Perception & Psychophysics
2007, 69 (6), 839-855

Similarity and categorization of environmental sounds

Brian Gygi
East Bay Institute for Research and Education
and Veterans Affairs Northern California Health Care System, Martinez, California

and

Gary R. Kidd and Charles S. Watson
Indiana University, Bloomington, Indiana

Four experiments investigated the acoustical correlates of similarity and categorization judgments of environmental sounds. In Experiment 1, similarity ratings were obtained from pairwise comparisons of recordings of 50 environmental sounds. A three-dimensional multidimensional scaling (MDS) solution showed three distinct clusterings of the sounds, which included harmonic sounds, discrete impact sounds, and continuous sounds. Furthermore, sounds from similar sources tended to be in close proximity to each other in the MDS space. The orderings of the sounds on the individual dimensions of the solution were well predicted by linear combinations of acoustic variables, such as harmonicity, amount of silence, and modulation depth. The orderings of sounds also correlated significantly with MDS solutions for similarity ratings of imagined sounds and for imagined sources of sounds, obtained in Experiments 2 and 3—as was the case for free categorization of the 50 sounds (Experiment 4)—although the categorization data were less well predicted by acoustic features than were the similarity data.

The perception of environmental sounds has only recently begun to receive the level of attention that speech and music perception have enjoyed for many years. Given the prevalence of environmental sounds (defined here as all naturally occurring sounds other than speech and music) in everyday life and, importantly, throughout the evolution of the mammalian auditory system, this class of sounds certainly deserves more attention. Although there are many issues to be addressed as more mature theories of the perception of environmental sounds are developed, a fundamental goal is to identify the acoustic information on which judgments of source objects and events are based. The challenge is either to find the invariant acoustic information that specifies each object or event or to determine how objects and events are identified in the absence of acoustic specificity.

This report is the third in a series of studies that addresses this problem using different approaches. In the first (Kidd & Watson, 2003), 145 environmental sounds were rated on 20 semantic dimensions. The intercorrelations among the ratings suggested that almost 90% of the variance was associated with four factors, interpreted as harshness, size, complexity, and appeal. In the second investigation (Gygi, Kidd, & Watson, 2004) passband limiting and vocoder techniques were used to determine the spectral regions on which listeners based their identification of selections from a set of 70 environmental sounds. Spectral regions essential for recognition of this set of sounds were between 1200 and 2400 Hz, similar to those utilized in speech processing, with somewhat more reliance on higher frequencies. In the present investigation, similarity ratings of environmental sounds were obtained to determine the major perceptual dimensions of listeners' psychological space for environmental sounds. An examination of the acoustic properties and source properties that underlie those psychological dimensions was carried out to identify the perceptually relevant information on which the identification of environmental sounds is based.

The use of similarity ratings in the present context follows a strategy that has been used with naturally occurring visual stimuli. Findings from those similarity studies have contributed to the development of several sophisticated theories on the optical bases of visual similarity (Biederman, 1987; Heinemann & Chase, 1990; Marr & Vaina, 1982), which have been used extensively in the study of computer vision. The similarity of interrelated objects is often represented geometrically as a psychological space whose dimensions reflect properties of the stimulus such as, in the case of visual objects, size, shape, color, and rotation. Studies that have attempted to find the underlying dimensions of auditory similarity have most often examined similarity of voices or musical instruments (Cleary, Pisoni, & Kirk, 2005; Goldinger, 1996; Grey & Moorer, 1977; Halpern, Zatorre, Bouffard, & Johnson,

B. Gygi, bgygi@ebire.org

2004; Iverson & Krumhansl, 1993; LeCompte & Watkins, 1993). Few studies have attempted to determine acoustic bases for similarity of environmental sounds.[1]

One indirect method for determining the perceptual dimensions of naturally occurring sounds is to have subjects rate the similarity of large numbers of exemplars of such stimuli, to apply multidimensional scaling (MDS) methods to reduce the high-dimensional acoustic space to a few underlying perceptual dimensions, and then to find acoustic features that the dimensions could possibly represent, often by correlating the ordering of objects along each dimension with an acoustic feature. This approach has been used to determine the acoustic bases for musical timbre for both actual and synthesized musical instruments (Grey, 1977; Kendall & Carterette, 1991; Krumhansl, 1989; McAdams, Winsberg, Donnadieu, De Soete, & Krimphoff, 1995; Miller & Carterette, 1975; Plomp, 1970). The validity of these exploratory procedures was recently confirmed by Caclin, McAdams, Smith, and Winsberg (2005).

Although these studies have used different stimuli and methods, in general the MDS solutions have had one dimension that corresponded to the spectral centroid of the sounds and another that reflected temporal properties, such as attack time or amplitude envelope shape. In the studies in which three-dimensional solutions were used, the third dimension tended to have less consistent interpretations, reflecting properties such as temporal variations in the spectral envelope (Grey, 1977) or to the fine structure of the spectrum (Krumhansl, 1989).

MDS has also been used successfully to determine the acoustic correlates of environmental sounds, although many studies have used stimuli from a narrowly defined class of sounds. For example, Cermak and Cornillon (1976) derived a two-dimensional MDS solution from dissimilarity ratings of traffic sounds. One axis correlated nearly perfectly ($r = .98$) with the energy-equivalent sound level (interpreted as loudness), and the other axis had a moderately strong correlation with a measure of source information, namely the proportion of trucks and buses in the sound samples. Howard (1977) obtained similarity ratings of underwater sounds recorded from the hulls of Navy vessels, such as cavitation sounds, engine noises, and rain. The first dimension of the resulting two-dimensional solution correlated highly ($r = .91$) with the degree of bimodality in the spectra of the sounds. The sounds with prominent bimodalities were described by the listeners as "sounding like two sounds." The second dimension seemed to reflect the presence of low-frequency (<1 Hz) periodicities in the sounds.

The ranges of sounds used in the studies described above are extremely circumscribed in comparison with the number of sounds that humans encounter in everyday life. Among the few studies involving a wider variety of environmental sounds, Vanderveer (1980) and Bonebright (2001) both used fairly large and varied catalogs of sounds (30 and 74 sounds, respectively) representing several different types of sound-producing events. Rather than directly measure pairwise similarity, they both used open sorting tasks, in which subjects grouped sounds according to any criteria they chose. A similarity matrix was then generated in both cases, based on the sorting data, with the rationale that a sorting task takes much less time and generates a comparable similarity matrix to that obtained with direct similarity ratings. The validity of this method was demonstrated using both laboratory-generated waveforms and speech sounds in Bonebright (1996). Analyzing the similarity matrix qualitatively, Vanderveer concluded that the proximity of sounds in the matrix seemed to be based on similarity of temporal structure more than spectral structure. For instance, sounds that were rhythmic, repetitive, or percussive (knocking and hammering) tended to cluster together.

The similarity matrix created from the sorting data in Bonebright (2001) yielded a three-dimensional MDS solution accounting for 80% of the variance. Bonebright projected the acoustic variables onto the three-dimensional MDS space and found that the acoustic measures which best characterized the first dimension were amplitude ceiling (highest amplitude level in the sound), average intensity, and change in frequency; for the second dimension duration was the primary descriptor, whereas the third dimension was best described by amplitude ceiling and peak frequency. Few acoustic measures projected solely onto a single dimension.

Another study utilizing a sorting task with a wide range of environmental sounds was described by Marcell, Borella, Greene, Kerr, and Rogers (2000). In this study, the categories themselves were of interest, rather than the underlying similarity space. Listeners provided labels for 120 sounds using a free classification procedure (adapted from McAdams, 1993), and the resulting categories were grouped by two judges into 27 categories that were more general. These general categories were primarily based on the type of object (e.g., bird, musical instrument), with a few categories defined by event types (e.g., accident, signal) and the location or context in which the object or event is generally heard (e.g., kitchen, bathroom). Categories based on sound quality (e.g., high pitched) were not used often or consistently enough to be included in the 27 general categories.

The categories observed by Marcell et al. (2000) suggest that, when listening to environmental sounds, listeners attend to acoustic properties that provide information about source properties and are less concerned with sound quality. Gaver (1993) has suggested that this source-oriented listening is a mode of listening that is distinct from that used when we listen to music or when we are asked to judge the quality of a sound. These two modes of listening were termed *everyday listening* and *musical listening*, respectively. To clarify the nature of the everyday listening stimuli, Gaver developed a taxonomy of environmental sounds, based on the physics of sound-producing events and listeners' descriptions of sounds, that represents the different categories of events that might be identified by acoustic information. The taxonomy consists of a hierarchical description of basic "sonic events" (deformation, dripping, exploding, impact, pouring, rolling, scraping, splashing, and wind-related sounds) that form more complex events in combination at higher levels of the hierarchy.

Although Gaver's (1993) taxonomy does not include some important classes of sounds, such as vocalizations

and electronically synthesized sounds, it does represent a plausible and fairly comprehensive classification system consistent with the results of Marcell et al. (2000). To the extent that everyday listening is used when making similarity judgments in an experimental setting, we would expect to find that the perceived similarity of environmental sounds is strongly influenced by acoustic properties that are important for the identification of the different categories in the taxonomy.

The primary goals of the present studies are to determine the structure of the psychological space for environmental sounds and to identify the acoustic or physical properties that are associated with the underlying dimensions. By using different experimental techniques and a more varied collection of sounds, it is possible to test the generality of the earlier findings, to more systematically examine the perceptual similarity of a wide range of environmental sounds, and to identify the acoustic and nonacoustic information on which perceived similarity is based.

The first study reported here examined the use of pairwise similarity ratings, rather than sorting, to assess the perceived similarity of 50 familiar, naturally occurring environmental sounds, representing a broad range of sources. No labels were presented for the sounds, and listeners were asked to make judgments based solely upon the sounds presented. For the resulting similarity matrix, a three-dimensional MDS solution provided a parsimonious account of the data. Various acoustic measurements of the sounds were made, and several were found to be predictive of the ordering of the sounds along the dimensions of the MDS solution.

Despite the instructions given to the subjects, these acoustic similarity ratings may not be entirely based on the perceived similarity of the acoustic features of the sounds (what the listener actually heard). Judgments may also be influenced by a priori knowledge of the sounds and the events that produced those sounds. Although there were no labels presented along with the sounds, the vast majority of the sounds were readily identifiable (as shown in Gygi et al., 2004). Thus, knowledge of source object and events and of the range of sounds produced by a given source could have influenced listeners' similarity judgments. That is, even if the first token of a bird chirping sounded quite different from the second token, listeners might still have been inclined to rate them as highly similar, simply because they were both bird chirps.

To determine the possible influence of these nonacoustic factors, two additional similarity experiments that did not involve listening to any sounds were performed. One presented labels for the group of 50 sounds to listeners and asked them to rate the similarity of the *sounds* as they imagined them. The other study presented the same labels but asked listeners to judge the similarity of the *events* the labels represented. Although auditory memory and knowledge of source properties are potential influences in both of these experiments, any differences in the psychological spaces derived using these two instruction conditions would be informative. If it is possible to judge the similarity of source properties independently from judging the properties of the sounds, and if the two underlying psychological spaces are different, then the differences should be revealed by these experiments. If it is not possible to make these judgments independently, then there would be no basis for the existence of separate psychological spaces. Taken together, the three experiments illuminate the influence on similarity ratings of acoustic information, of the internal representation of sounds, and of knowledge of the events that create the sounds.

The fourth experiment described in this article examines free categorization of the same set of 50 sounds used in the first similarity study. These sounds were presented without labels to listeners who were asked to group them in whatever way made sense to them. The listeners were then asked to supply labels for their groupings. Similar to the method used by Marcell et al. (2000), the individual groupings were then consolidated by two judges into a smaller set of basic categories. Then, unlike Marcell et al., the relationships between the categories and the acoustic features of the sounds were explored both qualitatively and quantitatively. Finally, a similarity matrix was derived from the groupings, as in Bonebright (2001), and correlated with the matrix from the paired similarity findings from the first experiment, as well as with the acoustic features found to be predictive in that experiment. The combination of categorization data, similarity data from three experiments, and data from an extensive acoustic analysis provides new information about the listeners' auditory perceptual space and how it may be influenced by the acoustic information present in the environment.

The Catalog of Environmental Sounds

The 50 environmental sounds used in the experiments reported here are listed in Table 1. These sounds are a subset of the 70 sounds used in Gygi et al. (2004). In that study, they were filtered with varying bandwidths. In the present study, at the widest filter settings, all of the sounds used were nearly perfectly identifiable. An effort was made to select a representative sampling of the different types of meaningful sounds encountered in everyday listening, partially based on the classes described by Gaver, 1993: nonverbal human sounds, animal vocalizations, machine sounds, the sounds of various weather conditions, and sounds generated by human activities. Two tokens or examples of sounds from each source event were selected (e.g., two coughs). To reflect the range of sounds associated with a given source event or event class, the two tokens within each pair were chosen to be as different acoustically as possible, given the range of sounds available for each source event. The similarity within pairs varied considerably across source events (e.g., the clock-ticking sounds were more similar to each other than the bird sounds), but all token pairs were chosen to be easily distinguishable. One token of each sound was judged by the experimenter to be the primary token for that sound (largely on the basis of having been used in Gygi et al., 2004), and the other token was designated the secondary token. The tokens were obtained from high-quality commercial sound effects recordings (Hollywood Edge and Sound FX The General) sampled at 44.1 kHz. The sounds were equated according to the root mean square (RMS) in a 100-msec window

Table 1
List of Sounds Used in the Acoustic Similarity and Categorization Experiments

Label	Short Name	Label	Short Name
Airplane flying	Airplane	Horse neighing	Neigh
Chopping wood	Axe	Ice dropping into glass	Ice drop
Baby crying	Baby	Typing on keyboard	Keyboard
Basketball bouncing	B-ball	Person laughing	Laugh
Bells chiming	Bells	Lighting a match	Match
Bird calling	Bird	Car accelerating	Car accel.
Bowling	Bowling	Phone ringing	Phone
Bubbling	Bubbling	Ping-pong ball bouncing	Ping pong
Car starting	Car start	Water pouring	Water pour
Cat meowing	Cat	Rain	Rain
Hands clapping	Claps	Rooster crowing	Rooster
Clock ticking	Clock	Scissors cutting paper	Scissors
Helicopter flying	Copter	Sheep baaing	Sheep
Person coughing	Cough	Siren blaring	Siren
Cow mooing	Cow	Person sneezing	Sneeze
Hitting cymbals	Cymbals	Splash	Splash
Dog barking	Dog	Thunder rolling	Thunder
Door opening & closing	Door	Toilet flushing	Toilet
Drumming	Drums	Cars honking	Honking
Electric saw cutting	Elec. saw	Train moving	Train
Footsteps	Footsteps	Typing on typewriter	Typewriter
Glass breaking	Glass break	Waves crashing	Wave
Gun shot	Gun	Whistle blowing	Whistle
Strumming harp	Harp	Windshield wipers	Wipers
Horse running	Gallop	Zipper	Zipper

Note—The short name is used to refer to the sounds as they appear in Figures 1–4.

around the peak amplitude in each sound and stored as binary files. The mean duration of the sounds was 2.3 sec, with the shortest sound being a cat meowing (579 msec) and the longest, a ping-pong ball bouncing (3,945 msec).

EXPERIMENT 1
Acoustic Similarity

Method

Subjects. Three of the subjects were male Indiana University students—two undergraduates aged 21 and a graduate student aged 30. A 4th subject was a male architect aged 24. All 4 subjects had normal hearing as measured by pure tone audiograms (thresholds $<$15 dB HL from 250–8000 Hz). The first 3 subjects were run in July, 1999 at Indiana University, and the fourth at the Veterans Affairs Medical Center in Martinez, CA, in August 2003. The subjects were paid for their participation. At both Indiana University and the VA Medical Center, 2 other listeners began the study; but all 4 dropped out before completing the experiment, and their data were not used.[2] The similarity of the data from all the subjects, as described in the Results, indicates that the attempt to replicate the conditions properly was successful.

Apparatus. At Indiana University, the experiment was conducted in the Group Experimentation Laboratory, a large, sound-treated chamber in which the subjects were tested simultaneously in individual stalls. The stimuli were generated using TDT 16-bit digital-to-analog converters (Tucker-Davis Technologies), amplified by a Macintosh 24 amplifier and delivered diotically to Etymotic ER-3A insert earphones. At the VA Medical Center, the subject was seated in a soundproof booth. The stimuli were generated from digital files by Echo Gina 24 sound cards, amplified by the TDT System 2 headphone buffer and presented through Sennheiser 250 II headphones. In both cases a 1-kHz calibration tone of the same RMS as the equated peaks of the sounds was set to 75 dB SPL at the headphones, and responses were recorded on individual PCs.

Procedure. The listeners were instructed to listen to each pair of sounds and to rate their similarity on a scale of 1 (*not similar at all*) to 7 (*as similar as they can possibly be*). They were told that, although many sounds might not seem similar at all, there was a gradation of similarity even among very dissimilar sounds. Examples were given using verbally presented items that were unlikely to evoke specific sounds (e.g., "How similar is the United States to Mexico?"). They were told to use the full scale when making their judgments and to try to make the average response about 4. The instructions did not include any mention of a possible distinction between event similarity and acoustic similarity, and there was no further guidance about the information on which the similarity judgments should be based.

A tone pulse preceded every trial. After every five trials the presentation software would produce two tone pulses to help listeners be certain they were on the correct trial.

The full matrix of 10,000 sound pairs, including each sound paired with itself and both orders of every pair, was presented randomly in blocks of 50 trials. Testing took place in 2-h sessions, with 2-min breaks after every block and a 5-min break after every fifth block. An additional block of trials was run at the beginning of the first day as practice. On average, listeners completed 10 blocks per day; testing required approximately three weeks to complete. Despite the inclusion of some makeup sessions, 245 similarity ratings are missing out of 40,000 (4 subjects $\times$ 10,000 ratings).

Results and Discussion

Overall, the mean similarity was 2.98, with an *SD* of 1.05. The distribution was close to normal, although slightly positively skewed because of the high similarity of sounds to themselves. Two of the listeners graded their responses fairly strongly to the low end of the 7-point scale, with means of 2.70 and 2.47, respectively. The means of the third and fourth listeners were more toward the middle of the scale at 3.46 and 3.26, respectively. The *SD*s of

all four were comparable, however, ranging from 1.20 to 1.45. For all listeners the similarity of a sound to itself was consistently judged to be 7, except for five instances: 4 from Subject 2 and 1 from Subject 4. In the MDS analysis to follow, those values were all replaced with 7.

The intersubject correlations were modest, as shown by the first value in each cell in Table 2. To account for some of the noise from measurement error, it is possible to combine ratings for each stimulus pair, ignoring the order of presentation (e.g., baby→horse and horse→baby). When the two ratings for each subject are averaged, the correlations become much stronger, as shown in the second value in each cell in Table 2. A common feature across all 4 listeners was that, in general, the two different tokens of each sound were judged to be extremely similar, with mean subject ratings of 6.07, 5.87, 6.01, and 6.57, respectively. The token pairs that were judged to be least similar were those for *airplane* (a jet and a prop plane) and *bird* (a robin and a loon), both of which had mean similarities of 4.88 across the 4 subjects.

Multidimensional Scaling Solutions

The similarity ratings for the 4 listeners were normalized to account for the intersubject differences in mean ratings. (As noted above, prior to normalization all the same-sound similarity ratings were changed to 7.) The four ratings were averaged to form a full similarity matrix. This matrix was then submitted to MDS analysis using an alternating least squares scaling method (ALSCAL) and a standard Euclidean model. Examination of the scree plot (a plot of stress as a function of dimensionality) revealed a slight elbow at two dimensions, but the stress was relatively high (.32), and the proportion of variance accounted for (RSQ) fairly low (.50). For three dimensions, both the stress (.24) and RSQ (.59) were better, and comparable to that found in other published studies using complex sounds (Allen & Scollie, 2002; Grey, 1977; Howard & Ballas, 1983; Howard & Silverman, 1976), so a three-dimensional solution was used. Two-dimensional plots of Dimension 1 versus Dimension 2 and of Dimension 1 versus Dimension 3 are shown in Figure 1.

A salient feature of the plot of Dimension 1 versus Dimension 2 is the clustering of sounds in distinct areas. The three clusters seem to correspond strongly to physical attribute of the sounds. Harmonic sounds are grouped along the upper half of Dimension 1, discrete impact sounds are on the lower half of both Dimensions 1 and 2, and continuous sounds are along the upper half of Dimension 2. The plot of Dimension 1 versus Dimension 3 does not exhibit such distinct clusters. There does seem to be a demarcation between vocalizations and nonvocalizations that cuts across the *x*- and *y*-axes. A feature of both plots is the proximity of similar sounds. Vocalizations and signaling sounds (baby crying, whistle, sheep), water-based sounds (rain, splash), rhythmic impacts (footsteps, typewriter, drums), and mechanical sounds (airplane, car accelerating, electric saw) each tend to group together.

One way to make explicit the clusters in this group of data is through hierarchical clustering analysis on the ordering of the sounds in the three-dimensional space. The clustering method used was average between-groups linkage, with a squared Euclidean distance metric (Loh & Shih, 1997). The resulting hierarchical cluster analysis diagram is shown in Figure 2, along with a tentative labeling scheme. One obvious difference among the clusters is the division between harmonic and nonharmonic sounds. The harmonic group contains vocalizations as well as musical instruments and signaling sounds. Within the nonharmonic sounds are groupings of impulsive sounds, continuous sounds, and a distinct subgroup of mechanical sounds. The qualitative difference between the groups was confirmed by a *t* test on pitch salience, a measure of harmonicity discussed below [$t(1,98) = -11.98, p < .01$].

Table 2
Intersubject Correlations of Similarity Ratings in the Acoustic Similarity Study

	All 10,000 Ratings/ *Mean Ratings for Each Pair* (*Both Orders*)		
	Subject 2	Subject 3	Subject 4
Subject 1	.43/*.61*	.63/*.75*	.46/*.63*
Subject 2		.46/*.64*	.41/*.57*
Subject 3			.51/*.66*

Acoustic Factors in Similarity

The data from Experiment 1 suggest an orderly relation between the similarity ratings of sounds and of their acoustical properties. An examination of the sounds with the greatest and least values on Dimension 1 reveals that the harmonic sounds tend to be at the upper end of the scale, whereas the inharmonic sounds are at the lower end. Similarly, as noted for Dimension 2, the percussive sounds are at the upper end of the range, whereas the continuous sounds are at the lower end, so a measure of periodicity or continuity might account for the ordering. Dimension 3 does not offer such a ready interpretation: Brief, spectrally static sounds tend to cluster at the upper end, whereas the lower end of the scale contains longer sounds with more dynamic spectra. This dimension may represent a measure of spectral–temporal complexity. To quantitatively evaluate these interpretations, measurements of various acoustic features of the original waveforms were made, and each of these measures was correlated with the MDS results along each dimension.

The variables measured reflected different spectral–temporal aspects of the sounds, including statistics of the envelope, autocorrelation statistics, and moments of the long-term spectrum. Most of the variables were used in Gygi et al. (2004) and Shafiro (2004), and several were found to be predictive of the identification of environmental sounds under conditions of a varying number of spectral channels. The measures and a brief description of each are listed below.

Envelope measures. (1) Long-term RMS/pause-corrected RMS (an index of the amount of silence); (2) number of peaks (transients, defined as a point in a vector that is greater in amplitude than the preceding point by at least 80% of the range of amplitudes in the vector); (3) number of bursts (amplitude increases of at least 4 dB

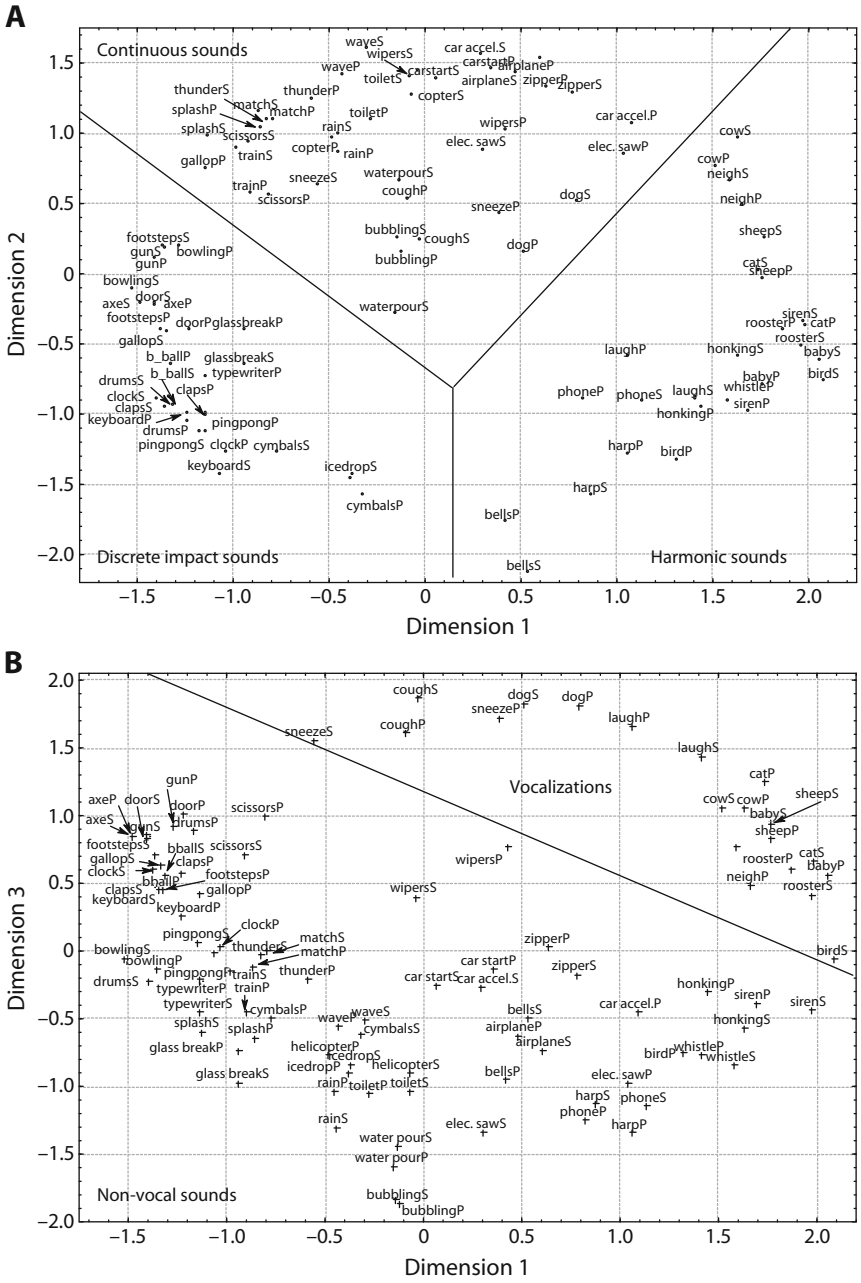

Figure 1. (A) Dimension 1 versus Dimension 2, and (B) Dimension 1 versus Dimension 3 from the three-dimensional multidimensional scaling solution for the acoustic similarity data. Tokens with a "P" after the sound type label are the primary tokens for that sound type, and those with an "S" are the secondary tokens. The major clusters for each chart are marked as described in the text.

sustained for at least 20 msec, based on an algorithm developed by Ballas, 1993); (4) total duration; and (5) burst duration/total duration (a measure of the "roughness" of the envelope).

Autocorrelation statistics. Number of peaks, maximum peak, mean peak, and *SD* of the peaks. Peaks (as defined above) in the autocorrelation function reveal periodicities in the waveform. The statistics of the peaks measure different features of these periodicities, such as the strength of a periodicity and the distribution of periodicities across different frequencies.

Correlogram-based pitch measures (from Slaney, 1995). Mean pitch, median pitch, *SD* pitch, maximum pitch, mean pitch salience, and maximum pitch salience. The correlogram measures the pitch and pitch salience by autocorrelating in sliding, 16-msec time windows. This captures spectral information and provides measures of the distribution of that information over time.

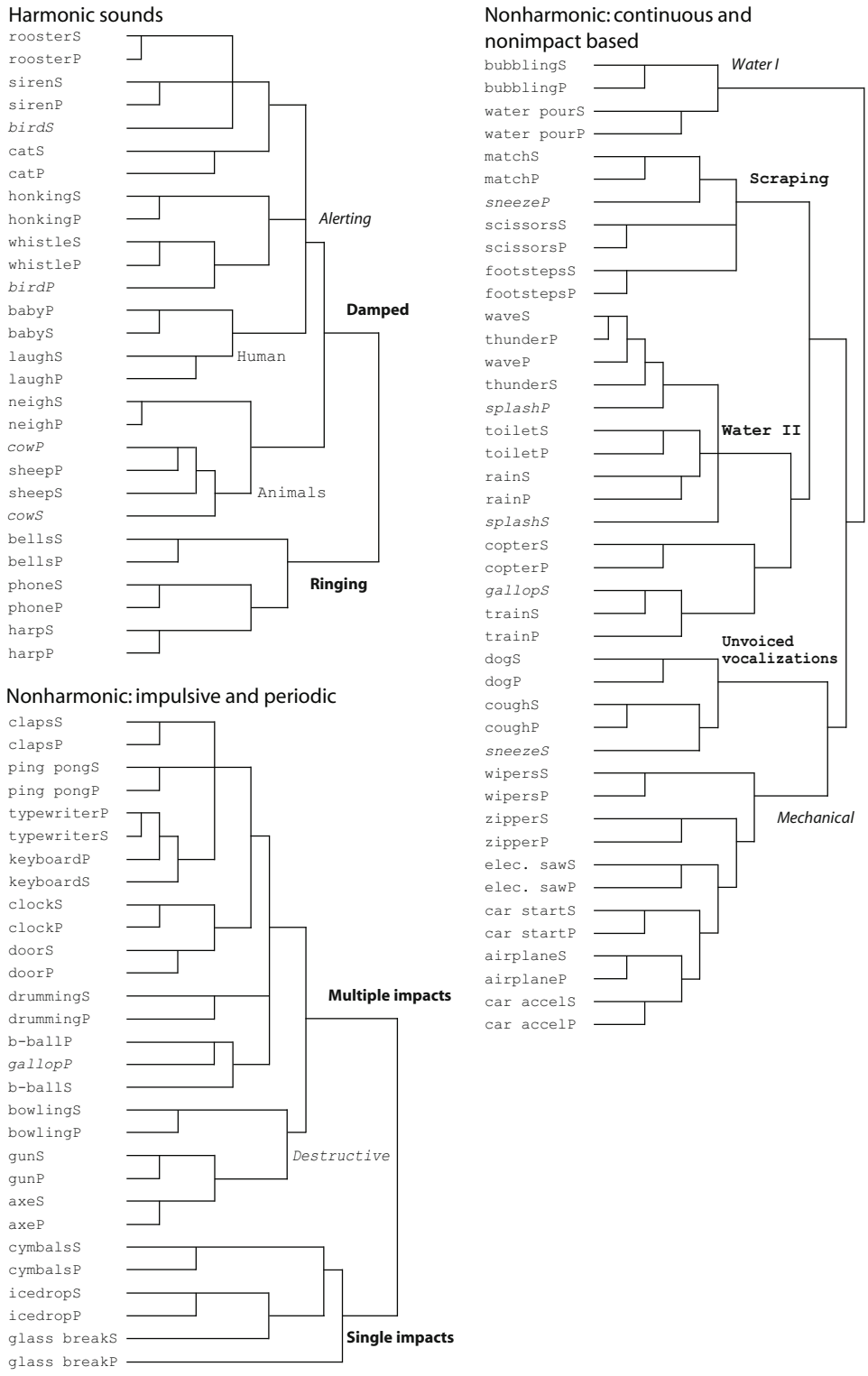

Figure 2. Hierarchical cluster analysis of the three-dimensional multidimensional scaling solution for the acoustic similarity data. Same-sound tokens that are not contiguous in the clustering are italicized.

Moments of the spectrum. Mean (centroid), *SD*, skew, and kurtosis.

RMS energy in octave-wide frequency bands from 63 to 16000 Hz.

Spectral shift in time measures. Centroid mean, centroid *SD*, mean centroid velocity, *SD* centroid velocity, and maximum centroid velocity. The centroid mean and *SD* are based on consecutive 50-msec time windows

throughout the waveform. The spectral centroid velocity was calculated by measuring the change in spectral centroid across sliding, 50-msec, rectangular time windows.

Cross-channel correlation. This is calculated by correlating the envelopes in octave-wide frequency bands (or channels) ranging from 150 to 9600 Hz. It measures the consistency of the envelope across channels.

Modulation spectrum statistics. The modulation spectrum, first suggested by Houtgast and Steeneken (1985), reveals periodic temporal fluctuations in the envelope of a sound. The algorithm used here, divides the signal into frequency bands approximately one critical band wide, extracts the envelope in each band, filters the envelope with low-frequency bandpass filters (upper f_c ranging from 1–32 Hz), and determines the power at that frequency. The result is a plot of the depth of modulation by modulation frequency. The statistics measured were the height and frequency of the maximum point in the modulation spectrum, as well as the number, mean, and variance of bursts in the modulation spectrum (using the burst algorithm described above).

Spectral flux statistics. Spectral flux is another measure of the change in the spectrum over time. As described by Lakatos (2000), it is the running correlation of spectra in short (50-msec) time windows. The mean, *SD*, and maximum value of the spectral flux were used in this analysis.

Acoustic Analysis Results and Discussion

Several acoustic variables were significantly correlated with the ordering on Dimension 1, as shown in Table 3. The strongest correlation was with the *mean pitch salience* (a measure of pitch strength or harmonicity), at $r = .75$. This confirms that Dimension 1 in large part reflects the pitch strength of a sound. Dimensions 2 and 3 were not nearly as strongly predicted by individual acoustic variables, and the common features are harder to discern.

Multiple regressions on the orderings on each dimension were conducted to determine whether combinations of acoustic variables better predicted the MDS findings, as summarized in Table 4. By and large the correlations between the predictors were low: Of the 903 correlations in the half matrix, only 99 had an absolute value greater than .4. There were some quite high correlations between similar types of predictors. The moments of the spectrum had high intercorrelations, as did the statistics of the autocorrelation matrix, the measures of harmonicity, the variables related to spectral velocity, and the modulation spectrum statistics. For Dimension 1, 79% of the variance could be accounted for by three variables: the mean pitch salience, the modulation spectrum maximum, and the spectral skew. These variables reflect distinct acoustic features: the concentration of spectral energy, the degree of harmonicity, and the depth of envelope modulation. Although pitch salience is by far the strongest individual predictor, listeners seem to use all three in making similarity judgments. In general, because of the nature of the source, vocalizations tend to (1) be harmonic in structure; (2) have more energy in the lower frequencies, and (3) tend to be continuous (i.e., lacking the large amplitude modulations of impulsive or intermittent sounds).

Multiple regression solutions on Dimensions 2 and 3 are less predictive than those for Dimension 1, and for neither dimension did any single acoustic variable correlate as strongly as with Dimension 1. Moreover, because of the high redundancy among the individual predictors, the best multiple regression solutions often did not include the best individual predictors. The strongest single predictor for Dimension 2 ($r = -.45$) was the RMS energy in a one-third-octave band centered at 250 Hz. It, however, was not included in the best multiple regression solution on Dimension 2, which retained six variables and accounted for 60% of the variance in the ordering on Dimension 2 (also included in Table 4). The acoustic features that are predictive in this case reflect two aspects of the waveforms: the long-term spectral composition (shown by the inclusion of *spectral centroid* and *spectral spread*) and the envelope structure (evidenced by the presence of *number of peaks*, *RMS/total energy*, and *burst duration/total duration*). Sounds in the lower portion of Dimension 2 tend to have greater high-frequency content and to be more periodic and impulsive, as shown in Figure 2. Impacts and repeated percussive sounds occur almost exclusively in the lower half of Dimension 2, whereas water-based sounds (dripping, pouring, and splashing) are in the upper half. Impacts will tend to have more bursts, more silence, and a greater proportion of high transients than nonimpact sounds. Water sounds are more continuous, with more low-frequency energy.

The tendency for Dimension 2 to be associated with envelope structure, whereas Dimension 1 is predominately associated with frequency information, is consistent with the findings from timbre studies mentioned previously (Grey, 1977; Kendall & Carterette, 1991; Krumhansl, 1989; McAdams et al., 1995; Miller & Carterette, 1975; Plomp, 1970). One factor that makes comparisons of these findings with the MDS studies on musical sounds difficult is that, although musical instruments present a broad range

Table 3
Strongest Correlations of Acoustic Variables With the Ordering on Each Dimension in the Acoustic Similarity Multidimensional Scaling Solution

Dimension 1		Dimension 2		Dimension 3	
Variable	*r*	Variable	*r*	Variable	*r*
Mean pitch salience	.75	RMS in band f_c = 250 Hz	.45	Duration	−.37
Spectrum *SD*	−.61	Autocorrelation function maximum	−.44	Spectral centroid	−.28
Maximum pitch salience	.58	Mean pitch salience	−.37	RMS in band f_c = 8 kHz	−.27
Modulation spectrum maximum	−.57	Burst duration/total duration	−.34	Adjusted RMS/total RMS	−.27
Spectral skew	.52	Median pitch	−.33	RMS in band f_c = 4 kHz	−.26
Mean autocorrelation function peak	−.49	RMS in band f_c = 500 Hz	.32	RMS in band f_c = 500 Hz	.26

Table 4
Multiple Regression Solutions for the Ordering of Sounds on the Three Dimensions of the Multidimensional Scaling Solution by Acoustic Variables

Variable	β	Description of Relation re: Maximal Values on Dimension
Dimension 1		Regression summary: $R = .890$, $R^2 = .792$
Mean pitch salience	.54	Greater pitch salience; greater harmonicity, "pitchier"
Modulation spectrum	−.45	Lesser depth of modulation
Maximum spectral skew	.25	More positive skew (greater proportion of energy @ lower frequencies)
Dimension 2		Regression summary: $R = .772$, $R^2 = .596$
Spectral centroid	−.67	Lower spectral centroids
Spectrum *SD*	.62	Broader spectra
Autocorrelation peaks *SD*	−.48	Narrower distribution of autocorrelation peaks
Pause-corrected RMS/overall RMS	.38	Less silence
Number of envelope peaks	−.30	Fewer peaks in the envelope
Burst duration/total duration	−.23	Bursts comprise a lesser proportion of the envelope
Dimension 3		Regression summary: $R = .750$, $R^2 = .562$
Duration	−.59	Briefer sounds
Number of bursts	.47	More bursts in the envelope
Mean spectral flux	.38	Greater spectral variation over time
Autocorrelation peaks *SD*	−.38	Narrower distribution of autocorrelation peaks
Median pitch	−.27	Lower pitched sounds
Spectral centroid	−.26	Lower spectral centroids

of sounds, most MDS studies with musical sounds have involved sustained, steady-state timbres. Very few have used brief percussive sounds with quick attacks and transients, so the findings tend not to refer to those types of sounds.

The emphasis on spectral features found here does contrast somewhat with Vanderveer's (1980) conclusion that temporal patterning is the dominant feature in the similarity of sounds, although she based her results on grouping data rather than on pairwise similarity ratings and did not perform any quantitative analyses. However, these findings are consistent with Vanderveer's, in that Dimension 2 seems to at least partly reflect a sensitivity to temporal patterns. Vanderveer's suggestion of a primary role of temporal structure—as opposed to spectral—is difficult to evaluate for several reasons, including the available measures of temporal structure and differences in the makeup of the sound catalogs. A stronger test of the role of temporal structure may require the development of new measures that can capture more of the perceptually salient differences in temporal patterning.

The correlations of single acoustic variables with the ordering on Dimension 3 are weaker overall than for those on Dimension 2. Even the strongest association—that with total duration—was rather weak ($r = -.37$). However, the most successful multiple regression solution for Dimension 3 had fairly good predictive power, accounting for 56% of the variance in the ordering. That solution retained six variables, shown in Table 5, which seem to indicate, as proposed above, that Dimension 3 reflects a measure of spectral–temporal complexity, including variables that measure envelope shape (bursts, autocorrelation peaks *SD*) and spectral change in time (mean flux, median pitch), in addition to overall duration.

Comparing the acoustic analyses done here with those of Bonebright (2001) is inexact because she did not correlate the features with the ordering on individual dimensions. Instead, as mentioned above, she projected each acoustic feature vector onto the three-dimensional MDS solution, with the result that the vectors cut across the dimensions and did not yield easily interpretable results for a single dimension. In addition, the strength of the correlations for each feature on the individual dimensions is not known. However, as mentioned above, she did find that (1) Dimension 3 seemed to reflect a spectral feature—peak frequency; (2) Dimension 2 had a relation with the duration of the sounds; and (3) instantaneous amplitude, average intensity, and frequency change were related to both Dimensions 1 and 3. Given the differences in the stimuli and the analysis methods, the general similarity to the present findings is perhaps more noteworthy than the differences are. More meaningful comparisons across studies may be possible once metrics that more directly capture perceptually relevant acoustic properties have been identified. Such properties may include higher order spectral, temporal, and spectral–temporal patterns that are only indirectly reflected in the current metrics.

EXPERIMENTS 2 AND 3
Similarity of Imagined Sounds and Imagined Events

The analyses indicate that the acoustic similarity ratings obtained in Experiment 1 were strongly influenced by the perceived similarity of the acoustic features of the sounds (i.e., what the listener actually heard). However, it is likely there were other factors involved, such as a priori knowledge of the sounds (long-term auditory memory), and a priori knowledge of the events that produced those sounds. To assess the potential influence of this a priori knowledge, two similarity experiments were performed that did not involve listening. The goal of these experiments was to provide comparisons of psychological spaces across all three experiments when listeners were asked to judge the similarity of presented sounds, imagined sounds, or the sources of sounds. Of course, if subjects are unable to attend to a single source of information (i.e., if auditory memory and source knowledge strongly influence similarity judgments in all conditions), differences among the

Table 5
Pairwise Correlations of the Orderings on the Dimensions Across the Different Similarity Studies

	Acoustic Similarity: Secondary Tokens	Sound Image Similarity	Source Image Similarity
Dimension 1			
Acoustic similarity: primary tokens	.97	−.81	.88
Acoustic similarity: secondary tokens	–	−.83	.90
Sound image similarity	–	–	−.93
Dimension 2			
Acoustic similarity: primary tokens	.94	.74	−.59
Acoustic similarity: secondary tokens	–	.74	−.57
Sound image similarity	–	–	−.47
Dimension 3			
Acoustic similarity: primary tokens	.95	.59	.03
Acoustic similarity: secondary tokens	–	.57	.03
Sound image similarity	–	–	.52

Note—The primary and secondary tokens for the acoustic similarity study are correlated separately with each of the image similarity studies.

spaces will be minimal. However, any differences among the MDS solutions will be informative.

The first of these studies requiring memory-based comparisons (sound image similarity) was designed to determine what listeners retain in memory of the sounds *as sounds*. The research on auditory imagery is not extensive. A fair amount of work has focused on the neurological correlates of imagined musical sounds (see, e.g., Halpern et al., 2004; Zatorre & Halpern, 2005). Other work has addressed psychoacoustic features (loudness and pitch) of imagined sounds (Intons-Peterson, 1980; Intons-Peterson, Russell, & Dressel, 1992) and the role of auditory imagery in remembering sounds actually heard (Sharps & Pollitt, 1998; Sharps & Price, 1992). This work indicates that, for most listeners, auditory imagery could provide a basis for reliable similarity judgments. The second study (source image similarity) was an examination of listeners' perceptual space for the events that produced the sounds.

In both experiments, stimuli were the text labels for the 50 sounds used in the acoustic similarity experiment, which are listed in Table 1. The difference between the two studies is that in the first, subjects were instructed to rate the similarity of imagined sounds (e.g., "How similar is the sound of a *baby crying* to the sound of a *car crash*?"); and in the second, sounds were not mentioned, and subjects were instructed to rate the similarity of their *image* of the events ("How similar is *a baby crying* to *a car crash*?"). The results of this pair of experiments will be considered together.

Experiment 2
Sound Image Similarity (Imagined Sound)

Method

Subjects. Five female college students took part in this study: three undergraduates and two graduate students, all between the ages of 21 and 24. All had normal hearing as measured by puretone audiograms (thresholds <15 dB HL from 250–8000 Hz), to rule out differences in imagined or remembered similarity due to hearing impairment.

Procedure. These two studies were conducted in the Group Experimentation Laboratory at Indiana University, where the first three subjects in the first study were tested. The instructions to the subjects were similar to those in the acoustic similarity study, except listeners were told to " . . . make similarity judgments based on your memory or knowledge of these sounds," and to use the similarity of a sound to itself as an anchor for the upper end of the scale. Subjects were provided the following example: Imagine the similarity of a dog barking and cow mooing, as opposed to that of a dog barking and a gunshot.

A pair of labels from the list in Table 1 was presented on each trial, and listeners were asked to rate the similarity of the sounds on a scale from 1 to 7 by typing in a response, as in the previous experiment. The experiment was self-paced, and listeners were allowed as much time as they needed to respond.

The 1,225 pairings from the half-matrix of 50 labels were presented in random order. A training block of 125 trials containing all the labels used as stimuli was presented at the beginning. Data from this training trial block were not included in the final analysis. A 2-min break was given after every 8 min, and a 5-min break after every 40 min. The study was completed within one 2-h session.

Experiment 3
Source Image Similarity (Imagined Event)

Method

Subjects. Four females and two males participated in this study. The females and one male were Indiana University students between the ages of 21 and 23, and one male aged 29 was not a student. As in the sound image similarity study, all had normal hearing as measured by puretone audiograms (thresholds <15 dB HL from 250 to 8000 Hz).

Procedure. As noted above, the only difference in procedure between this and the previous study is in the instructions given to the subjects: "Make similarity judgments based on how similar these events are, compared to all the other events in the list." There was no reference to the sound of these events, and no examples or further instructions were provided to guide subjects in making the similarity judgments.

Results

The individual subjects' mean similarity in the sound image study ranged from 1.79 to 3.27, with an overall mean similarity of 2.34, which was lower than that of the acoustic similarity study (2.98). The *SD* for the sound image study ratings (1.33) was greater than that for the acoustic similarity study (1.09). The range of mean similarities for the source image study was from 2.81 to 4.21, with an overall mean similarity of 3.58 and an *SD* of 1.31, both greater than those found in the acoustic simi-

larity study. The intersubject correlations in both of the image similarity studies were much lower than those in the acoustic similarity study ($M = .48$, without averaging across token orders). The range of correlations in the sound image study was .11 to .49, with a mean of .30; for the source image study, the range was .06 to .29, with a mean of .18. Thus, not surprisingly, there was much less agreement among listeners when the stimuli were imagined than when there were physical examples.

As in the acoustic similarity study, the similarity ratings for each subject were normalized, the similarity half-matrices were assembled, and the ALSCAL MDS algorithm applied. For consistency with the acoustic similarity results, a three-dimensional MDS solution was used in both cases. The stresses and variance accounted for are quite close to those of the acoustic similarity MDS solution (stress = .23, RSQ = .57 for sound image similarity; stress = .25, RSQ = .48 for source image similarity).

The first two dimensions of both three-dimensional solutions are quite similar to the first two dimensions of the acoustic similarity solution, as shown in Figure 3. (Scatterplots of the third dimension are not shown due to the near-zero correlation of that dimension across the three solutions.) Several of the clusters noted in the first experiment are present, such as vocalizations, impacts, and continuous sounds; the approximate boundaries of these clusters are shown. The correspondence among the three solutions is confirmed by correlations of the ordering of sounds on each dimension across the three solutions, as shown in Table 5. The pairwise correlations on Dimension 1 are quite high, ranging from $r = -.81$ to $-.93$.[3] On Dimension 2, the correlations are not nearly as strong, but are still well above chance, ranging from $r = -.47$ to $-.74$. As expected, the correlations on Dimension 3 are the weakest, from $r = -.03$ (n.s.) to $-.59$. The near-zero correlation for Dimension 3 between the acoustic similarity data and the source image similarity data is worthy of note. If the earlier interpretation for Dimension 3 for the acoustic similarity data is correct—that it represents spectral–temporal complexity—it is possible that this information is less accurately recalled when attempting to access the source events in memory. Apart from this difference, however, the three solutions agree reasonably well.

Discussion

The perceived similarity of imagined sounds and imagined source events has much in common with the similarity of those sounds when actually heard. Of course, it is impossible to verify that subjects judge stimuli entirely on the basis of the factor indicated in the instructions. Some combination of acoustic factors and source factors may have influenced judgments in all three experiments, even though subjects did report a reliance on visual imagery in Experiment 3. That the perceptual spaces did not differ substantially in the three conditions suggests that the salient source properties and the salient acoustic properties are so closely associated that either observers are unable to judge them separately or the separate judgments are made in essentially the same perceptual space.

Humans clearly learn the types of sounds reliably produced by various environmental events. While the same type of physical event can produce quite different sounds, the two different tokens are far more similar to each other than any sounds produced by different events. This regularity is used by listeners to form mental models of sound-producing events. The results indicate that these mental models may include information about both acoustic properties and event properties that influences the perceived similarity of presented sounds. The results of Experiments 2 and 3 indicate that it is probably impossible to listen to familiar sounds simply *as sounds* while ignoring knowledge of their sources and of the sounds produced by other instances of the same events.

A second finding is that there are clusters of environmental sounds that are similar whether one is listening to the sounds, imagining them, or imagining the events that produced them. An open question in the study of environmental sounds is whether these sounds constitute a unitary class. Examining ideal coding strategies in the nervous system for different types of sounds, Lewicki (2002) suggested that *animal vocalizations* and *other environmental sounds* are qualitatively distinct classes of sounds. The clusters found in the three-dimensional perceptual space derived in the present studies show a separation between what are labeled here *vocalizations*, *continuous sounds*, and *impact sounds*. The sounds that fall in between these clusters reflect the ambiguity of the world—e.g., is the sound of bowling an impact event or a continuous event?

The finding that relatively complex acoustic factors are predictive of perceived similarity leads to the question of "why those factors, and why not more fundamental ones, like loudness, duration, or frequency?" One answer is that those simple features are not invariant under transformation, to borrow the language of ecological psychology. A more revealing approach draws on the connection with sound-producing events and posits that listeners will focus on the acoustic features that best differentiate between relevant sound sources. In the hierarchical cluster analysis of the acoustic similarity MDS solution, the main distinction was between harmonic and nonharmonic sounds. The common thread among the harmonic sounds in this battery is that they nearly all have some communicative function, being either a vocalization, a signaling sound (car honking, whistle) or a musical sound (harp, bells). More simply, harmonic sounds imply highly relevant sound sources, often conveying intention. This may explain why the primary dimension of the MDS solution is best predicted by pitch salience. The other predictive acoustic features for Dimension 1—*spectral skew* and *modulation depth*—further indicate the presence of vocalizations that, as noted above, tend to have lower modulation indices (in comparison with repetitive sounds, like typing) and have spectra that tend to be positively skewed.

A similar argument for the more numerous predictors of Dimension 2 is more difficult to make, but it does seem that Dimension 2 differentiates between continuous sounds (such as water sounds) and impact sounds. The predictors generally seem to reflect properties that distinguish between these two types of sounds. For example, *spectral centroid*

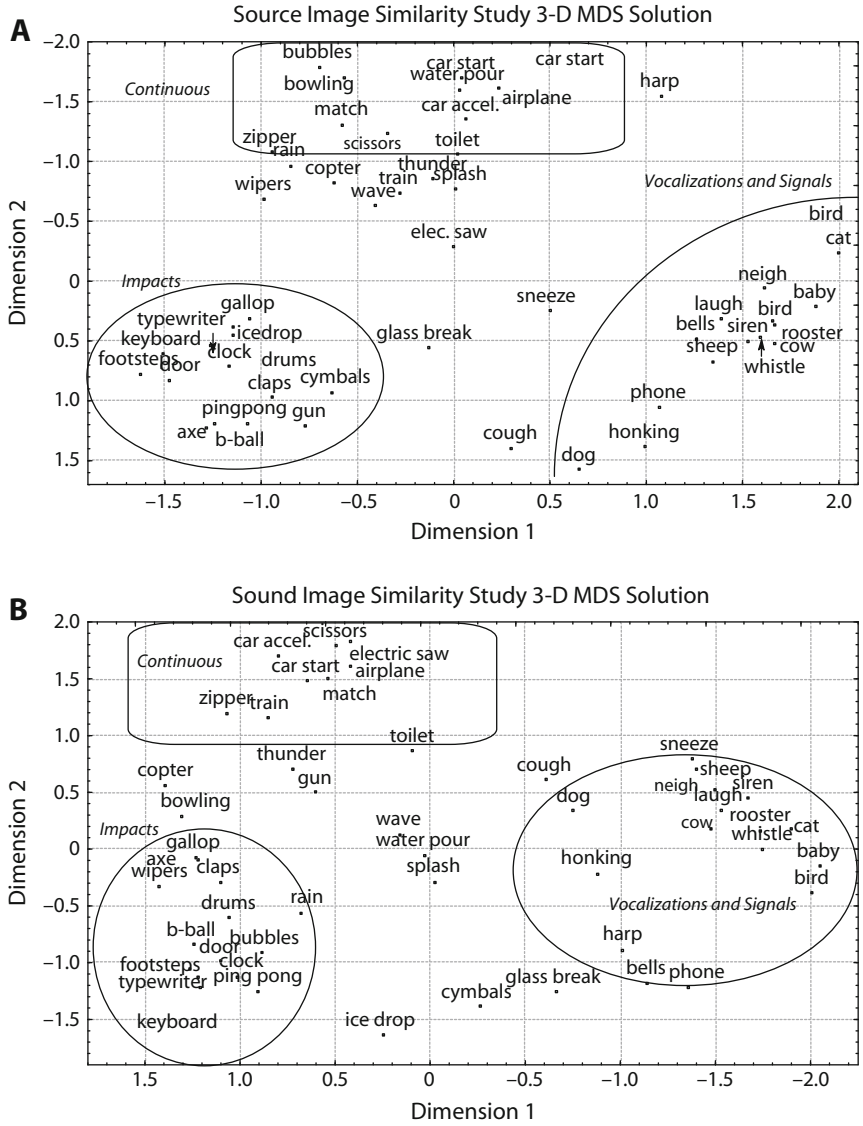

Figure 3. Three-dimensional multidimensional scaling (MDS) solutions for (A) Dimension 1 versus Dimension 2 of the source image similarity study and (B) of the sound image similarity study. The major source types are circled and marked in each case. In each plot, one axis is rotated to make the clusters line up.

(water and wind sounds have lower centroids than impact sounds, which have high transients), *spectral SD* (water sounds have broader spectra), and *envelope measures* (more steady-state envelopes with fewer peaks) are all relevant to this distinction. Since Dimension 3 does not seem to differentiate among different source types, an explanation of the role of those variables will not be attempted.

EXPERIMENT 4
Free Categorization of Environmental Sounds

A topic that is closely related to similarity scaling is categorization. In theory, the more similar two items are, the more likely they are to be grouped in the same category. It is this presumed linkage that enabled Bonebright (2001) and Vanderveer (1980) to generalize from their sorting tasks to similarity ratings. However, empirical data have shown that the relationship between categorization and similarity is complex, as discussed in some detail in Goldstone (1994). Reviewing the extensive literature on the subject, Goldstone concluded,

> Neither *similarity* nor *category* is a unitary construct—there are variations of each that are importantly different. Similarity cannot ground all category types. Still, the class of categories for which overall similarity provides a partial account are an important class because of their wide inductive potential. (p. 151; emphasis in original)

Categorization is usually considered a more cognitively based function than similarity, and it can be influenced by numerous "higher order factors" such as goals, theories,

and within-category variability (Barsalou, 1991; Fried & Holyoak, 1984; Murphy & Medin, 1985). The observation of certain consistent clusters in the similarity ratings from the first three experiments does not necessarily mean that listeners would replicate those clusters when explicitly asked to categorize the sounds. It may be expected that different characteristics of the sounds and their sources would have a stronger influence on the categorization of sounds than in the similarity judgments. To ascertain the relation between groupings derived from similarity ratings and those based on explicit categorization, a free categorization study was performed with the primary tokens used in the acoustic similarity study.

Method

Subjects. Seventeen subjects included 9 women and 8 men, all under the age of 30, and all with normal hearing, as defined earlier. All were paid for their participation.

Apparatus. This study was conducted in the Speech and Hearing Research Center at the Department of Veterans Affairs in Martinez, California. (See the Method section for Experiment 1 for a description of the apparatus.) As mentioned previously, the stimuli were the 50 primary tokens from the acoustical similarity study, which were presented at 80 dB SPL.

Procedure. Subjects were told that their task was to group environmental sounds and that they should put sounds together that "seem to belong together." They were to make no fewer than five groups and no more than 12. Examples were given using visual stimuli. Subjects were then familiarized with the sounds by hearing the complete list of 50 played through once in random order, with labels such as *Sound 1* and *Sound 2* presented as the sounds played. Subjects were told that as they listened to the sounds they should think about ways those sounds could be grouped. No other instructions were given. Consideration of either sound properties or source properties was not mentioned to the subjects.

Following the familiarization, a screen for grouping the sounds appeared. The sound labels from the familiarization procedure were arranged on the left side of screen, and grouping areas were arranged on the right side. Subjects could hear a sound by double clicking on the label. They would then group the sound by dragging its label to one of the boxes in the grouping area. Subjects were initially supplied with five grouping boxes, but could create new ones up to 12 in number. When they were finished, they saved their results.

In a follow-up session, the subjects' groupings were presented on screen for them, and they were asked to look over their groupings and make any changes they wanted. Then they were given a sheet of paper and asked to write down descriptive labels for their categories to indicate the reason behind their groupings. The average time between the first session and last session was one week; in no case was it less than three days, and only once was it as long as three weeks. The grouping changes in the second session were minimal; however, only the data from the second session were used in the analysis.

Results and Discussion

One subject used his categories to create a narrative involving the sounds; although this was creative, his data were difficult to interpret and were not used in the analysis. The 16 remaining subjects created a total of 140 different categories, with a mean of 8.75 each. The distribution was fairly even: Only one subject used the minimum five categories, and two used the maximum 12. Two independent judges, who were not part of the experiment but who had a modest familiarity with the sounds used, independently grouped the 140 different labels into a smaller set of categories into which labels with very similar meanings (e.g., *animal activity noises* and *animal sounds*) were sorted. Although the number and wording of the categories differed between the two judges, they were in very close agreement, and the list was reduced to 13 general categories based on separate discussions with the two judges. This number was fewer than the 23 general categories identified by the judges in Marcell et al. (2000); however, in that study a greater number of subjects (38) categorized a larger number of sounds (120), so a greater variance in common category labels is to be expected. There are some commonalities in the types of categories, although the categories in Marcell et al. tend to be more specific (e.g., *four-legged animal*, *farm animal*, *insect*, *pet*, and *reptile*/*amphibian*, rather than just *animal*).

The resulting list of general categories is shown in Table 6, along with the number of listeners who had created a label in each category. The nature of the general categories is quite illuminating: The most frequently used categories are ones referring to source types, such as *animals*/*people* (used by all 16 listeners), *vehicles/mechanical*, *musical*, and *water*. Lesser used were categories which grouped sounds by a context, such as *outdoor*, *sports*, and *location specific* (which included sounds grouped under *household*, *office*, and *bar* categories). Infrequently used are categories referring to simple acoustic features, such as *pitched* or *rumbling*, and emotional responses (e.g., *startling*/*annoying* and *alerting*). The tendency to categorize sounds based on source types was also found in Marcell et al. and supports Gaver's hypothesis that everyday listening is primarily oriented to the sources of sounds.

A similarity matrix can be constructed from the category assignments by tabulating the number of times sounds were grouped in the same categories (the method used in Bonebright, 2001). Some sounds, such as *wave/rain*, *cough/laugh*, and *airplane/car starting*, were grouped in the same category by all listeners, although overall the mean similarity in the matrix was quite low (2.0), reflecting the fact that nearly half the cells had a zero value (i.e., those sounds were never put in the same category). This matrix was analyzed with the same MDS procedures as the previous similarity matrices. Since these are categorical data, it is to be expected that the fits for this MDS solution would be much better than for the matrices based on the similarity ratings: The stress for the three-dimensional MDS solution is .163, and the RSQ is .806. The scatterplot for the first and second dimensions is shown in Figure 4. (Because of the near-zero correlations between Dimension 3 of the categorization data and all of the similarity studies, Dimension 3 is not included in any of the scatterplots presented here.)

An obvious difference between Figure 4 and the similarity rating scatterplots is that it is much less diffuse; the clusters are much tighter and separated from each other by greater distances. The memberships of several of the clusters are similar to those of the earlier solutions. There is a very tight cluster of animal sounds, which is contiguous to but separate from the cluster for human sounds. There are also clusters for water sounds, rhythmic sounds, and

Table 6
General Categories From the Categorization Experiment and the Number of Listeners Who Made a Basic-Level Category That Fell Into a Superordinate Category

General Category	Number of Listeners
Animals/people	16
Vehicles/ mechanical	14
Musical	11
Water/weather	10
Impact/explosion	8
Location-specific	6
Sports	6
Outdoor	4
Pitched	3
Rhythmic	3
Rumbling	3
Startling/annoying	2
Alerting	2

mechanical sounds. However, the ordering of the sounds on the three dimensions of the solution did not correlate as well with the similarity rating solutions as the similarity solutions did with each other, as shown in Table 7. The highest correlation was $r = .66$, with Dimension 1 of the sound image similarity solution. Correlations with data from all three similarity studies were considerably lower on Dimension 2 and lower still on Dimension 3. Further, the dimensions of the category-based solution correlated poorly with the acoustic variables listed earlier. No acoustic variable had a correlation on any of the dimensions higher than $r = .50$. Multiple regression solutions using the acoustic variables showed a similar predictive value. The best multiple regression model was for the third dimension, using three variables—maximum autocorrelation, number of autocorrelation peaks, and the number of bursts in the modulation spectrum—and accounted for about 40% of the variance.

The moderate resemblance of the category data to the similarity data should not be too surprising. Goldstone (1994) noted that "a good deal of evidence has found dissociations between categorization and similarity assessments, with similarity assessments grounded more in perception and categorization depending more on a categorizer's theories, goals, culture and other high-level factors" (p. 131). In general, the main basis for categorization of sounds reported here was the nature of each sound's source, as it was with similarity. However, since categories based on acoustic features such as *pitched sounds* were infrequently used, it appears that the actual acoustics of the sounds were not as important a factor in this task as it was for the similarity judgments. Moreover, other criteria used for grouping, such as *context* or *emotional affect* of the sounds, were not evident in the similarity data.

GENERAL SUMMARY AND CONCLUSIONS

The perceived similarities among environmental sounds are strongly determined by the acoustic features of those sounds, specifically including harmonicity, spectral spread, continuity, periodicity, and envelope modulation. In addition, the MDS solutions for similarity data show clusterings by type of sources, such as vocalizations, impacts, and water sounds. The experiments reported here also suggest that listeners focus on acoustic features that enable identification of important sound sources. For example, harmonic sounds often indicate either a vocalization or a signaling sound. Sounds produced by water

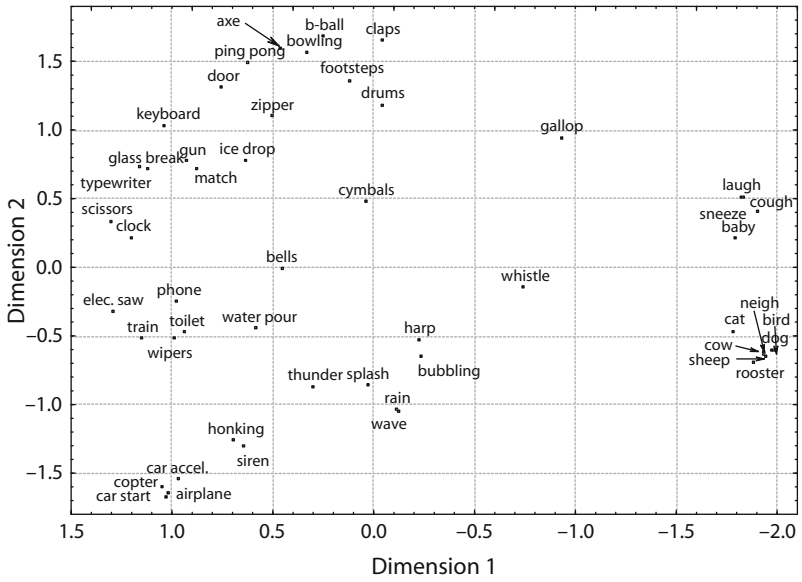

Figure 4. Dimensions 1 and 2 from the three-dimensional multidimensional scaling solution for the categorization data from Experiment 4.

Table 7
Correlations of the Orderings on the Dimensions From the Categorization Multidimensional Scaling Solution With Those of the Corresponding Dimensions of the Similarity Studies

Dimension	Acoustic Similarity: Primary Tokens	Acoustic Similarity: Secondary Tokens	Sound Image Similarity	Source Image Similarity
1	−.52	−.55	.66	−.61
2	−.42	−.46	−.40	.43
3	*.05*	*.05*	*.23*	*.08*

Note—Correlations in italics are not significant.

and wind tend to be slowly modulated broadband noises. Highly periodic noise bursts with large amounts of silence are indicative of machine sounds.

This clustering by source type is consistent across the MDS solutions for similarity ratings of actual sounds, imagined sounds, and imagined events. The first two dimensions of the solutions for each set of data are highly correlated, showing a connection between real and imagined acoustic properties, as well as between acoustic properties and the properties of imagined source events. The similarity of perceptual spaces for actual sounds and imagined sounds suggests that listeners' auditory memory for these familiar sounds is in positive agreement with the acoustic properties of the actual sounds, and it may influence judgments made when the sound is actually present. The similarity of the perceptual space for imagined events to the spaces for both real and imagined sounds suggests that either memory for complex multisensory events is largely auditory or—perhaps more likely—that event properties that influence similarity judgments are highly correlated with the salient acoustic properties (i.e., perceptually similar events make similar sounds). Yet another possibility is that when judging sound similarity, listeners are unable to ignore their ideas concerning the similarity of the underlying events, even though they are asked to judge only the sound. In other words, the data do not reveal whether the similarity between judgments of sounds and of events is due to the inability to judge sounds and events independently or to a close correspondence between the salient nonacoustic properties of events and the salient acoustic properties.

The hierarchical cluster analysis showed a primary clustering of harmonic versus nonharmonic sounds. The largest group of harmonic sounds are vocalizations, and it may be that prompt recognition of that class of sounds is a fundamental proclivity of the auditory system. This is a likely consequence of evolution, since many vocalizations are produced by things that, in the language of evolutionary biologists, are either "edible, lethal, or lovable." The free-categorization data support this conclusion, with the modal grouping being *human and animal sounds*. In general, it appears that judgments of the similarity of sounds, as well as categorization of those sounds, reflect salient properties of the sound-producing events that have significance for a listener's potential interactions with the sound's sources.

When subjects were allowed to group sounds based on their own conceptions of which sounds "belonged" together, the majority tended to group sounds based on the different types of sound sources, rather than on affective responses to those sounds. The preferred sound-source types resembled many of those which were evident in the MDS solutions for similarity rankings. These data also provide some support for Gaver's (1993) hypothesis that "everyday listening" is focused on the acoustic properties that provide information about source identity, rather than on the subjective qualities of sounds, as when listening to music. However, a test of this hypothesis would require that we reliably categorize acoustic properties as informative or uninformative with regard to source identities and compare their relative influence on perceptual judgments, perhaps utilizing similarity, identification, and discrimination tasks.

Taken together, the three experiments in this series addressing environmental sounds provide some insight into listeners' sensitivity to the information conveyed by environmental sounds and offer converging evidence about the importance of source properties in the perception of environmental sounds.

In the first study (Kidd & Watson, 2003), subjects were asked to rate the subjective qualities of a set of sounds similar to those used in the present study. Using a semantic differential technique with 20 rating scales, Kidd and Watson obtained estimates of properties such as pleasantness, harshness, and clarity. The results were factor analyzed, and the acoustic properties correlated with each factor were identified. Although the four factors (characterized as *harshness*, *complexity*, *appeal*, and *size*) do not directly correspond to the MDS dimensions identified in the present study, similar acoustic features were associated with those factors. Both *harshness* and *appeal* were associated with greater energy at higher frequencies, but appealing sounds tended to have greater pitch salience and energy variation. *Complex* sounds tended to be longer, with more variation in spectrum and amplitude over time. The *size* factor was associated with lower frequency sounds and greater total energy. Only this last factor is clearly related to judgments of a source property. However, all judgments were influenced by knowledge of source properties in that the correlation between acoustic properties and listener ratings was greatly affected by the category of sound (e.g., *impact*, *wind*, and *scrape*) being judged. Thus, whether listeners are judging the similarities of sounds (regardless of instructions to consider only the sound or only the source) or rating the quality of sounds on a wide range of semantic scales, source attributions appear to have a strong influence on judgments.

The second study in the series (Gygi et al., 2004), showed that the frequency information used for environmental sound identification tended to be in the 1200- to 2400-Hz range, the most important frequency band for speech comprehension in the Articulatory Index (French & Steinberg, 1947), with a somewhat greater influence of higher frequencies (between 6 and 8 kHz) for environmental sound identification. That study also showed that when spectral information was limited using vocoder techniques, some environmental sounds (roughly a third of the 70 sounds evaluated) required considerably more fine-grained spectral information (more channels) than that required for near-perfect identification of speech sounds (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). In addition, because of the greater influence of high frequency information, the frequency range required for environmental sound identification was broader than that required for speech comprehension. These differences likely reflect the greater range of source properties included in the broad category of environmental sounds in comparison with the source properties associated with speech sounds. However, these findings suggest that speech may have evolved to take advantage of the same sort of auditory sensitivities that have been used to identify environmental events long before speech became part of the auditory environment. This idea is consistent with recent findings showing a general familiar sound recognition ability for both speech and nonspeech sounds (presented in noise) that is distinct from other auditory abilities measured with novel laboratory-generated sounds (Kidd, Watson, & Gygi, 2007; Watson & Kidd, 2002).

Identification of the information used for auditory object and event recognition in terms of frequency ranges and general features such as harmonicity and envelope characteristics provides some useful guidelines as we search for more specific relations between source properties and acoustic properties. The similarity of the perceptual spaces derived in the different instruction conditions suggests that either the psychological spaces for source properties and acoustic properties are very similar, or subjects are unable to make independent judgments of sources and sounds (of course the two hypotheses are not mutually exclusive and may reinforce each other). Either way, the dimensions and clusters obtained in the present study provide a useful guide to the relations between perceptually relevant source and acoustic properties, and this guide may provide the basis for the identification of environmental sounds. The grouping of sounds in the perceptual spaces identified in the present study reveal subsets of sounds with many acoustic differences but with an underlying commonality in terms of their perceptually relevant (or salient) acoustic and source information. Finding the common spectral–temporal patterns associated with subsets of similar sounds will require an examination of higher order relations present in these sounds, as well as the manipulation of those relations in discrimination and identification experiments.

AUTHOR NOTE

This research was supported by Grant RO1 DC00250 from the National Institute on Deafness and Other Communicative Disorders, Grant MH12436-01 from the National Institute of Mental Health, and Grant RO1 07998 from the National Institute of Aging. Conversations with Robert Goldstone contributed materially to this research. Correspondence concerning this article may be addressed to B. Gygi, East Bay Institute for Research and Education, 150 Muir Road 151-I, Martinez, CA 94553 (e-mail: bgygi@ebire.org).

REFERENCES

ALLEN, P., & SCOLLIE, S. (2002). Stimulus set effects in the similarity ratings of unfamiliar complex sounds. *Journal of the Acoustical Society of America*, **112**, 211-218.

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

BARSALOU, L. W. (1991). Deriving categories to achieve goals. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 27, pp. 1-64). New York: Academic Press.

BIEDERMAN, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, **94**, 115-117.

BONEBRIGHT, T. L. (1996). An investigation of data collection methods for auditory stimuli: Paired comparisons versus a computer sorting task. *Behavior Research Methods, Instruments, & Computers*, **28**, 275-278.

BONEBRIGHT, T. L. (2001). *Perceptual structure of everyday sounds: A multidimensional scaling approach*. Paper presented at the 2001 International Conference on Auditory Display, Espoo, Finland.

CACLIN, A., MCADAMS, S., SMITH, B. K., & WINSBERG, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America*, **118**, 471-482.

CERMAK, G. W., & CORNILLON, P. C. (1976). Multidimensional analyses of judgments about traffic noise. *Journal of the Acoustical Society of America*, **59**, 1412-1420.

CLEARY, M., PISONI, D. B., & KIRK, K. I. (2005). Influence of voice similarity on talker discrimination in children with normal hearing and children with cochlear implants. *Journal of Speech, Language, & Hearing Research*, **48**, 204-223.

FRENCH, N. R., & STEINBERG, J. C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America*, **19**, 90-119.

FRIED, L. S., & HOLYOAK, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 234-257.

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

GOLDINGER, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **22**, 1166-1183.

GOLDSTONE, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, **52**, 125-157.

GREY, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, **61**, 1270-1277.

GREY, J. M., & MOORER, J. A. (1977). Perceptual evaluations of synthesized musical instrument tones. *Journal of the Acoustical Society of America*, **62**, 454-462.

GYGI, B., KIDD, G. R., & WATSON, C. S. (2004). Spectral–temporal factors in the identification of environmental sounds. *Journal of the Acoustical Society of America*, **115**, 1252-1265.

HALPERN, A. R., ZATORRE, R. J., BOUFFARD, M., & JOHNSON, J. A. (2004). Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia*, **42**, 1281-1292.

HEINEMANN, E. G., & CHASE, S. (1990). A quantitative model for pattern recognition. In M. L. Commons, R. J. Herrnstein, S. M. Kosslyn, & D. B. Mumford (Eds.), *Computational and clinical approaches to pattern recognition and concept formation* (Quantitative Analyses of Behavior, Vol. 9, pp. 109-126). Hillsdale, NJ: Erlbaum.

HOUTGAST, T., & STEENEKEN, H. J. M. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligi-

bility in auditoria. *Journal of the Acoustical Society of America*, **77**, 1069-1077.

HOWARD, J. H. (1977). Psychophysical structure of eight complex underwater sounds. *Journal of the Acoustical Society of America*, **62**, 149-156.

HOWARD, J. H., & BALLAS, J. A. (1983). Perception of simulated propeller cavitation. *Human Factors*, **25**, 643-655.

HOWARD, J. H., & SILVERMAN, E. B. (1976). A multidimensional scaling analysis of 16 complex sounds. *Perception & Psychophysics*, **19**, 193-200.

INTONS-PETERSON, M. J. (1980). The role of loudness in auditory imagery. *Memory & Cognition*, **8**, 385-393.

INTONS-PETERSON, M. J., RUSSELL, W., & DRESSEL, S. (1992). The role of pitch in auditory imagery. *Journal of Experimental Psychology: Human Perception & Performance*, **18**, 233-240.

IVERSON, P., & KRUMHANSL, C. L. (1993). Isolating the dynamic attributes of musical timbre. *Journal of the Acoustical Society of America*, **94**, 2595-2603.

KENDALL, R. A., & CARTERETTE, E. C. (1991). Perceptual scaling of simultaneous wind instrument timbres. *Music Perception*, **8**, 369-404.

KIDD, G. R., & WATSON, C. S. (2003). The perceptual dimensionality of environmental sounds. *Noise Control Engineering Journal*, **51**, 216-231.

KIDD, G. R., WATSON, C. S., & GYGI, B. (2007). Individual differences in auditory abilities. *Journal of the Acoustical Society of America*, **122**, 418-435.

KRUMHANSL, C. L. (1989). Why is musical timbre so hard to understand? In S. Nielzén & O. Olsson (Eds.), *Structure and perception of electroacoustic sound and music* (Excerpta Medica, Vol. 846, pp. 43-53). Amsterdam: Elsevier.

LAKATOS, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception & Psychophysics*, **62**, 1426-1439.

LECOMPTE, D. C., & WATKINS, M. J. (1993). Similarity as an organising principle in short-term memory. *Memory*, **1**, 3-22.

LEWICKI, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, **5**, 356-363.

LOH, W.-Y., & SHIH, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, **7**, 815-840.

MARCELL, M. M., BORELLA, D., GREENE, M., KERR, E., & ROGERS, S. (2000). Confrontation naming of environmental sounds. *Journal of Clinical & Experimental Neuropsychology*, **22**, 830-864.

MARR, D., & VAINA, L. [M.] (1982). Representation and recognition of the movements of shapes. *Proceedings of the Royal Society of London: Series B*, **214**, 501-524.

MCADAMS, S. (1993). Recognition of sound sources and events. In S. McAdams and E. Bigand (Eds.), *Thinking in sound: The cognitive psychology of human audition* (pp. 146-198). Oxford: Oxford University Press, Clarendon Press.

MCADAMS, S., WINSBERG, S., DONNADIEU, S., DE SOETE, G., & KRIMPHOFF, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, **58**, 177-192.

MILLER, J. R., & CARTERETTE, E. C. (1975). Perceptual space for musical structures. *Journal of the Acoustical Society of America*, **58**, 711-720.

MURPHY, G. L., & MEDIN, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, **92**, 289-316.

PLOMP, R. (1970). Timbre as a multidimensional attribute of complex tones. In R. Plomp & G. F. Smoorenburg (Eds.), *Frequency analysis and periodicity detection in hearing* (pp. 397-414). Leiden: Sijthoff.

SHAFIRO, V. (2004). Perceiving the sources of environmental sounds with a varying number of spectral channels. *Dissertation Abstracts International*, **64**, 6361.

SHANNON, R. V., ZENG, F.-G., KAMATH, V., WYGONSKI, J., & EKELID, M. (1995). Speech recognition with primarily temporal cues. *Science*, **270**, 303-304.

SHARPS, M. J., & POLLITT, B. K. (1998). Category superiority effects and the processing of auditory images. *Journal of General Psychology*, **125,** 109-116.

SHARPS, M. J., & PRICE, J. L. (1992). Auditory imagery and free recall. *Journal of General Psychology*, **119**, 81-87.

SLANEY, M. (1995). *Auditory toolbox: A MATLAB toolbox for auditory modeling work* (Apple Tech. Rep. No. 45). Cupertino, CA: Apple Computer.

VANDERVEER, N. J. (1980). Ecological acoustics: Human perception of environmental sounds. *Dissertation Abstracts International*, **40**, 4543.

WATSON, C. S., & KIDD, G. R. (2002). On the lack of association between basic auditory abilities, speech processing and other cognitive skills. *Seminars in Hearing*, **23**, 83-93.

ZATORRE, R. J., & HALPERN, A. R. (2005). Mental concerts: Musical imagery and auditory cortex. *Neuron*, **47,** 9-12.

NOTES

1. A literature search indicated that multidimensional scaling (MDS) studies of similarity of speech sounds also seem to be extremely rare. The reason for that is likely because more direct measures of the proximity of speech sounds exist (e.g., phonetic confusions).

2. Although the number of subjects is low in comparison with that in other MDS studies, the length of the study (three weeks) made it difficult to recruit and retain subjects. The large number of items to be rated, however, tends to yield results with greater intersubject agreement.

3. Since MDS solutions are rotation invariant, the sign of the correlation is not crucial.

(Manuscript received April 17, 2006;
revision accepted for publication November 20, 2006.)