted. These values were tested against the null hypothesis that pairwise ratings would be the same, independent of the presentation order (i.e., $AB = BA$), resulting in an absolute magnitude value of zero. Bonferroni-corrected one-sample t-tests were conducted on the absolute magnitude of the asymmetry values; all comparisons were significantly greater than zero (all $p < 0.001$). Note that absolute magnitudes are shown in place of directional magnitudes since averaging negative and positive values in the directional magnitudes does not provide a meaningful statistic for characterizing the stimulus set.

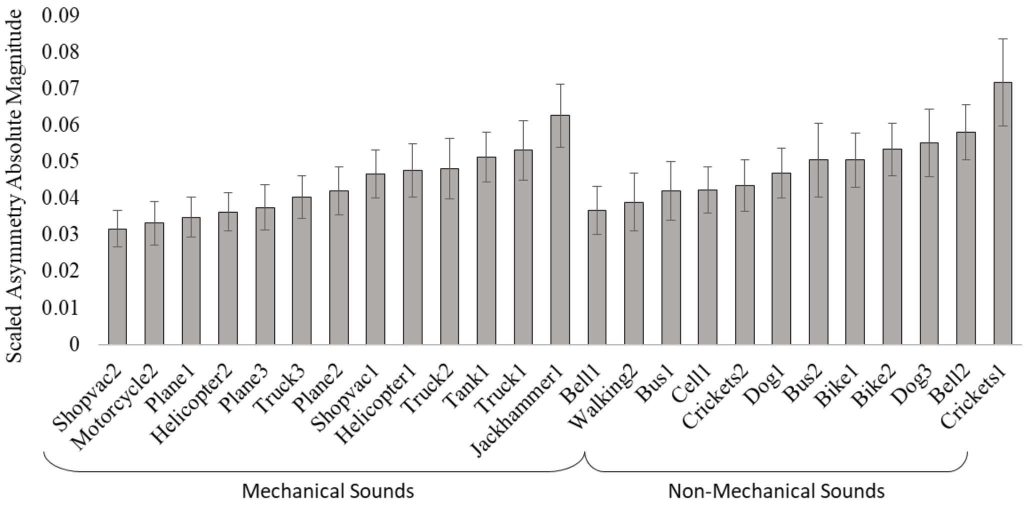


Fig. 4 Asymmetry absolute magnitude values for all stimuli in the set; error bars indicate standard error. For all stimuli, these values were significantly greater than zero.

5.2 Directional Effect Impact on MDS Solutions

In addition to determining the presence of asymmetry in the stimulus ratings, Experiment 1 also addresses the practical question regarding the extent to which asymmetries in pairwise similarity ratings will impact the resultant MDS solutions. This analysis was conducted to explore the potential impact on data quality of the time-saving practice of collecting only half of the pairwise comparisons (i.e., one triangle of the distance matrix containing only the *ij* or *ji* comparisons). Figure 5 shows the MDS solutions constructed using the *ij* and *ji* unidirectional ratings for the present stimulus set. The two MDS solutions are highly intercorrelated (x axis: r^2 = 0.957, y axis: r^2 = 0.783, and *m*: r^2 = 0.868), but there are some notable differences from the averaged MDS solution. Clustering over the averaged solution segregated the mechanical sounds from the more natural or tonic sounds, just as in the averaged solution. While the coarse topology remains largely similar, items are located in different locations in each of the solutions, and these differences may hold important implications for perceptual modeling.

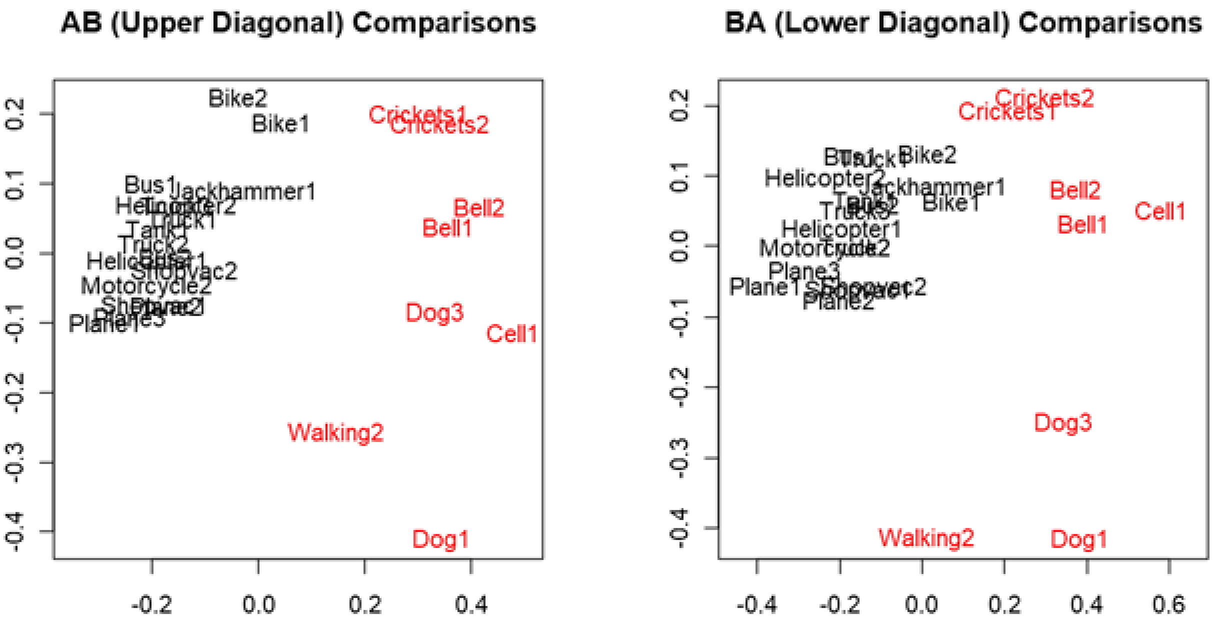


Fig. 5 MDS solutions generated using the *AB* and *BA* comparisons, colored according to cluster membership. While the two solutions cluster similarly, stimuli are placed differently within the solutions, especially for stimuli in the looser (red) cluster. This can have implications for modeling based upon the distances among stimuli in the space.

To summarize, the results of Experiment 1 demonstrate clear order effects in the rated similarity for environmental sounds, manifesting as an asymmetry effect in the similarity ratings for stimulus pairs. Further, this asymmetry has a clear impact on the MDS solution configurations depending on which comparisons (*AB* vs. *BA*) are used to generate the solution. This suggests that, as a practical matter, researchers should consider evaluating stimuli in both directions if pairwise similarity ratings are to be used for norming or informing interpretation of performance for environment sound items. What Experiment 1 does not reveal, however, is if the asymmetrical ratings follow any meaningful or logical pattern. The goal of Experiment 2 was to investigate these stimulus pairs for underlying structure that might explain the asymmetrical similarity ratings.

6. Experiment 2

The goal of Experiment 2 was an attempt to explain the observed asymmetries from Experiment 1 by characterizing the *AB–BA* stimulus order relationships, and to explore the potential impact of systematically ordered stimulus pairings on resulting MDS solution quality. The framework applied herein originates from seminal work in analogical transfer literature, where stimulus order relationships are characterized in terms of their systematicity, or the extent to which mutually constraining relations can be drawn between the base item (stimulus *A*) and the target item (stimulus *B*) (Gentner 1980, 1983). Put another way, a lifetime of experience with environmental sounds constrains the listeners' expectations about the “logical” order of items. For example, you expect to hear a buzzing insect followed by a flyswatter crack, but not necessarily the other way around. Systematicity has been operationalized in the literature differently according to stimulus modality and study domain, including stimulus saliency (Tversky 1977), prominence (Johannesson 1997), frequency (Nosofsky 1991), informativity (Bowdle and Gentner 1997), breadth of knowledge (Pothos and Busemeyer 2011), and the extent to which one stimulus provides a natural reference point for the other (Rosch 1975).

Here, we apply a dynamic semantic processing account of auditory processing in which semantic representations are constructed online in real time as information becomes available (van Petten et al. 1999). The retrieval of a semantic representation, based upon incoming sensory information, determines the extent to which that stimulus is identifiable to the participant. In turn, this allows the listener to compare the target and base items on the basis of their semantic similarity rather than merely perceptual similarity. In Experiment 2 we used listener estimates of stimulus identifiability as a declarative measure of semantic activation and operationalized systematicity as the relative identifiability of each stimulus in the pair. Accordingly, we expected that two stimuli would be rated as more similar (i.e., produce higher positive directional asymmetry magnitude values) when the more identifiable stimulus was presented in the *base* (*B*) position, because that order would permit more higher-order connections to be drawn between it and the *target* (*A*) item.

6.1 Method

6.1.1 Participants

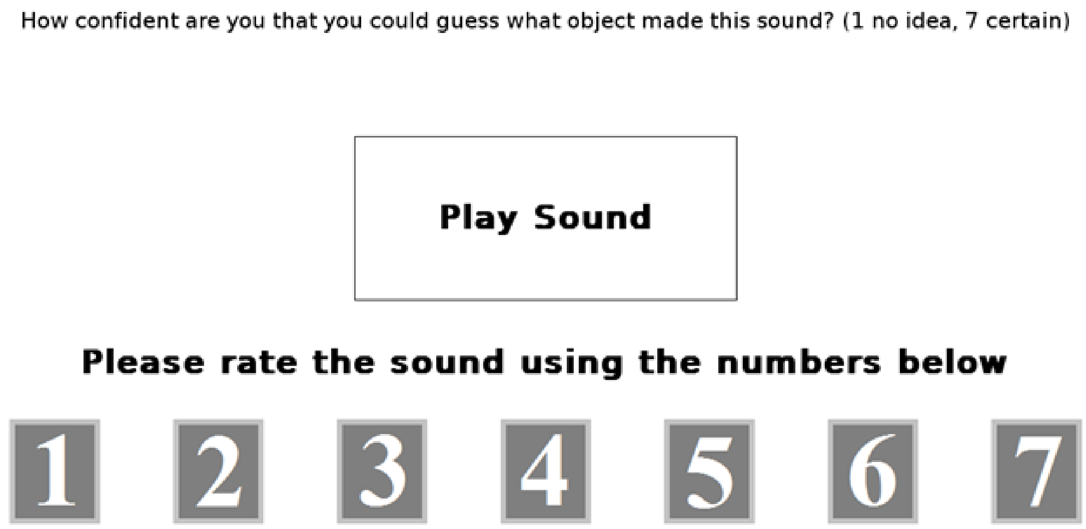
The same 26 Binghamton University undergraduate participants who completed Experiment 1 also completed Experiment 2.

6.1.2 Stimuli and Apparatus

Experiment 2 used the same 25 environmental sounds as Experiment 1, with the exception that stimuli were presented alone rather than in a pairwise fashion. An experiment program, coded in the Psychology Experiment Building Language (PEBL v. 0.14) (Mueller and Piper 2014), presented the participants with the stimuli and collected their responses using the same PC and headphones they used in Experiment 1.

6.2 Procedure

Each trial began with the presentation of a single environmental sound, presented from the full stimulus set in a random order. Participants could then repeat that sound as many times as desired. After listening to the environmental sound, participants were asked to rate the identifiability of that sound on a seven-point Likert scale (see Fig. 6). Participants provided a single rating for each sound over 25 total trials. Item-wise identifiability was calculated using aggregate statistics of these ratings.



How confident are you that you could guess what object made this sound? (1 no idea, 7 certain)

Please rate the sound using the numbers below

Fig. 6 Visual representation of the Likert scale used for rating the identifiability of each stimulus. Participants were instructed to use the entire range, rating each stimulus from 1 = no idea what object made the sound, to 7 = certain what object made the sound. Each trial began when participants clicked the Play Sound button, which triggered the appearance of the scale buttons. Participants could play each sound as many times as they liked before making the rating.

7. Results and Discussion

Participants generally rated the stimuli as identifiable (*M* = 5.23, *SD* = 0.86). Comparing stimulus identifiability by cluster (see Fig. 2), the mechanical sounds (black) were rated as slightly more identifiable (*M* = 5.34, *SD* = 0.86) than the melodic and natural sounds (red) (*M* = 4.99, *SD* = 0.84). The stimulus-wise identifiability ratings are shown in Fig. 7.

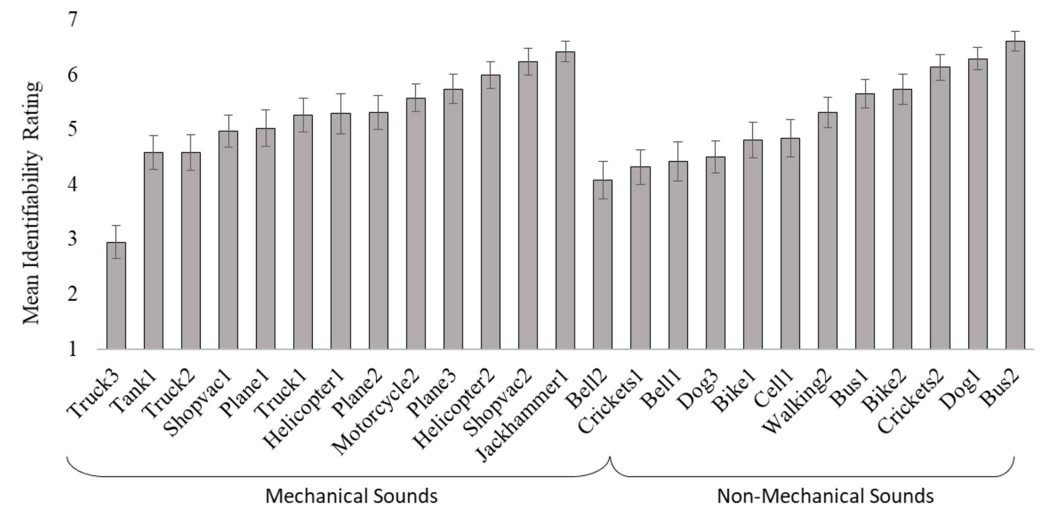


Fig. 7 Stimulus-wise identifiability for the environmental sound set. Note that in some cases sounds made by semantically similar objects differed radically in terms of identifiability (for example, the Truck, Cricket, and Dog).

To test for systematicity behind the asymmetries in the pairwise similarity ratings described in Experiment 1, the data structure associated with the identifiability data was constructed as follows. An identifiability relationship matrix was constructed with identical dimensions to the asymmetry matrix generated during Experiment 1. In this matrix, rows and columns indicate each stimulus in the set, and a particular cell corresponds to the relative identifiability of stimuli *i* and *j*. If the two were rated as equally identifiable, the cell received a zero. If stimulus *i* was rated as more identifiable than *j*, the cell received a 1, and if it was rated as less identifiable, it received a –1. The resulting matrix was a skew-symmetrical matrix (that is, if the value in cell *ij* was 1, the value in cell *ji* would be –1). This identifiability relationship matrix was used to bin the values in the asymmetry matrix, with one bin containing the pairings that place the more identifiable item in the *target* (*A*) position and the other placing the more identifiable item in the *base* (*B*) position. The resulting asymmetry distributions were compared using a t-test. Consistent with our hypothesis, placing the more identifiable item in the *base* (*B*) position produced higher similarity ratings, as indicated by the positive directional asymmetry magnitude values ($M = 0.019$, $SD = 0.599$), as opposed to placing it in the *target* (*A*) position [($M = –0.009$, $SD = 0.585$), $t(292.67) = 4.08$, $p < 0.001$, and *Cohen's D* = 0.047]. This result indicates an effect of identifiability, whereby the relative identifiability of the two stimuli produces systematic order effects in the rated similarity of those two stimuli.

We believe that this identifiability effect arises from differences in the extent to which higher order connections can be drawn from the *base* (*B*) item to the *target* (*A*) item. Using identifiability as a declarative indicator of semantic activation, the more identifiable item serves as a more natural reference point than the reverse. Environmental sounds are different from stimuli typically used in asymmetry studies because rather than explicitly indicating a noun or situation, they indicate physical interactions among objects in the world. Due to the broad stimulus pool used in this study, it is difficult to precisely quantify the similarity between the stimuli since they do not often share first order relationships. For example, by what criterion is it reasonable to gauge the similarity between a barking dog and a passing truck? The physical characteristics of each sound, and the sound sources themselves, are very different. However, higher-order causal relationships can influence a listener's perception of similarity: Perhaps the dog is barking as it chases the truck. As auditory perception research moves from the laboratory into the real world it is important to remember that the tools and methods employed for scientific discovery (e.g., pairwise similarity ratings) must be robust to the relationships among high-dimensional complex real-world stimuli.

Finally, a significant motivation for this study was to test whether order effects could impact experiments that use pairwise similarity ratings to create MDS solutions. We demonstrated with this stimulus set that an MDS solution constructed using randomly assigned *AB* comparisons was highly intercorrelated with one constructed using the *BA* comparisons. Having a measure of systematicity permitted us to test the result of introducing systematic order effects into the pairwise comparisons as well as the results of these effects on the resultant MDS solutions. Two new symmetrical distance matrices were created using the identifiability relationship matrix as a guide, one that placed the more identifiable item in the *target* (*A*) position and another which placed it in the *base* (*B*) position. These matrices represent worse-case scenarios in which experimenters introduced systematic relationships between the *A* and *B* variables. The resulting MDS solutions (Fig. 8) were less intercorrelated (x axis: $r^2 = 0.964$, y axis: $r^2 = 0.583$, and *m*: $r^2 = 0.773$) than those using the randomly assigned order relationships (*m*: $r^2 = 0.868$), indicating systematic assignment of stimuli to the *A* and *B* positions, whether intentional or unintentional, can impact the structure of the MDS solutions. Therefore, these results show that random assignment appears to be effective in reducing the impact of these asymmetries.

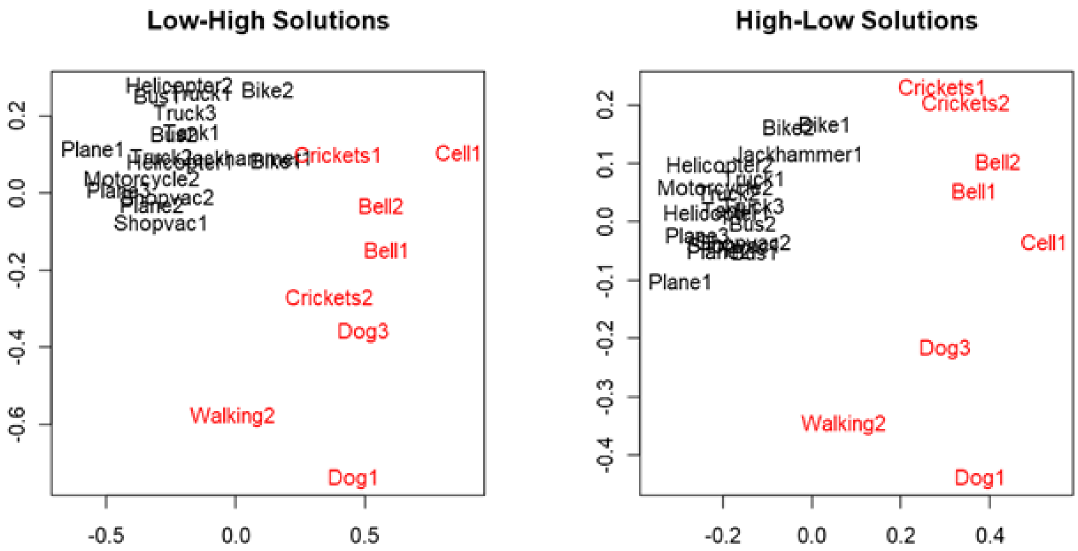


Fig. 8 MDS solutions generated using the Low-High and High-Low identifiability comparisons, colored according to cluster membership. These solutions are less intercorrelated than those produced using randomized *AB–BA* comparisons.

8. Discussion

This report presented the results of two experiments conducted to evaluate the presence and potential cause of asymmetries in pairwise similarity ratings of real-world environmental sounds. First we detected asymmetries in the rated similarity of environmental sounds and examined the effect of those asymmetries on experiment-relevant characteristics of the data (Experiment 1). Next we demonstrated the role that subjective assessments of identifiability played in producing those order effects and tested the ability of randomization to mitigate these effects (Experiment 2). The findings reported represent, to our knowledge, the first evidence of systematic order effects in the rated similarity of environmental sounds. The results of these two experiments suggest that presentation order can significantly impact participants' similarity ratings, and that one factor in producing this effect is the extent to which each of the items can be processed on the basis of their semantic characteristics: category membership and identifiability.

The practical research question driving this investigation concerned the impact that order effects, caused by directional semantic relationships between the stimuli, might have on MDS solution quality. These directional effects are a potential concern for stimulus selection when using environmental sounds because they could influence the results of perceptual studies; but also, efforts to mitigate these effects using MDS-based stimulus characterizations would be undermined by attempts to reduce the trial burden by running comparisons for items in only a single direction (*AB* vs. *BA*). In the stimulus set tested in the present study, the impact of directional asymmetries was not particularly large; however, it followed a logical, semantically interpretable pattern and thus illustrates how presentation order could influence the results of an MDS-based assessment of stimulus similarity. Experimenters using this method for stimulus norming, down selection, or other decisions related to stimulus selection should consider carefully the risks of running only unidirectional comparisons against the cost and effort of doubling the task trial load. This may be of particular concern for studies using a widely varied stimulus set. For this widely varied environmental sound pool, the MDS solutions produced from the *AB* and *BA* ordered stimuli were highly intercorrelated and topologically very similar. That said, stimuli were located differently between the MDS solutions, and this can have implications for stimulus pooling during research studies. Here, we discuss the circumstances under which we might expect to see asymmetrical ratings, and discuss potential methods for mitigating or eliciting these biases.

The present study used a widely varied pool of different environmental sounds. Hierarchical accounts of stimulus processing such as the one presented by Misdariis et al. (2010) suggest that auditory stimuli can be compared (and rated for similarity) at multiple levels. Higher levels consist of *categorical level* processing, which would include semantic properties of the stimulus representation elicited by the sound. Within categories, stimuli can be evaluated at the *continuous level*, the level at which stimuli in the group are semantically related, but rather differ on physical stimulus attributes. Comparing the similarity between a dog bark and a fire alarm, for example, is a judgment made at the categorical level. Conversely, rating the similarity of the deep bark of a large dog to the high-pitched bark of a small dog is a judgment that can be made on a small number of physical attributes of the sounds themselves (e.g., spectral characteristics). Studies employing a relatively constrained stimulus pool, such as underwater noises (e.g., Howard 1977) or impact sounds (Klatzky et al. 2000) may encourage participants to rate the similarity of stimuli on the basis of their physical attributes if they belong to the same semantic category or are unidentifiable. Conversely, broad sets of environmental sounds such as the type used by Gygi et al. (2007) and the present study permit a degree of semantic content to contribute to participants' similarity ratings. The result is the aforementioned skew in the ratings toward dissimilar, as well as a potential increase in the types of order effects observed here.

Recognizing a potential connection between stimulus breadth and asymmetrical similarity ratings, is it possible to mitigate these asymmetries when dealing with broad stimulus pools? In the present study, the *AB* and *BA* presentation orders were selected randomly rather than systematically, which corresponds to random assignment of the more identifiable stimuli to either the *A* or *B* positions, and thus random assignment of that stimulus to either the upper or lower portion of the similarity matrix. We also used the identifiability relationship matrix to create pairings in which the stimulus presentation order was assigned according to identifiability. MDS solutions generated from randomized comparisons shared 10% more variance than those generated with systematic assignment, indicating that randomization can provide an effective means of controlling for the influence of these order effects. If the goal of the research is simply to map the topology of the similarity space, then randomly assigning stimuli to *A* and *B* positions and collecting only half of the observations (i.e., *AB* comparisons only) may be sufficient. However, for more-sophisticated computational modeling or evaluating the psychophysical characteristics of the similarity spaces, collecting both *AB* and *BA* comparisons may be necessary because of the underlying asymmetries in any realistically complex environmental sound set.

9. Conclusion

In the research community there are generally two competing approaches to inferring invariant knowledge from experiments, one that seeks to control as many factors as possible in the hope of reducing the number of potential confounding variables, and another that aims to maximize the conditions under which the effect is observed (either through a high *n*, large number of trials, or randomization of experiment elements). As researchers increase complexity and realism in their experimental materials and tasks, it is untenable to control for all potentially confounding variables that could covary with the intended study manipulations. Approaches such as those outlined in this report, and the others in this series (Dickerson et al. 2015, 2016, 2018; McArdle et al. 2017), advocate the latter approach. We argue that the everyday environment is an exceptionally high-dimensional space with respect to the number of variables that could influence a given behavior. Even for a single environment it would be impossible to systematically control and manipulate all factors across all dimensions. Alternatively, we propose that testing experiment hypotheses pertaining to real-world behavior can be improved by observing the environment and documenting stimulus attributes that are stable and meaningful to the observer. That is, we can understand real-world behavior by characterizing aspects of the sensory environment. In this way, additional interpretive power is provided to assessments of behavior in dynamic, complex, and realistic environments. Further, studies such as those outlined in this report create rich datasets that begin the difficult work of providing quantifiable semantic features that could be used in future adaptive artificial intelligence development. Understanding high-level semantic relationships is critical for understanding real-world behavior in humans, and provides information regarding the types of representations that permit an intelligent agent to function in complex real-world environments.

10. References

Aldrich KM, Hellier EJ, Edworthy J. What determines auditory similarity? The effect of stimulus group and methodology. The Quarterly Journal of Experimental Psychology. 2009;62(1):63–83.

ANSI S3.6-2010. Specification for audiometers. Washington (DC): American National Standards Institute; 2010.

Bonebright TL. Perceptual structure of everyday sounds: a multidimensional scaling approach. Proceedings of the 2001 International Conference on Auditory Display; 2001. p. 73–78.

Bowdle BF, Gentner D. Informativity and asymmetry in comparisons. Cognitive Psychology. 1997;34(3):244–286.

Dickerson K, Gaston JR, McCarty-Gibson S. Parameterizing sound: design considerations for an environmental sound database. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2015 Apr. Report No.: ARL-TR-7198.

Dickerson K, Foots A, Moser A, Gaston J. Correlation between identification accuracy and response confidence for common environmental sounds. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2018 June. Report No.: ARL-TR-8383.

Dickerson K, Perelman BS, Sherry L, Gaston JR. Two classification methods for grouping common environmental sounds in terms of perceived pleasantness. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2017 Feb. Report No.: ARL-TR-7960.

Dickerson K, Gaston JR, Perelman BS, Mermagen T, Foots AN. Sound source similarity influences change perception in complex scenes. Proceedings of Meetings on Acoustics 169. 2015:23(1):050006.

Garner WR, Haun F. Letter identification as a function of type of perceptual limitation and type of attribute. Journal of Experimental Psychology: Human Perception and Performance. 1978;4(2):199.

Gaston JR, Letowski TR. Listener perception of single-shot small arms fire. Noise Control Engineering Journal. 2012;60(3):236–245.

Gentner D. The Structure of analogical models in science. Cambridge (MA): Bolt Beranek and Newman Inc.; 1980.

Gentner D. Structure-mapping: a theoretical framework for analogy. Cognitive Science. 1983;7(2):155–170.

Gentner D. The mechanisms of analogical learning. In: Vosniadou S, Ortony A, editors. Similarity and analogical reasoning. New York (NY): Cambridge University Press; 1989. p. 199–241.

Gregg MK, Samuel AG. The importance of semantics in auditory representations. Attention, Perception, & Psychophysics. 2009;71(3):607–619.

Gygi B, Kidd GR, Watson CS. Similarity and categorization of environmental sounds. Perception & Psychophysics. 2007;69(6):839–855.

Holman EW. Monotonic models for asymmetric proximities. Journal of Mathematical Psychology. 1979;20(1):1–15.

Howard Jr JH. Psychophysical structure of eight complex underwater sounds. The Journal of the Acoustical Society of America. 1977;62(1):149–156.

Johannesson M. Modelling asymmetric similarity with prominence. Lund University Cognitive Studies, LUCS 55. Lund (Sweden): Lund University; 1997.

Klatzky RL, Pai DK, Krotkov EP. Perception of material from contact sounds. Presence: Teleoperators & Virtual Environments. 2000;9(4):399–410.

Leisch F. FlexMix: A general framework for finite mixture models and latent glass regression in R. Journal of Statistical Software. 2004;11(8):1–18.

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

Lemaitre G, Susini P, Winsberg S, McAdams S, Letinturier B. The sound quality of car horns: a psychoacoustical study of timbre. Acta acustica. 2007;93(3):457–468.

McArdle J, Foots A, Stachowiak C, Dickerson K. Strategies for characterizing the sensory environment: objective and subjective evaluation methods using the VisiSonic Real Space 64/5 Audio-Visual Panoramic Camera. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2017 Nov. Report No.: ARL-TR-8205.

Misdariis N, Minard A, Susini P, Lemaitre G, McAdams S, Parizet E. Environmental sound perception: metadescription and modeling based on independent primary studies. EURASIP Journal on Audio, Speech, and Music Processing. 2010;(1):362013.

Mueller ST, Piper BJ. The psychology experiment building language (PEBL) and PEBL test battery. Journal of Neuroscience Methods. 2014;222:250–259.

Nosofsky RM. Stimulus bias, asymmetric similarity, and classification. Cognitive Psychology. 1991;23(1):94–140.

Parizet E, Koehl V. Application of free sorting tasks to sound quality experiments. Applied Acoustics. 2012;73(1):61–65.

Pothos E, Busemeyer J. A quantum probability explanation for violations of symmetry in similarity judgments. Proceedings of the Annual Meeting of the Cognitive Science Society; 2011. Red Hook (NY): Curran Associates, Inc.; c2011.

Rosch E. Cognitive representations of semantic categories. Journal of Experimental Psychology: General. 1975;104(3):192.

Siedenburg K, Jones-Mollerup K, McAdams S. Acoustic and categorical dissimilarity of musical timbre: evidence from asymmetries between acoustic and chimeric sounds. Frontiers in Psychology. 2016;6:1977.

Terasawa H, Slaney M, Berger G. Perceptual distance in timbre space. Proceedings of the 11th Meeting of the International Conference on Auditory Display; 2005. p. 61–68.

Truax B. Soundscale: acoustic communication and environmental sound composition. Contemporary Music Review. 1996;15(1):49–65.

Tversky A. Features of similarity. Psychological Review. 1977;84(4):327.

Van Petten C, Coulson S, Rubin S, Plante E, Parks M. Time course of word identification and semantic integration in spoken language. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1999;25(2):394.

Venables WN, Ripley BD. Random and mixed effects. In: Modern applied statistics with S. New York, (NY): Springer-Verlag; 2002. p. 271–300.

List of Symbols, Abbreviations, and Acronyms

2-D	two dimensional
BIC	Bayesian information criterion
FMM	flexible mixture modeling
MDS	multidimensional scaling
NMDS	nonmetric multidimensional
PC	personal computer
PEBL	Psychology Experiment Building Language

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

2 DIR CCDC ARL
(PDF) IMAL HRA
RECORDS MGMT
FCDD RLD CL
TECH LIB

1 GOVT PRINTG OFC
(PDF) A MALHOTRA

1 ARMY RSCH LAB – HRED
(PDF) FCDD RLH B
T DAVIS
BLDG 5400 RM C242
REDSTONE ARSENAL AL
35898-7290

1 USA ARMY G1
(PDF) DAPE HSI M SAMS
300 ARMY PENTAGON
RM 2C489
WASHINGTON DC 20310-0300

1 USAF 711 HPW
(PDF) 711 HPW/RH K GEISS
2698 G ST BLDG 190
WRIGHT PATTERSON AFB OH
45433-7604

1 USN ONR
(PDF) ONR CODE 341 J TANGNEY
875 N RANDOLPH STREET
BLDG 87
ARLINGTON VA 22203-1986

1 USA NSRDEC
(PDF) RDNS D D TAMILIO
10 GENERAL GREENE AVE
NATICK MA 01760-2642

1 OSD OUSD ATL
(PDF) HPT&B B PETRO
4800 MARK CENTER DRIVE
SUITE 17E08
ALEXANDRIA VA 22350

12 CCDC ARL
(PDF) FCDD RLH
J LANE
J CHEN
P FRANASZCZUK
K MCDOWELL
K OIE
FCDD RLH BD
D HEADLEY
FCDD RLH FA
A DECOSTANZA
FCDD RLH FB
A EVANS
FCDD RLH FC
J GASTON
FCDD RLH FD
K DICKERSON
B PERELMAN
A MARATHE