Published in final edited form as:
J Exp Psychol Hum Percept Perform. 2011 April ; 37(2): 551–565. doi:10.1037/a0020671.

The Incongruency Advantage for Environmental Sounds Presented in Natural Auditory Scenes

Brian Gygi and
Veterans Affairs Northern California Health Care System, Martinez, CA

Valeriy Shafiro
Communications Disorders and Sciences, Rush University Medical Center, Chicago, IL

Abstract

The effect of context on the identification of common environmental sounds (e.g., dogs barking or cars honking) was tested by embedding them in familiar auditory background scenes (street ambience, restaurants). Initial results with subjects trained on both the scenes and the sounds to be identified showed a significant advantage of about 5 percentage points better accuracy for sounds that were contextually incongruous with the background scene (e.g., a rooster crowing in a hospital). Further studies with naïve (untrained) listeners showed that this Incongruency Advantage (IA) is level-dependent: there is no advantage for incongruent sounds lower than a Sound/Scene ratio (So/Sc) of −7.5 dB, but there is about 5 percentage points better accuracy for sounds with greater So/Sc. Testing a new group of trained listeners on a larger corpus of sounds and scenes showed that the effect is robust and not confined to specific stimulus set. Modeling using spectral-temporal measures showed that neither analyses based on acoustic features, nor semantic assessments of sound-scene congruency can account for this difference, indicating the Incongruency Advantage is a complex effect, possibly arising from the sensitivity of the auditory system to new and unexpected events, under particular listening conditions.

We live and operate in a world of dense acoustic environments, which provide us with information about a multitude of objects and events in our vicinity. The nature of that information depends on the type of signal we are listening to. Speech signals convey linguistic messages contained in a rule-driven symbolic system, along with other information relating to properties of the speaker, called indexical information (Pisoni, 1993). However, for familiar, non-speech non-musical sounds (here termed environmental sounds) in most cases what is necessary for the listener is to recover the nature of the physical source(s) that produced the sound, e.g., what it is, how big it is, or where it is, and this information is carried in the acoustics of the sound wave. Some research has uncovered acoustic features that enable source identification in the clear and under different filtering conditions (Ballas, 1993; Gygi, Kidd, & Watson, 2004; Shafiro, 2008) and which reveal properties of the source such as the distance, the shape, how hard it is, and in the case of footsteps, the gender of the walker (Carello, Anderson & Kunkler-Peck, 1998; Freed, 1990; Kunkler-Peck & Turvey, 2000; Li, Logan & Pastore, 1991; Repp, 1987; Warren & Verbrugge, 1984).

Of course, in everyday life multiple sound sources are often co-occurring and so the acoustic features that enable identification in isolation might be masked by other sounds. Numerous studies have extensively researched the effects of competing backgrounds on speech

Correspondence concerning this article should be address to Brian Gygi, Speech and Hearing Research, Veterans Affairs Northern California Health Care System, 150 Muir Road, Martinez, CA USA 94553. bgygi@ebire.org.

perception (often referred to as the “Cocktail party effect”, or CPE (Cherry, 1953), but few have addressed this question for environmental sounds (Ballas & Mullins, 1991; Leech, Gygi, Aydelott, and Dick, 2009).

In traditional psychophysical studies, background sounds are usually described as noise and modeled by quasi-random processes. However, in real world listening situations, acoustic backgrounds are seldom random but are determined by the different sources present in each setting. A recording of a park ambience will have a different spectrum than one of a street because of the different types of sound-generating objects (birds in the former case, cars in the latter). An automatic recognition system was developed (Aucouturier & Defreville, 2007) that could accurately identify the two different settings based solely on spectral information derived from the recordings (specifically the Mel Frequency Cepstral Coefficients (MFCCs)). It is reasonable to think that the auditory system would evolve to take advantage of these regularities. For instance, sounds in real settings often have slow amplitude modulation because of microturbulence in the atmosphere (Boersma, 1997; Richards & Wiley, 1980). Nelken, Rotman and Josef (1999) hypothesized that these inherent fluctuations enabled recognition of individual bird vocalizations because of comodulation masking release (CMR): a single bird call will tend to be coherently modulated at quite a different rate than the background mixture, facilitating separation.

An auditory setting can also provide information regarding the kinds of sounds that a listener could reasonably expect to encounter, analogous in some ways to the beneficial effects of grammatical and semantic context on the perception of speech sounds used in several speech tests e.g., the Speech in Noise Test (Bilger, Nuetzel, Rabinowitz & Rzeczkowski, 1984) in which high-probability words have much lower identification thresholds than low-probability words, the facilitation enabled by semantically similar words (“semantic priming”, Marslen-Wilson, 1987), or the effect of phonemes on preceding or following phonemes (e.g., phonetic context effects, Raphael, 1972; “phonemic restoration” Warren, 1970). However, grammatical context is rule-based, which severely limits the range of possible options, e.g. rules of noun – verb order eliminating some words as lawful candidates. Some research (Broadbent, 1958) has shown that in speech the facilitatory effect of grammatical context depends on the probability of the word occurring in a particular context, i.e., if there are many alternatives the facilitation is lessened. Thus, in some situations the effects of linguistic context become completely deterministic, whereas situational context in the real world is always probabilistic. We have learned through our daily listening experience what the likelihood of certain sounds occurring in certain settings are, and that there are extremely few impossible sequences[1].

Along with the definitions of context listed above, context has also been defined as stimuli that help direct listening to specific spectral-temporal regions of a target, similar to the standard definition of a prime. In the psychoacoustic literature such contexts have been found to aid detection of pure tones (Greenberg & Larkin, 1968; Schlauch & Hafter, 1991), tonal sequences (Watson, 1987) and multi tone complexes (Hafter & Saberi, 2001). However, whether naturalistic auditory contexts (here called “auditory scenes”) can facilitate identification of environmental sounds has to date not been ascertained. Some studies that have examined the effect of auditory context to date have shown that perception of sound producing objects and their properties is affected by sounds that precede or follow it in time. Ballas and Howard (1987) reported an experiment in which the same metallic clang sound was perceived as a car crash when combined with a screechy valve turning

With the ability to record and replay sounds electronically the constraints imposed on specific scenes has become less defined (e.g., a horse may be heard in a restaurant if a TV is on and is showing horses). However, some statistical boundaries still remain (as shown by the Web-based congruency ratings).

sound or as a factory machine sound when combined with water drip and noise burst. Fowler (1990) found that upon hearing the same sound of a steel ball rolling off a ramp listeners judged the steepness of the ramp differently depending on the duration of the following sound made by the ball when it continued to roll on a flat surface.

Aside from behavioral studies, there is a large body of neurophysiological research which has used environmental sounds juxtaposed with other semantically related stimuli, such as words (Orgs, Lange, Dombrowski, and Heil, 2006; Plante, Van Petten, and Senkfor, 2000; Van Petten and Rheinfelder, 1995) or pictures (Cummings, Ceponiene, Koyama, Saygin, Townsend, and Dick, 2006; Schneider, Debener, Oostenveld, and Engel, 2008). Commonly these studies utilized electroencephalographs (EEGs) and had as dependent variables the amplitude and/or latency of the N400 component. This component, which peaks at around 400 ms post-stimulus, has been found, in reading adults, to be a sensitive indicator of the semantic relationship between a stimulus and the context in which it occurs, i.e., unexpected stimuli elicit larger and earlier N400 waves than do semantically appropriate stimuli or non-semantic irregularities in a sequence (in the case of words, see Kutas and Hillyard, 1984). The studies using environmental sounds have often showed a diminished or late N400 to environmental sounds paired with semantically related sounds, words or pictures as opposed to unrelated ones, indicating that a consistent semantic context may facilitate environmental sound perception. However, the tasks involved are often either detection of a novel stimulus (the “oddball paradigm”) or a lexical decision task, which are quite different from identification of sound source events or properties which is the goal of most everyday listening.

Given the differences in definition of context, goals and methodologies, the findings of the above studies may be difficult to apply to auditory scenes directly. Nevertheless, overall these studies seem to suggest that identification of environmental sounds would be easier if they were in a context that was congruent in some manner. However, the few published studies that further examined the effects of auditory contexts have yielded quite different results. Ballas and Mullins (1991) presented target sounds that were embedded in a sequence of other sounds that provided a scenario, such as a match lighting, a fuse burning, and an explosion. The sequences were designed to be either consistent with the true source of the sound, biased towards an acoustically similar sound, or random.

The effect of consistent context significantly raised performance above that of the biased sequences and the random sequences; however it did not improve performance over the sounds presented in isolation. The authors concluded this is because each of the sounds in the contextual sequence also needed identifying, and since identification is not perfect, this would tend to reduce the beneficial effects of context. The authors interpreted this finding as showing that “…the only positive effect of consistent contexts is to offset the negative effects of embedding a sound in a series of other sounds.” This is similar to the conclusion reached in Howard and Ballas (1980), in which the stimuli were environmental sounds organized in an artificial grammar, and implies that context does not help resolve an ambiguous signal.

However, there are many factors which diminish the applicability of these results to real-world listening as well. In Ballas and Mullins’ paradigm there was no peripheral masking of the target sounds, and perhaps most importantly, no competition for attentional resources from simultaneously occurring sounds, as is typically the case in everyday listening situations, when multiple sound sources are acting at the same time (which was also a shortcoming of some of the other studies discussed earlier). Another problem, which was mentioned in conjunction with the neurophysiological studies, is that it is not clear how compelling a scene was created by two sounds presented consecutively in isolation.

A recently published study (Leech et al., 2009) used more naturalistic stimuli, environmental sounds mixed in familiar auditory scenes (which included some of the stimuli used in the Experiments described below), and tested, among other things, how the sound/scene congruency affected the detectibility of a target sound associated with a picture presented to listeners. It was found that under certain circumstances sounds which were unlikely to occur in a given context actually had about five percentage points better identification accuracy than sounds which were likely to occur in that context. Further, this effect was found with both adults and children. Thus, it seems that for detection of environmental sounds, there is the opposite of the conventional effect: incongruent contexts can actually aid detection of target sounds.

A major goal of the studies described here is to more closely simulate real-world listening, both in stimulus conditions and tasks. Thus, it would be instructive to see how more naturalistic auditory scenes and more realistic tasks affect the identification of target sounds (which, as noted above, is the more common goal of environmental sound perception). Based on the results using speech and psychoacoustic stimuli, the prediction would be that scenes would be better at priming expectations of sounds which are likely to occur in that context. However, the findings of Leech et al., which are most relevant to the stimuli and listening conditions in the present study, suggest that sounds which are incongruent with the auditory context in which they appear should be identified more accurately than congruent sounds. For identification of environmental sounds, then there are reasons to expect an advantage both for congruent or incongruent sounds. This study, which was conducted separately from and concurrently with the Leech et al. study, attempted a systematic exploration of the effects of real-world auditory contexts on environmental sound identification in order to clarify conflicting hypotheses based on previous research.

Experiments with Sounds in Scenes

The experiments described below attempted to test the identification of environmental sounds mixed with familiar, naturally occurring auditory scenes. The target sounds were either contextually “congruent”, i.e., likely to occur in a particular background scene (such as hammering at a construction site) or contextually “incongruent” (such as a horse galloping in a restaurant).

Sounds and Scenes Used

Fourteen recordings of common natural scenes were used, all of which were found to be highly identifiable in Gygi (2004a). The scenes were collected from actual field recordings donated by professional sound recordists. The recordings were all in stereo and recorded at a sampling rate of 44.1 kHz or higher (if higher, they were downsampled to 44.1 kHz). The scenes were edited to be ten seconds long, equated for overall root mean square energy (rms), with a correction for pauses of greater than 50 ms (see Gygi et al., 2004 for a description of the pause-corrected rms) and stored as raw binary files. The scenes were chosen to be a balance of indoor/outdoor and urban/rural settings. There was little intelligible human speech in the scenes and whatever speech did occur was not specific to the context of the scene and was not expected to help the listener in identification of the scene. The baseline identifiability of the scenes and method used in that study are described in Appendix A.

Thirty-one familiar environmental sounds which had been used in previous studies, (e.g., Gygi, 2004b, and Gygi et al., 2004, 2007) were mixed with the scenes. An effort was made to represent the different major sound source categories which were obtained in Gygi et al. (2007): harmonic sounds, continuous non-harmonic sounds and impact sounds. They were taken from high-quality sound effects CDs (Hollywood Leading Edge and Sound FX The

General) sampled at 44.1 kHz. The target sounds ranged in length from 457 ms (Sneeze) to 3561 ms (Baby Crying). As with the scenes, the sounds were equated for overall rms, with a correction for pauses of greater than 50 ms and stored as raw binary files. The sound/scene pairs, listed in Table 1, were designed to be either congruent or incongruent. Congruency was determined by the experimenter, in consultation with two other judges familiar with the study but who had not taken part in it. In a follow-up test, the (in)congruency of the stimulus pairs has been further validated in a separate laboratory-based study and also in another Web-based congruency rating study, (see Discussion). A complete counterbalancing of sounds and scenes was not possible – for example, it is difficult to think of a scene in which the sounds of footsteps would be unexpected. Conversely, outside of a musical performance setting, it would be very unusual to hear a harp being strummed. Consequently, although each scene had two congruent and incongruent sounds, for a total of 56 sound-scene combinations, twenty -five sounds were presented in both congruent and incongruent situations, and six in one or the other. A further constraint was that the target sound should not appear anywhere else in the scene. For example, if the target sound was a cow mooing and the scene was a barnyard scene, there would be no other cow mooing in the barnyard scene.

The target sounds were mixed in the scenes centred in the middle, beginning four seconds into the scene. The onsets and offsets of the target sounds were marked with “ding” sounds which were also centred in the mix. There was 100 ms between the first ding and the onset of the target sound. The function of the dings was to act as markers for the listening interval, explained below. The listening intervals were all 3.5 seconds long, which was longer than the longest target sound length, so the final ding always followed the offset of the target sound by at least 100 ms. Although these acoustic markers were quite salient and might have the effect of distracting the listener, pilot studies suggested that, given the length of the scenes and variety of sound sources, there was a much greater amount of uncertainty among the listeners as to what the target may be, when a more specific listening interval in the scene was not marked. The stimuli were mixed at different Sound to Scene ratios (So/Sc) described below, which were determined by comparing the pause-corrected rms of the target sounds to the pause-corrected rms of the section of the scene in which the target sounds appeared.

Procedure

In all the experiments described here, the trial procedure was the same. Listeners were seated in a soundproof booth in front of a computer terminal and keyboard. The listeners were given a sheet listing the sounds to be presented and three-letter codes with which to respond The labels for the sounds were intended to give a sense of the source objects and the events involved (e.g., CAT for cat meowing). The list provided to listeners had 90 possible codes, far more than the number of target sounds, representing a broad range of naturally-occuring sounds. The number of possible responses was greater than the number of target sounds to avoid closed set effects such as listeners using process of elimination to reduce the effective possible responses to one or two, while still ensuring standardization of responses. This same three-letter code format was used successfully for identification in Gygi, Kidd, and Watson (2004). The nature of the study was explained to the listeners and they were given two practice trials.

They were not told that some sounds would be congruent with the scenes and some not, but they were told that “some sounds may seem perfectly natural, but some sounds might be a little unusual in a particular scene.” After that they were given a short but rigorous familiarization session with the codes used in responding, in which for each possible target sound listeners were presented with a label describing a sound and told to type in the corresponding code. This was done to reduce the time listeners would spend searching for the appropriate key code.

On every trial one sound/scene mixture was presented diotically through headphones. At the same time, on the terminal a label for the scene the listener was hearing was displayed (e.g., “Street at Night”). The purpose of the label was to ensure there was no confusion on the part of the participant as to what scene they were listening to, so that he/she would focus on the cues to identifying the target rather than the scene. Since the scenes themselves had different identifiability rates, as shown in Appendix A, if the listener would not have to identify the scene, this would partially compensate for the potential confound of different scene identifiability. During the presentation of an auditory scene, at the onset of the listening interval that contained the target sound, the text box changed color from yellow to orange and remained that way until the offset of the interval, thus informing the listener of when to listen for the target sound. Listeners were instructed to identify the target sound by typing the appropriate code on the keyboard. The list of codes was always within view in front of each subject. If the listener did not respond within seven seconds, a prompt would flash on the screen, encouraging them to respond. If the listener responded with a code that was not valid, they were prompted to reenter the code. No feedback was provided. All responses were recorded electronically and saved on a file server. Reaction time was recorded but since it included the scan time for the proper code, the data will not be reported.

In each experiment which employed multiple So/Sc, all the stimuli at the lowest So/Sc were presented first in random order. Then all the stimuli at the next highest So/Sc were presented in a new random order. Listeners were given a break after completion of a particular So/Sc level. Stimuli were generated from digital files by Echo Gina 24 sound cards, amplified by the TDT System 2 headphone buffer and presented through high-quality Sennheiser 250 II headphones. Prior to testing a 1-kHz calibration tone of the same rms as the equated level of the scenes was set to 75 dB SPL at the headphones.

Experiment 1 – Experienced Listeners

In this experiment the identification of sounds in scenes was the last in a battery of four tests designed to familiarize listeners with both the sounds and scenes to be used.

- On the first day listeners were tested on their baseline identification in the quiet of 195 environmental sounds (which included the target sounds used in this study), representing multiple tokens of 51 different sound sources. The listeners identified the sounds using the same codes described above. A short familiarization session with the codes, described above, preceded the identification trials. No feedback was given on these baseline trials. These data will not be reported here, since it was largely to train the listeners on the target sounds and the codes used in responding; normative values for these sounds were reported in Gygi (2004b).
- In the next testing session the listeners rated the typicality of each of the tokens on a scale of 1 to 7.
- The next test had listeners identify the scenes themselves, using the method described in Appendix A. No feedback was given on this day of testing.

Each test required a one-hour test session to complete. Test sessions were run on separate days. Overall, after the listeners performed identification of isolated sounds and scenes in quiet, they were highly familiar and practiced with both the sounds to be identified and the scenes the sounds were embedded in before beginning the experiment described below.

Subjects

Fourteen young listeners, seven male and seven female between the ages of 18–30, were tested at the Speech and Hearing Research Laboratory at the Veterans Affairs Northern California Health Care System Martinez Outpatient Clinic in Martinez, CA. All had normal hearing as measured by pure tone audiograms (thresholds < 15 dB HL from 250 – 8000 Hz). Subjects were recruited by flyers and paid for their participation. Two of the listeners only completed trials at one So/Sc and their data were omitted from the analysis.

Stimulus Conditions

The sounds were mixed in the scenes at So/Sc of −12 and −15 dB. These levels were found in pilot studies to give a good range of performance across stimuli and subjects. As described above, all the stimuli at −15 dB were first presented in randomized order, and then the stimuli at −12 dB were presented in a different randomized order.

Results and Discussion

The p(c) for each listener in each So/Sc and congruency condition were computed, converted to rationalized arcsine units (RAU) and subjected to a repeated measures 2×2 ANOVA. There was main effect of So/Sc, $F(1, 11)=96.04$, $p<0.00001$, indicating that listeners' identification performance significantly decreased as the So/Sc ratio decreased. There was also a main effect of Congruence, $F(1, 11) =4.84$, $p<0.05$, indicating that identification performance of individual environmental sounds was affected by the type of auditory background scenes in which these sounds were heard. However, consistent with the Leech et al. (2009) study, there was an advantage in the Incongruent condition. This Incongruency Advantage (IA) was about 0.05 percentage points overall, 0.59 vs. 0.54. The results by Congruency condition are shown in Figure 1, which also makes clear that there was no interaction between So/Sc and Congruence; the Incongruency Advantage was present at both So/Sc.

Although the magnitude of IA is not large, it is significant and consistent for both So/Sc conditions. As such IA contradicts the predictions based on several studies noted earlier that congruent context should facilitate identification of individual environmental sounds. However, since this effect has not been previously reported for identification (Leech et al. studied detection), a closer examination of potential confounding factors is warranted. The subject results showed that, although there was a fair spread of performance among the subjects (two subjects achieved p(c) over 0.9 at the highest So/Sc), this result was not due to a few listeners being exceptionally good in the incongruent condition or poor in the congruent condition. At the highest So/Sc, only one listener performed better in the Congruent condition, all others showed an Incongruency advantage, which p(c) ranged from −0.03 to 0.21. The performance by subject in each condition is shown in Table 2. Nor were there some sounds that skewed the results by exceptional or poor identification in one or the other condition, as the scatterplot in Figure 2 of target sound p(c) in congruent vs. incongruent settings at −12 dB So/Sc shows equal number of sounds above and below the regression line. The correlations of p(c) for target sounds in the incongruent vs. congruent conditions were moderate, $r = 0.65$ at −12 dB and $r = 0.47$ at −15 dB. In general, it appears the Incongruency Advantage for both sounds and scenes was based on moderately better performance across a number of stimuli rather than exceptionally better performance for a few. Because of the low variance of the responses to some of the sound-scene pairs, a conventional item analysis was not possible. However, as a partial test of the effect of outlying stimuli, the responses for the target sounds that were far above and below the regression line in both directions (Crickets, Baby Crying, Splash in Water and Ice Dropping) were eliminated from analyses. This actually increased the IA at −12 db So/Sc to .08 and left the IA at −15 db So/Sc virtually unchanged.

Another potential variable that could have affected the results was that, although the target sounds were for the most part counterbalanced between the two congruency conditions, it is possible that the ones which were not counterbalanced are responsible for the effect. However, if the analysis limited to only those stimuli which included counterbalanced target sounds, the analysis is virtually unchanged, and the Incongruency advantage is still present (if anything, it is slightly more pronounced).

Therefore, it appears the Incongruency Advantage for these target environmental sounds indeed resulted from the presence of congruent or incongruent backgrounds. This finding is in line with some results from the vision literature, such as Gordon (2004) in which inconsistent objects in a visual scene had shorter RTs than consistent objects, or Loftus and Mackworth (1978) in which attention was directed more rapidly to the location of inconsistent objects (although this result is controversial). However, because all listeners in Experiment 1 were highly familiar with both individual sound and scene tokens, one potential confounding factor may be the amount of listener experience with the stimuli. Since auditory IA is a newly-reported effect, it is necessary to explore other factors that may have contributed to its manifestation.

Experiment 2 – Naive Listeners

Since the listeners in Experiment 1 were very familiar with both the specific sounds and the scenes used in constructing the stimuli, a possible confound is that they had already constructed mental scenarios regarding the sounds and scenes or based their decisions on specific details of the stimuli. If this were the case, the auditory IA effect may be limited to experienced listeners who have had extensive practice with the test stimuli, while the IA effect would be absent among naïve listeners who have never heard these environmental sounds or scenes. The first goal of Experiment 2 was to examine whether the IA is present in naïve listeners with no training on the sounds and scenes. Another goal of Experiment 2 was to test a wider range of So/Sc, to see if the IA holds across a wide range of target sound levels.

Subjects and Stimulus Conditions

The same test as in Experiment 1 was administered to new young normal hearing listeners. Except for a familiarization session with the response labels to ensure that all subjects are comfortable with the experimental procedure (identical to the one in Experiment 1), they were given no prior training or testing on the sounds or the scenes. In addition, a wider range of So/Sc was employed: −4.5, −6.0, −7.5, −9.0, −12.0 and −15.0 dB. However, no listener heard the stimuli at more than two different So/Sc and the levels given to a single subject were always 3 dB apart, resulting in three groups of listeners. Group 1 heard the stimuli −6 and −9 dB So/Sc (N = 17), Group 2 at −7.5 and −4.5 dB, (N = 19), and Group 3 heard them at −12 and −15 dB, (N=15). The listeners were tested at the Auditory Research Laboratory at Rush University Medical Center, Chicago, IL under conditions which replicated those of Experiment 1. The 51 listeners (3 males, 48 females) were between the ages of 18–30, and all had normal hearing as defined in Experiment 1. In both Groups 1 and 2 there were two listeners who did not complete the battery of tests, so their data were excluded from the analysis. As in Experiment 1, for the testing the listeners were seated in a soundproof booth before a computer terminal and the same stimuli and experimental software were used. Stimuli were presented at 70 dB SPL through Sennheiser 250 II HD headphones, and responses were made on a keyboard.

Results and Discussion

Since there were three groups of subjects, each of which were only tested at two So/Sc, an omnibus ANOVA was not possible because of missing cells. Therefore, three different repeated measures ANOVA were performed on the RAU for each subject in each condition, and significant differences between congruent and incongruent conditions at a particular So/Sc were tested using planned comparisons. The results are plotted in Figure 3. For So/Sc of −12 dB and below, there were no significant differences between the congruent and incongruent conditions, and at −15 dB the performance in the congruent condition was slightly, but not significantly higher than in the incongruent condition. However in the −9 dB So/Sc condition the advantage in the incongruent condition was 4.2 percentage points and approached significance, $F(1,14) = 1.55$ $p < 0.078$. At −7.5 dB, there appeared a significant Incongruency Advantage of a similar magnitude to that found for the experienced subjects, 79.7% vs. 75.0% $F(1,16) = 6.53$ $p < 0.021$. A slightly smaller IA was also present at −6.0 dB (76.1% vs. 71.8%, $F(1,13)=4.64$, $p<0.047$), and at −4.5 dB the IA was 1.94 percentage points and not significant. In fact the level of performance at −7.5 was slightly greater than at either −6.0 or −4.5 dB, but the differences in p(c) across levels were not significant, so this likely reflects random perturbations in p(c) at near-ceiling levels.

As was the case with experienced listeners, there were no specific combinations of sounds and scenes which could account for the Incongruency Advantage. Further, the ordering of the sounds across levels was remarkably consistent, with correlations on the p(c) for each sound-scene pair across all So/Sc ranging from $r = 0.6$ to 0.9. Therefore, Experiment 2 provides confirmatory evidence for the Incongruency Advantage, and further demonstrates that auditory IA may be level dependent: at lower So/Sc, below at about −7.5 dB, there does not appear to be any effect of the congruence of the sound with the scene. On the other hand, the presence of IA in experienced subjects in Experiment 1 at lower So/Sc rations (i.e., −12 and −15 dB), likely indicates the effect of practice with the specific sound and scene tokens, which can be identified at substantially less favorable conditions, but show a similar IA.

Both experienced and naïve listeners showed a small, about five percentage points in p(c), but significant advantage for sounds that are incongruous with their background scene. For naïve listeners this effect becomes apparent between −9.0 and −7.5 dB So/Sc. Experienced listeners showed this effect even at −12 and −15 dB So/Sc. Similarly, the Leech et al. (2009) data, which were also gathered using untrained listeners, showed an Incongruency Advantage for detection at −6 dB So/Sc, indicating that detection and identification may show similar effects for untrained listeners. No significant benefit for congruent sounds was found in any condition; although there does seem to be a tendency in that direction at −15 So/Sc where there was actually a −3 IA (i.e., three percentage points better performance in the congruent condition). It was not, however, significant because of the variance in the results. So it seems that in some situations at least there is a consistent tendency to notice unexpected or unusual sounds and to miss expected sounds.

Experiment 2 was different from Experiment 1 in that there were almost no male participants, which may have contributed to the failure to find the IA at similar So/Sc as Experiment 1. The gender breakdown for Experiment 1 showed that while the males (somewhat surprisingly) performed overall about 3–4% better than the females on sound identification, the small sample sizes mean that the differences were not significant, and, more importantly, there was an Incongruency Advantage for both groups at the same S/N (−12 and −15 dB). So the fact that Experiment 2 had all females would not likely explain why the participants in those groups did not show an IA at those So/Sc.

Experiment 3 – Experienced Listeners with an Expanded Stimulus Set

Although the Incongruency Advantage was obtained with different groups of subjects in different experimental conditions, the possibility exists that it is somehow unique to the particular sound-scene pairs used as stimuli. To test this, the same procedure as in Experiment 1 was used to test a new group of listeners, only this time the inventory of sound-scene pairs was expanded to include additional sound tokens of environmental sounds (not included in the previous experiments) and to embed them in different sound scene segments. If the IA was specific to the scene-token combinations used then no IA should be found with new sound tokens in new scene segments. Conversely, if IA still obtains with new stimuli it will demonstrate that this phenomenon is more general and robust. Hence, each sound source in Experiment 3 was represented by 3 – 5 tokens, drawn from the pool of sounds tested during baseline identification in Exp. 1.

All tokens for each of the 31 target sounds used in Experiment 1 were mixed with the 14 scenes used in Experiments 1 and 2 at So/Sc of −18, −15 and −12 dB, for a total of 217 sound-scene combinations at each So/Sc. Further, there were 36 foils made from sounds and scenes which listeners had been exposed to in the preliminary tests of the test battery, but which had not been used in the first version of the context study. These are listed in Table 1. The purpose of the foils was to mitigate any learning by participants of the specific sound-scene pairs which were used in the previous experiments, by increasing uncertainty as to the probable response alternatives. The foils were not designed to be contextually neutral with regard to context; rather, they were chosen without taking into consideration their congruence or incongruence. All told, there were 253 sound-scene combinations at each of three So/Sc, or 759 total stimuli. This study had two objectives: to expand the possible catalog of sounds subjects would listen for (in effect, increasing uncertainty), and to test for level effects on IA in highly experienced subjects.

Subjects

Fourteen new subjects (five male, nine female) were recruited, all young normal-hearing subjects as in the previous study. This study was run at the same setting as Experiment 1, using the same training and procedure except as noted in the previous paragraph.

Results and Discussion

The RAU across subjects in each condition (including the foils) were subjected to a 3x3 repeated measures ANOVA (three So/Sc and three congruency levels) and are plotted in Figure 4. As can be seen on the figure, there were no significant differences between the three congruency conditions at −18 dB So/Sc. However, at both −15 and −12 dB there were significant IAs of 4.8 percentage points ($F(1,13) = 5.67$, $p < 0.037$) and 4.9 percentage points (F (1,13) =7.36, $p < 0.018$), respectively, with the p(c) for the foils falling in between the congruent and incongruent results, which is intriguing, given that the foils were not deliberately chosen to be contextually neutral, as mentioned earlier. Further, the overall level of performance is much higher at both −15 and −12 dB than for the subjects in Experiment 1 which may be due to the increased exposure to the scenes due to the greater number of trials, enhancing the familiarity with the stimuli.

The IA across all three experiments are plotted in Figure 5 as a function of So/Sc. When viewed in conjunction with the lower p(c) at the same So/Sc for the naïve listeners, it seems that greater experience with the stimuli not only improves performance (which is to be expected, given improved performance on speech recognition with frozen noise and babble e.g., Felty, Buchwald, and Pisoni, 2009;Langhans and Kohlrausch, 1992) but does not diminish the IA; on the contrary, it causes the IA to appear at approximately 6 dB lower So/

Sc. It might be expected that greater familiarity with the scenes would somewhat mitigate the IA since the scenes would tend to be more “background”, requiring less attention. However, in testing English speech perception in babble Van Engen and Bradlow (2007) found that native English listeners were more adversely affected by English babble than by Mandarin Chinese babble. This suggests that a more familiar background may produce greater interference, possibly by increasing uncertainty for what is signal and what is background. It also may be that the more the listeners heard the scenes the less they listened to the details and just formed a template of the scene. This is examined in more detail in conjunction with the description of the “modal misses’ in the General Discussion. In any case, it seems that increasing uncertainty and overtraining the listeners not only improved overall performance, it enhanced the Incongruency Advantage at lower So/Sc ratios, although the effect still does not seem to appear at the lowest So/Sc used, −18dB. However, the difference between experienced and naïve listeners is one of degree rather than type, since the magnitude of the IA is similar in the two groups. This issue is examined further in the General Discussion.

Logistic psychometric functions plotted to the mean p(c) from Experiments 2 and 3 at each So/Sc in the two congruency conditions are shown in Figure 6 (Experiment 1 only had data at two So/Sc and so the fits would be less reliable). For the Experiment 3 data, while the So/Sc necessary for a p(c) is virtually identical in the two conditions, ~16 dB, at the 0.80 isoperformance level, there is a 2.2 dB advantage for incongruent target sounds. The differences between the two psychometric functions were tested using the psignifit toolbox for Matlab (see http://bootstrap-software.org/psignifit/) which implements the maximum-likelihood method described in Wichmann & Hill (2001). A goodness of fit test showed no significant differences in the alpha parameters, which determines the intercept of the function, but there was for the beta parameters, $p < 0.023$, resulting in a steeper slope for the incongruent target sounds, 0.047 compared to 0.037 for the congruent target sounds.

The fits to the data from Experiment 2 showed a very similar pattern: the incongruent function had a significantly steeper slope 0.060 vs. 0.049 ($p < 0.0074$), but nearly identical intercept to the congruent function. However, the difference in the 0.80 isoperformance points for the Experiment 2 functions is less than for Experiment 3, about 1.77 dB. Nevertheless, the psychometric functions for both experiments suggest that at low So/Sc there may be some benefit from congruency, which was present in the Experiment 2 data at −9 So/Sc, although not significant, and as mentioned the IA dominates at higher So/Sc. The potential congruency benefit at low So/Sc may tend to be obscured in these studies by generally poor identification performance at low So/Sc conditions.

Comparing the psychometric functions for the experienced versus the naïve listeners, the intercepts for both the congruent and incongruent functions are about 4.5 dB lower for the experienced listeners, suggesting experience with the stimuli moves the psychometric function to the left, which is somewhat less than the 6 dB advantage suggested by the raw data. However, since the Experiment 2 data were not from equally spaced So/Sc (four data points between −4.5 and −9 dB; two data points between −12 dB and −15 dB) it is possible that visual inspection might lead one to expect a greater benefit than is really suggested by the data. The naïve listeners had significantly steeper slopes for both congruent and incongruent functions than the experienced listeners, indicating they performed more like the experienced listeners at higher So/Sc.

As in the first two Experiments, the group IA seems to come about largely from stimuli showing moderate IAs. The mean IA by target sound at −12 dB So/Sc was 7.1 percentage points and 6.3 percentage points at −15 dB So/Sc.

The performance on the new tokens for each sound source in Experiment 3 was not significantly different from the tokens which were previously used in the Experiments 1 and 2. Table 3 shows the p(c) in each So/Sc and congruency condition for Experiment 3 for both sets of tokens. All of the differences between Congruent and Incongruent conditions examined separately for the old and new sound tokens were significant for each So/Sc condition except at So/Sc = −12 dB, which approached significance, $p < 0.07$. By the same token, none of the differences in identification accuracy between the old tokens and the new tokens were significant, although at −18 dB, the Congruent/Incongruent difference approached significance, $p < 0.079$. Overall, the newly introduced tokens were comparable to the previously used ones in identifiability, showing that what caused the IA was not the particular tokens that were selected.

Comparing the results of the new sound-scene pairs with the sound/scene pairs originally presented in Experiments 1 and 2 shows no significant differences between the mean p(c) in any condition, and the correlations between subjects' performance on the two classes of stimuli are extremely high, r=0.92, at both the −12 and −15 dB So/Sc. Since there was quite a bit of acoustic variability among the different tokens for a given target sound type, it seems the Incongruency Advantage is not based on the specific features of a certain target sound, but on more general cognitive aspects of a target sound class.

General Discussion

From the data presented here, it seems the Incongruency Advantage is a robust effect, being found in both experienced and naïve subject populations, and in groups tested in different locations and with different stimulus sets. Moreover, the magnitude of the effect are surprisingly consistent across different So/Sc: the IA seems to be about 4–6 percentage points at its greatest, and does not present at lower So/Sc (below −15 dB in experienced subjects and −7.5 dB in naïve subjects). Comparing the performance on sound-scene combinations common to Experiments 1 and 3 showed a strong correlation, r = 0.77, so the ordering of performance on the sound-scene combinations was fairly constant even across different groups of listeners, although of course the listeners in Experiment 3 performed much better overall. It appears that increased training on the scenes and target sounds does not necessarily affect the Incongruency Advantage, although it does improve performance overall.

Although, as detailed in the Introduction, while some of the existing literature in auditory research would initially lead one not to predict the IA, rather its opposite (with the exception of Leech et al., 2009), there is a substantial body of work which would suggest that sensory systems in general are designed to detect contrasting events, which has been evident most often in vision research (Gordon, 2004; Loftus and Mackworth 1978; Marr, 1982). One aspect common to all sensory systems is adaptation to frequently presented stimuli (see Hood, 1950 for a description of auditory adaptation), which leads to increased responses to rarely presented stimuli. Numerous neuropsychological studies have used as a dependent measure Mismatch Negativity (MMN), an auditory potential that is evoked by rare sounds, and its size depends upon the probability of the rare sounds. MMN is present even under anesthesia, showing that it is a fundamental property of the auditory system and not dependent on more central factors such as attention (Csépe, Karmos, & Molnár 1987). One model of auditory cortex posits a major function of the cortex as a change or novelty detector (Ulanovsky, Las, & Nelken 2003). Kluender, Coady and Kiefte (2003) reviewed the evidence that sensory systems function as contrast enhancers, to facilitate detection of change, and used that to explain perception of coarticulated speech.

One possible confound of the current findings mentioned earlier is the actual congruency or incongruency of the sound-scene pairs. Although the pairings were arrived at by the experimenter and verified by two independent judges, it is possible the judgments of the degree of congruency were biased. The validity of (in)congruency pairings between sounds and scenes was tested in two separate experiments, one Web-based and one laboratory-based. In each experiment, listeners were first presented the target sound for the sound-scene combination in isolation, then heard the sound mixed in the scene (at a So/Sc of 0 dB) and then made judgments as to the congruency of the sound-scene pair on a seven-point scale (1=not congruent at all, 7=totally congruent). The experimental interface and instructions given to the subjects in both cases were identical. The Web-based study was carried out over the Internet with twenty anonymous participants, none of whom took part in any of the Experiments described here. The laboratory-based study was carried out at the VA in Martinez on the subjects who had taken part in Experiment 3, but only after they had completed Experiment 3. Despite the different subject groups and experimental settings, the results were remarkably consistent: the mean rating was virtually identical, 4.279 vs. 4.282 and a correlation of r = 0.95 between the two sets of rankings.

The results for the Web study are plotted in the histogram in Figure 7. It is clear from this that the distribution of ratings for the congruent and incongruent sounds, is strongly bimodal. Most of the sound/scene pairs judged by the experimenters to be incongruent had a mean rating of 3 or below (overall mean rating 2.83), and most of the “congruent” pairings and a mean rating of 5 or above (overall mean rating 5.69). There were some significant exceptions, e.g., a Match being struck mixed in a Bowling Alley, which was judged to be congruent by the experimenters, had only a mean congruency rating of 3.8 (perhaps because the experimenters are old enough to remember when smoking was ubiquitous in bowling alleys). Conversely, Ice Dropping in a Glass in a Beach scene, grouped with the incongruent pairings, received a mean congruency of 4.2. However, eliminating the grossly misclassified pairings in each group did not change the results of the analysis, or the magnitude of IA. In addition, the congruency ratings data for the foil stimuli were obtained in this study. Since the foils were selected to increase listener uncertainty, rather than specifically to be contextually neutral, the ratings ranged from highly congruent to highly incongruent; nevertheless, the majority of ratings for the foils are in the 2.5–5 range, with a mean of 3.57 and *SD* of 1.50.

Given the previously mentioned sensitivity of the auditory system to novel events, it would be reasonable to expect that the congruency of a sound-scene pair as rated by the listeners would be related to the Incongruency Advantage: the more incongruous a sound in scene, the more likely it is to be identified correctly. However, the correlation between the mean congruency ratings of a sound-scene pair from the Web study and the p(c) were weak and non-significant in all three experiments, ranging from $r = -0.11$ to 0.28. This may be due to the possibility that listeners rated the sound-scene congruency based not on the perceived congruency of the specific target and specific auditory scene presented to them on a given trial, but rather to how likely the type of sound would be to occur in a scene in an abstract sense.

A further possible confound is the inherent identifiability of the sounds themselves. As noted in the Procedures, on the first day of testing all listeners were tested on the baseline identifiability of the target. As reported previously (Gygi, 2004b) the baseline identifiability of the sounds in quiet ranged from poor (p(c) = 0.38 for Shovel) to perfect for a number of sounds. Correlations of the p(c) for each So/Sc and congruency condition in all three Experiments with the baseline identifiability were low, from r = 0.13 to 0.30 (all *n.s.*), so it appears the identifiability of the sounds in the quiet is not a major factor in the identifiability of the sounds in natural scenes.

Despite the weak relationship between identifiability and the IA, it is possible that some sounds are particularly identifiable in certain scenes, due to spectral-temporal differences between the target sound and the scene, e.g., a quasi-harmonic sound with a strong amplitude peak such as a sneeze might stand out particularly well in a relatively steady state background such as an auto race. Previous environmental sounds research has uncovered acoustic features, such as pitch saliency and burst duration/total duration, which have accounted for a fair amount of the variance in identification results (Gygi, et al. 2004; Shafiro 2007). However, in those studies the acoustic analyses were performed on sounds presented in the clear, even though the sounds were filtered or vocoder-processed. Since in this study, the target sounds are masked by much-louder scenes, the acoustic analyses were performed on the interval of the scene containing the target sound. The analyses were similar to those done in Gygi et al. 2004 and Gygi, Kidd and Watson 2007, and a multiple regression solution was obtained using these variables to predict the identification results. The acoustic features analyzed are described in Appendix B.

The best multiple regression solution for all the stimuli which retained only significant predictors used five variables, listed in Table 4, but which only accounted for about 30% of the variance (multiple r = 0.54). This was not particularly predictive, but when the solutions were obtained independently for the congruent and incongruent stimuli, the predictive power of the best solution for the congruent sounds was much stronger (best multiple r = 0.73.51% of the variance) than for the incongruent sounds (best multiple r = 0.61, 38% of the variance). However, there were no variables in common across all three solutions, making interpretation of the results difficult. A discriminant analysis was performed to determine if there was one set of acoustic variables that could differentiate between the congruent and incongruent stimuli. The results were poor at best, with a model using four variables only achieving a p(c) = 0.57 overall (with chance = 0.50).

If it is the case that hearing a certain scene in effect activates a schema for the listener as to what types of sounds might reasonably occur, then looking at the incorrect responses might be instructive. If listeners are expecting certain sounds, then when they are uncertain about which sound was the actual target, they might be expected to default to a common sound in that scene. A list of the most common incorrect responses by scene is listed in Table 5. All of the “modal misses” for each scene are sounds that might reasonably be expected to occur in that scene, with the possible exception of Basketball bouncing in the Train Station scene (although the possible scenarios in that case are interesting to contemplate). It should be noted that none of the sounds represented by the modal misses were actually present in any of the relevant scenes. Overall, when listeners were incorrect, they tended to respond with a congruent sound. For Experiment 3, out of 4155 incorrect responses, 1661 or 40% were congruent sounds, 1383 (32%) were incongruent, and 1111 (28%) were neutral. It seems listeners did indeed have some schema in mind for these scenes, and are confabulating sounds that are likely to be present. This suggests that the Incongruency Advantage may arise out of the interplay of two separate functions of perceptual systems. One is the tendency to suppress redundant or low-information stimuli, such as predictable occurrences that do not require responses. This is a necessary component for systems that have to conserve resources. The other is the need to recognize high-information stimuli that do require responses. Perceptual systems constantly have to strike a balance between these two demands.

Finally, that the Incongruency Advantage is level-dependent might suggest some ways in which the auditory system balances those two competing requirements. First, it seems that when all the sounds in a scene are clearly audible, which occurs at high So/Sc, listeners do not need to divide their attention among all the sounds; rather, there may be a tendency to synthesize congruent sounds into a single background, so that the individual sounds are less well noticed. In this case, a sound that is unexpected or unusual will tend to “pop out” from the background. Further increases in So/Sc ratio, however, do not produce measurable IA due to nearly perfect identification performance for target sound identification which prevents the effect from being detected. On the other hand, when processing resources are limited, there may be an advantage to not having to focus on every sound in a certain auditory scene. However, at low So/Sc, when the target sounds are much less audible, subjects may need to be more analytic in their listening, to actively focus on the individual sounds, and the effort is increased, so there is no benefit from incongruency, and indeed there may be an advantage for congruent sounds, which is suggested by the psychometric functions. This is one possible reason why the IA is only found at a fairly narrow range of high So/Sc.

This may also yield an explanation for the fact that experienced listeners showed an IA at lower So/Sc, but the magnitude of the benefit was not greater. The effect of experience seems to be to allow listeners to engage in the synthetic listening mode at lower So/Sc. This is somewhat intuitive: when the scenes and target sounds are well known to the listener, the listener is better able to group the scene into a background, which allows for easier identification of unusual sounds. However, this does not seem to translate into a greater benefit in terms of p(c). It may be that there is a maximum benefit to be gained from synthetic listening. A greater understanding of underlying processes is needed to provide an adequate explanation of IA mechanisms.

There are several directions future research in this area can take. Since the effect is somewhat small in absolute terms, it would be desirable to find conditions that maximize the effect, such as multi-channel highly immersive presentation or combined audio-visual scenes. One necessary issue to explore is the effect of informative backgrounds versus non-informative background that are acoustically similar, such as speech-shaped noise. This would help to separate the peripheral and informational masking effects and address the issue that Ballas and Mullins raised regarding the effects of congruent context. There is the obvious question of how subjects know what sounds to expect. Do spectral–temporal properties of all sounds typical of a scene get activated in some way? There are numerous cognitive issues involved here. So “cocktail party effects”, similar to those found with speech, could be investigated, testing listeners’ ability to track one sound in either a congruent or incongruent setting. Finally, the effects of visual and lexical context are important to know, especially since the listeners in these experiments were altered to the identity of the scene through two different modalities, orthographic and acoustic. So it would useful to know if, e.g., does displaying the word “restaurant” or seeing a clip of a restaurant scene when identifying plates clinking embedded in white noise make the identification process better? This so far little-studied area of auditory processing can provide new information on principal questions of perception and cognition in real-world settings.

Acknowledgments

The authors would like to thank Blyth Holmes, Kendra Cooper and Kim Ho for assistance with listener testing, and Dr. Charles Watson for helpful comments on the manuscript.

This work was supported by a Merit Review Training Grant from the Department of Veterans Affairs Research Service, VA File # 06-12-00446 to Brian Gygi, and the National Institute on Aging. Partial support was also provided by the National Institute of Deafness and Other Communication Disorders (DC0008676) to Valeriy Shafiro.

References

Aucouturier, J-J.; Defreville, B. Sounds like a park: A computational technique to recognize soundscapes holistically, without source identification. Paper presented at the 19th International Congress on Acoustics; Madrid, Spain. 2007.

Ballas JA. Common factors in the identification of an assortment of brief everyday sounds. Journal of Experimental Psychology: Human Perception & Performance. 1993; 19(2):250–267. [PubMed: 8473838]

Ballas JA, Howard JH. Interpreting the language of environmental sounds. Environment & Behavior. 1987; 19(1):91–114.

Ballas JA, Mullins T. Effects of context on the identification of everyday sounds. Human Performance. 1991; 4(3):199–219.

Bilger RC, Nuetzel JM, Rabinowitz WM, Rzeczkowski C. Standardization of a test of speech perception in noise. Journal of Speech and Hearing Research. 1984; 27:32–48. [PubMed: 6717005]

Boersma HF. Characterization of the natural ambient sound environment: Measurements in open agricultural grassland. Journal of the Acoustical Society of America. 1997; 101(4):2104–2110.

Broadbent, DE. Perception and communication. New York: Pergamon Press; 1958.

Carello C, Anderson KL, Kunkler-Peck AJ. Perception of object length by sound. Psychological Science. 1998; 9(3):211–214.

Cherry C. Some experiments on the recognition of speech with one and with two ears. Journal of the Acoustical Society of America. 1953; 26:975–979.

Csépe V, Karmos G, Molnár M. Evoked potential correlates of stimulus deviance during wakefulness and sleep in cat—animal model of mismatch negativity. Electroencephalogr Clin Neurophysiol. 1987; 66:571–578. [PubMed: 2438122]

Cummings A, Ceponiene R, Koyama A, Saygin AP, Townsend J, Dick F. Auditory semantic networks for words and natural sounds. Brain Research. 2006; 1115(1):92–107. [PubMed: 16962567]

De Coensel B, Botteldooren D, De Muer T. 1/f noise in rural and urban soundscapes. Acta Acustica United with Acustica. 2003; 89:287–295.

Felty RA, Buchwald A, Pisoni DB. Adaptation to frozen babble in spoken word recognition. The Journal of the Acoustical Society of America. 2009; 125(3):EL93. [PubMed: 19275281]

Fowler CA. Sound-producing sources as objects of perception: Rate normalization and nonspeech perception. Journal of the Acoustical Society of America. 1990; 88(3):1236–1249. [PubMed: 2229661]

Freed D. Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events. Journal of the Acoustical Society of America. 1990; 87(1):311–322. [PubMed: 2299041]

Gordon RD. Attentional allocation during the perception of scenes. Journal of Experimental Psychology: Human Perception & Performance. 2004; 30:760–777. [PubMed: 15301623]

Greenberg GZ, Larkin WD. Frequency-response characteristics of auditory observers detecting signals at a single frequency in noise: The probe-signal method. Journal of the Acoustical Society of America. 1968; 44:1513–1523. [PubMed: 5702025]

Gygi, B. Speech Separation and Comprehension in Complex Acoustic Environments. Montreal, Quebec; Canada: 2004a. Parsing the blooming buzzing confusion: Identifying natural auditory scenes.

Gygi B. Studying environmental sounds the watson way. Journal of the Acoustical Society of America. 2004b; 115(5):2574.

Gygi B, Kidd GR, Watson CS. Spectral-temporal factors in the identification of environmental sounds. Journal of the Acoustical Society of America. 2004; 115(3):1252–1265. [PubMed: 15058346]

Gygi B, Kidd GR, Watson CS. Similarity and categorization of environmental sounds. Perception and Psychophysics. 2007; 69(6):839–855. [PubMed: 18018965]

Hafter ER, Saberi K. A level of stimulus representation model for auditory detection and attention. The Journal of the Acoustical Society of America. 2001; 110(3):1489. [PubMed: 11572359]

Howard JH, Ballas JA. Syntactic and semantic factors in the classification of nonspeech transient patterns. Perception & Psychophysics. 1980; 28(5):431–439. [PubMed: 7208253]

Hood JD. Studies in auditory fatigue and adaptation. Acta Oto Laryngologica. 1950; (Suppl):92.

Houtgast T, Steeneken HJM. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. Journal of the Acoustical Society of America. 1985; 77(3):1069–1077.

Kayser C, Petkov CI, Lippert M, Logothetis NK. Mechanisms for allocating auditory attention: An auditory saliency map. Current Biology. 2005; 15(21):1943. [PubMed: 16271872]

Kluender KR, Coady JA, Kiefte M. Sensitivity to change in perception of speech. Speech Communication. 2003; 41(1):59–69.

Kunkler-Peck AJ, Turvey MT. Hearing shape. Journal of Experimental Psychology: Human Perception & Performance. 2000; 26(1):279–294. [PubMed: 10696618]

Kutas M, Hillyard SA. Brain potentials during reading reflect word expectancy and semantic association. Nature. 1984; 307(5947):161. [PubMed: 6690995]

Lakatos S, Cook PC, Scavone GP. Selective attention to the parameters of a physically informed sonic model. Journal of the Acoustical Society of America. 2000; 107(5 Pt1):L31–L36. [PubMed: 10830403]

Langhans A, Kohlrausch A. Differences in auditory performance between monaural and diotic conditions. I: Masked thresholds in frozen noise. The Journal of the Acoustical Society of America. 1992; 91(6):3456. [PubMed: 1619122]

Leech R, Gygi B, Aydelott J, Dick F. Informational factors in identifying environmental sounds in natural auditory scenes. The Journal of the Acoustical Society of America. 2009; 126(6):3147. [PubMed: 20000928]

Loftus GR, Mackworth NH. Cognitive determinants of fixation location during picture viewing. Journal of Experimental Psychology: Human Perception and Performance. 1978; 4:565–572. [PubMed: 722248]

Li X, Logan RJ, Pastore RE. Perception of acoustic source characteristics: Walking sounds. Journal of the Acoustical Society of America. 1991; 90(6):3036–3049. [PubMed: 1787243]

Marr, D. Vision. New York: W.H. Freeman; 1982.

Marslen-Wilson WD. Functional parallelism in spoken word-recognition. Cognition. 1987; 25(1–2): 71. [PubMed: 3581730]

Nelken I, Rotman Y, Yosef OB. Responses of auditory-cortex neurons to structural features of natural sounds. Nature. 1999; 397:154–157. [PubMed: 9923676]

Orgs G, Lange K, Dombrowski JH, Heil M. Conceptual priming for environmental sounds and words: An erp study. Brain and Cognition. 2006; 62(3):267. [PubMed: 16793186]

Pisoni DB. Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. Speech Commun. 1993; 13(1–2):109–125.

Plante E, Van Petten C, Senkfor AJ. Electrophysiological dissociation between verbal and nonverbal semantic processing in learning disabled adults. Neuropsychologia. 2000; 38(13):1669–1684. [PubMed: 11099725]

Raphael LJ. Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in american english. The Journal of the Acoustical Society of America. 1972; 51(4 2):1296–1303. [PubMed: 5032946]

Repp BH. The sound of two hands clapping: An exploratory study. J Acoust Soc Am. 1987; 81(4): 1100–1109. [PubMed: 3571727]

Richards DG, Wiley RH. Reverberations and amplitude fluctuations in the propagation of sound in a forest: Implication for animal communication. American Naturalist. 1980; 115:381–399.

Schlauch RS, Hafter ER. Listening bandwidths and frequency uncertainty in pure-tone signal detection. Journal of the Acoustical Society of America. 1991; 90:1332–1339. [PubMed: 1939898]

Schneider TR, Debener S, Oostenveld R, Engel AK. Enhanced eeg gamma-band activity reflects multisensory semantic matching in visual-to-auditory object priming. NeuroImage. 2008; 42(3): 1244. [PubMed: 18617422]

Shafiro V. Identification of environmental sounds with varying spectral resolution. Ear and Hearing. 2008; 29(3):401–420. [PubMed: 18344871]

Slaney, M. Auditory toolbox: A matlab toolbox for auditory modeling work. 1995.

Ulanovsky N, Las L, Nelken I. Processing of low-probability sounds by cortical neurons. Nature Neuroscience. 2003; 6:391–398.

Van Engen KJ, Bradlow AR. Sentence recognition in native- and foreign-language multi-talker background noise. The Journal of the Acoustical Society of America. 2007; 121(1):519. [PubMed: 17297805]

Van Petten C, Rheinfelder H. Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. Neuropsychologia. 1995; 33(4):485–508. [PubMed: 7617157]

Voss RF, Clarke J. “1/f noise” in music: Music from 1/f noise. The Journal of the Acoustical Society of America. 1978; 63(1):258.

Warren RM. Perceptual restoration of missing speech sounds. Science. 1970; 167:393–395. [PubMed: 4311809]

Warren WH, Verbrugge RR. Auditory perception of breaking and bouncing events: A case study in ecological acoustics. Journal of Experimental Psychology: Human Perception & Performance. 1984; 10(5):704–712. [PubMed: 6238128]

Watson, CS. Uncertainty, informational masking, and the capacity of immediate auditory memory. In: Yost, W.; Watson, CS., editors. Auditory processing of complex sounds. New Jersey: Lawrence Erlbaum Associates, Inc; 1987. p. 267-277.

Wichmann FA, Hill NJ. The psychometric function: I. Fitting, sampling and goodness-of-fit. Perception and Psychophysics. 2001; 63(8):1293–1313. [PubMed: 11800458]

Appendix A. Identifying Natural Auditory Scenes

Thirty-three naturally-occurring auditory scenes were collected from various sources, such as field recording professionals, sound effects CDs or on-line databases. The files were all stereo 44100 Hz or greater and digitized at 16-bit or greater. If greater, they were downsampled to 44.1 kHz, 16-bit files. They represented different canonical scene types, combining indoor and outdoor as well as urban and rural settings. The scenes were selected to be familiar, representative and to contain multiple sources, as well as containing little or no intelligible human speech. The original durations ranged from 10 – 42 s but were edited down to an average duration of 10.42 s. They were normalized for level, as described in the Methods for Experiment 1.

The scenes were presented binaurally through headphones in a soundproof room to twenty NH listeners at 80 db SPL. Listeners were first asked to make an immediate identification in their own words. Then they could play back the scene and were asked to note as many different sounds sources as they could identify. Finally, they could change their initial identification if they so desired (almost none did). All responses were made on computer. Two judges rated the answers for correctness on a fractional scale and tabulated the number of sources correctly identified (r of the judges’ ratings = 0.96). The p(c), number of sources listed, and the number of correct sources identified by each listener are shown in Table A1. Twenty-five of the thirty-three scenes were identified with a p(c) of 0.70 or greater. There was a modest but significant correlation between the p(c) and the number of correct sources noted, $r = 0.40$, $p < .05$.

In an attempt to find acoustic variables which could predict the identification results, instantaneous power and instantaneous pitch of the scenes were calculated in the manner of Voss & Clarke (1978) and De Coensel (2003), and the slopes fitted to polynomials. The moments of the long-term power spectra were calculated and the slopes of the long-term power spectra were similarly fitted. The slopes were all calculated in log-log units. Significant correlations were found between the p(c) and instantaneous power slope, r =0.42, and between the p(c) and both the long-term power spectra *SD*, $r = 0.39$ and the instantaneous power slope, $r = 0.36$.

Appendix B

Using Matlab 7.1, the corpus of environmental sound tokens used in this study was measured for specific acoustic variables. The variables obtained reflected different spectral-temporal aspects of the sounds including statistics of the envelope, autocorrelation statistics, and moments of the long-term spectrum. The measures and a brief description are 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 last 80% of the range of amplitudes in the vector), (3) Number of Bursts (amplitude increases of at least 4 dB sustained for at least 20 ms, based on an algorithm developed by Ballas, 1993), (4) Total Duration, (5) Burst Duration/Total Duration (a measure of the ‘roughness of the envelope

Autocorrelation Statistics

Number of Peaks, Maximum, Mean Peak, *SD* of the Peaks. Peaks (as defined above) in the autocorrelation function reveal periodicities in the waveform, and the statistics of these 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, Max pitch, Mean pitch salience, Max pitch salience. The correlogram measures the pitch and pitch salience by autocorrelating in sliding 16 ms time windows. This captures spectral information and provides measures of the distribution of that information over time.

Moments of the Spectrum

Mean (Centroid), SD, skew, 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, Max Centroid Velocity. The centroid mean and SD are based on consecutive 50-ms time windows throughout the waveform. The spectral centroid velocity was calculated by measuring the change in spectral centroid across sliding 50-ms 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 a critical band wide, extracts the envelope in each band, filters the envelope with low-frequency bandpass filters (upper f_c ranging from 1 to 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 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 ms) time windows. The mean, *SD* and maximum value of the spectral flux were used in this analysis.

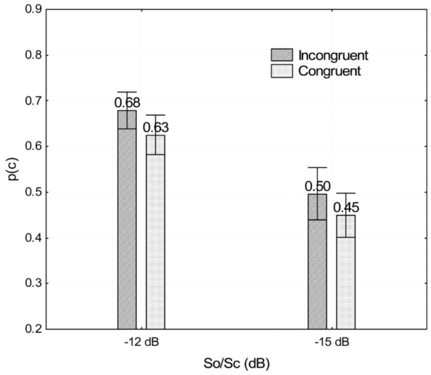

Figure 1.

Identification by congruency condition in Experiment 1.

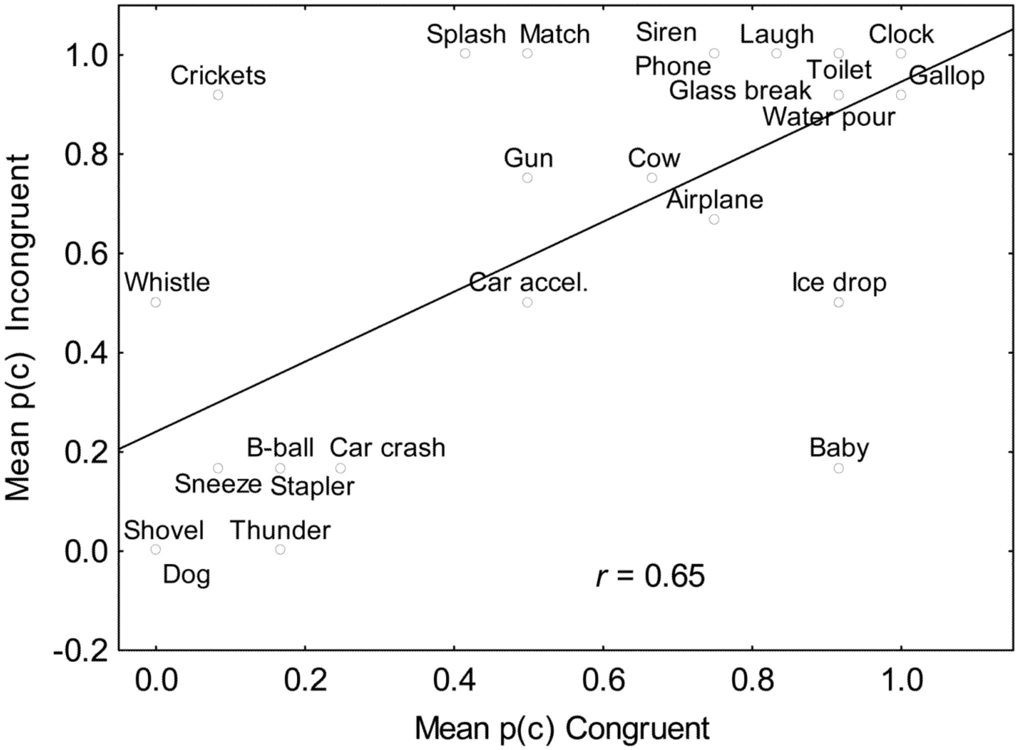

Figure 2.

Correlation of p(c) by target sound across congruency conditions in Experiment 1 at −12 dB So/Sc. The regression line is included. The positive correlation means that the relative identifiability of the target sounds did not change greatly between the two conditions.

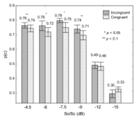

Figure 3.

Identification results for naive subjects in Experiment 2.

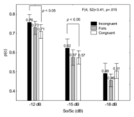

Figure 4.

Results for experienced listeners with an expanded catalog of sound-scene pairs from Experiment 3.

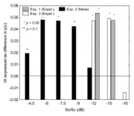

Figure 5.

IA across all three experiments as a function of So/Sc. The IA was calculated from the difference in mean p(c) between the congruent and incongruent conditions. An asterisk or double asterisk denotes a probability of being different from zero of $p < 0.05$ and $p < 0.10$, respectively. Experiments 1 & 3, using experienced listeners, showed an IA at much lower So/Sc than Experiment 2, which included naïve listeners.

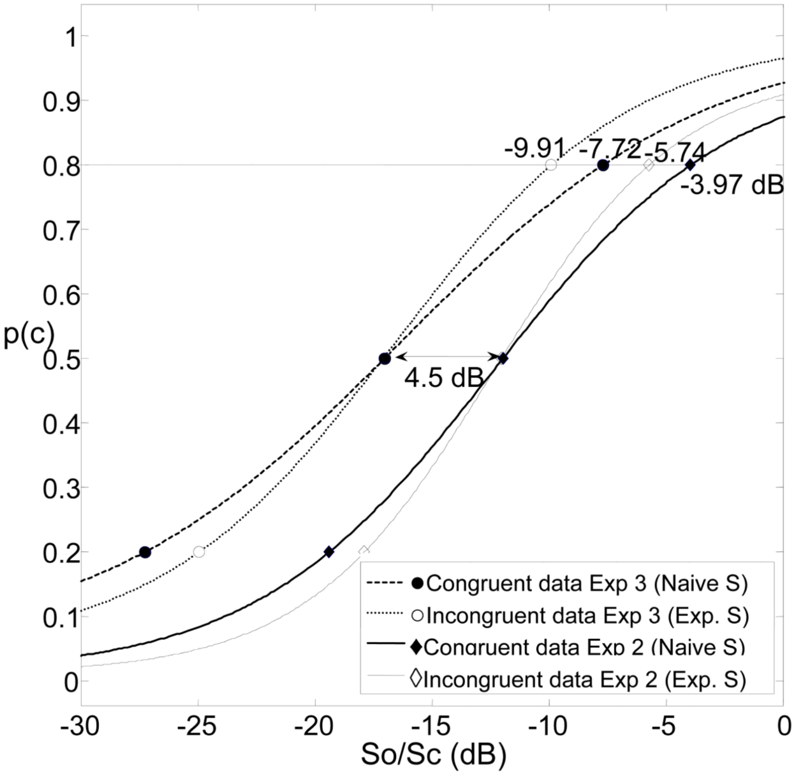

Figure 6.

Psychometric functions for the mean performance at each So/Sc in the congruent and incongruent conditions for Experiment 3. While the two functions are virtually identical at the 0.50 p(c) point, for p(c) =0.81 there is a 2.3 dB advantage for incongruent target sounds.

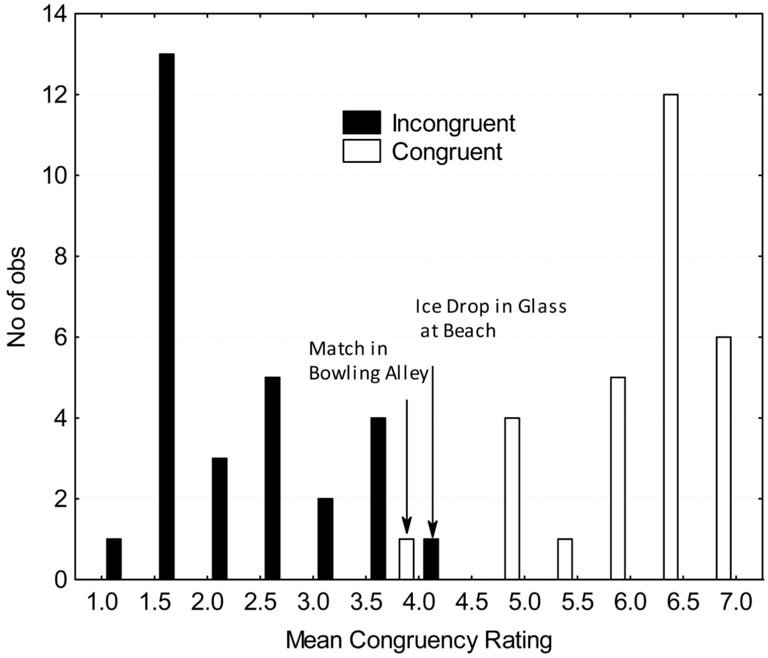

Figure 7.

Histogram of listener’s congruency ratings for sound-scene pairs vs. judges’ labeling from the Web study. Two of the outlier points for both the congruent and incongruent sounds, noted in the text, are labeled.

Table 1

List of sounds and scene used in this study. The italicized target sounds are ones which were not used in both a congruent and incongruent scene

CONGRUENT/INCONGRUENT TARGETS

Scene	Congruent Target 1	Congruent Target 2	Incongruent Target 1	Incongruent Target 2
Auto race	Car accelerating	Car crash	Sneeze	Splash in water
Beach	Dog barking	Splash in water	Ice dropped in glass	Clock ticking
Bowling alley	Laughing	Match striking	Cow mooing	Siren
Construction site	Shoveling	Whistle blowing	Baby crying	Water poured in glass
Farm	Cow mooing	Horse running	*Drums*	Match striking
Fire	Glass breaking	Siren	Laughing	Telephone
Forest	Gun firing	Crickets	Car crash	Toilet
Grocery	Baby crying	Sneeze	Car accelerating	Plane passing by
House foyer	Clock ticking	Toilet	Basketball	Shoveling
Office	Telephone	Stapler	Whistle blowing	Thunder
Playground	Thunder	Basketball	*Water bubbling*	Stapler
Restaurant	Ice dropped in glass	Water poured in glass	Horse running	Crickets
Tennis match	Plane passing by	Applause	Glass breaking	Dog barking
Train station	*Child coughing*	*Footsteps*	Gun firing	*Harp being strummed*

FOILS

Scene	Foil Target 1	Foil Target 2
Auto race	Train passing by	Bird call
Beach	Helicopter	Horse neighing
Bowling alley	Door opening and closing	Cymbals being struck
Construction site	Scissors cutting paper	Train passing by
Farm	Ping-pong ball bouncing	Cat meowing
Fire	Horse neighing	Door opening and closing
Forest	Sheep baaing	Bells sounding
Grocery	Typing on a typewriter	Scissors cutting paper
House foyer	Cat meowing	Ping-pong ball bouncing
Office	Hammering	Gargling
Playground	Bird call	Helicopter
Restaurant	Cymbals being struck	Typing on a typewriter
Tennis match	Gargling	Sheep baaing
Train station	Bells sounding	Hammering
Factory	Whistle	Cow mooing
Kitchen	Match being struck	Harp being strummed
Street at night	Footsteps	Toilet flushing
Video arcade	Laughing	Water poured in glass

Table 2

Listener performance in Experiment 1 by condition. The last two columns indicate the Incongruency Advantage for each So/Sc.

L	−12dB Incong.	−12dB Cong.	−15dB Incong.	−15dB Cong.	IA −12dB	IA −15dB
1	0.71	0.68	0.54	0.50	0.04	0.04
2	0.64	0.61	0.36	0.43	0.04	−0.07
3	0.54	0.46	0.29	0.32	0.07	−0.04
4	0.71	0.61	0.50	0.39	0.11	0.11
5	0.57	0.46	0.32	0.32	0.11	0.00
6	0.71	0.68	0.64	0.46	0.04	0.18
7	0.61	0.46	0.54	0.29	0.14	0.25
8	0.93	0.86	0.75	0.50	0.07	0.25
9	0.54	0.54	0.32	0.39	0.00	−0.07
10	0.96	0.93	0.93	0.93	0.04	0.00
11	0.61	0.57	0.39	0.36	0.04	0.04
12	0.61	0.64	0.39	0.50	−0.04	−0.11
Mean	0.68	0.63	0.50	0.45	0.05	0.05
SD	0.14	0.15	0.20	0.17	0.05	0.12

Table 3

P(c) in Experiment 3 for the newly-introduced tokens for each sound source as opposed to the tokens which had been previously used in Experiments 1 and 2. All of the differences between Congruent and Incongruent were significant ($p < 0.05$) except for the comparison marked in italics which approached significance, $p < 0.07$. None of the differences between the previous tokens and the new tokens were significant, although the comparison in boldface approached significance, $p < 0.062$.

So/Sc	−12 dB		−15dB		−18dB	
	Incongruent	Congruent	Incongruent	Congruent	Incongruent	Congruent
Old Tokens	*0.73*	*0.70*	0.60	0.55	0.47	**0.47**
New Tokens	0.76	0.71	0.62	0.58	0.45	**0.51**

Table 4

Multiple Regression Analyses Using Acoustic Predictors

All stimuli

R= 0.54 R^2= 0.30 Adjusted R^2= 0.28

Acoustic Variable	Beta	B
Spectral SD	0.36	0.00
No. of Peaks	0.33	0.03
Range of Autocorrelation Peaks	−0.15	−2.09
rms in Fc = 2000 Hz	−0.40	−2.83
rms in Fc = 8000 Hz	−0.24	−3.33

Congruent Stimuli

R= 0.61 R^2= 0.38 Adjusted R^2= 0.35

Acoustic Variable	Beta	B
Spectral SD	0.49	0.00
No. of Bursts	0.61	0.46
Max. Modulation Spectrum Burst	−0.56	−0.05
Range of Autocorrelation Peaks	−0.42	−8.28
Mean Cross-Channel Correlation	−0.42	−2.15

Incongruent Stimuli

R= 0.73 R^2= 0.53 Adjusted R^2= 0.51

Acoustic Variable	Beta	B
Mean Saliency	−1.17	−2.94
Max. Saliency	1.25	2.59
No. of Bursts	−0.36	−0.32
Mean Autocorrelation Peak	0.49	21.67
rms in Fc = 500 Hz	0.32	1.52

Table 5

Modal incorrect responses by scene

Scene	Response	No. or Responses
Beach	Waves crashing	156
Bowling alley	Bowling	97
Construction site	Dog barking	86
Tennis match	Hitting tennis ball	83
Office	Whistle blowing	77
Playground	Glass breaking	73
Grocery	Baby crying	55
Forest	Bird calling	46
House foyer	Door opening & closing	43
Train station	Basketball bouncing	36
Farm	Cow mooing	28
Restaurant	Person laughing	27
Auto race	Airplane flying	24
Street at night	Dog barking	20
Fire	Tree falling	16
Factory	Toilet flushing	8
Kitchen	Frying food	6
Video arcade	Person coughing	5

Table A1

Identification of natural scenes used as backgrounds in Experiments 1–3.

Scene	p(c)	No. of Correct Sources Named	No. of Total Sources Named	Scene	p(c)	No. of Correct Sources Named	No. of Total Sources Named
Train station*	0.95	2.60	3.80	Rural night	0.80	3.80	4.10
Forest fire*	0.95	2.50	2.60	Video arcade**	0.80	3.20	3.30
Playground*	0.95	3.10	3.70	Bar	0.75	3.20	3.70
Farm*	0.90	2.70	4.00	Casino	0.75	3.10	3.50
Bowling alley*	0.90	3.50	4.20	Xmas gathering	0.75	3.00	3.60
Beach*	0.90	3.30	3.80	Horse race	0.75	2.40	3.10
Forest*	0.90	2.70	3.60	Hockey game	0.70	3.30	3.90
Hospital	0.90	3.60	3.70	Garage	0.70	3.00	3.80
Street protest	0.90	3.50	3.60	Market	0.65	3.60	3.70
Restaurant*	0.90	2.70	3.20	Construction site*	0.60	2.80	3.00
Tennis match*	0.90	3.20	3.60	Factory**	0.56	2.56	3.22
Fireworks	0.85	3.00	3.50	ATM	0.45	3.20	4.10
Grocery*	0.85	3.90	4.60	Street at night**	0.40	1.50	2.00
House foyer*	0.85	4.40	5.10	Laundromat	0.35	1.50	2.00
Kitchen**	0.80	1.80	2.80	Library	0.20	2.50	3.60
Crosswalk	0.80	2.50	3.60	Office*	0.20	3.30	3.80
Auto race*	0.80	2.50	2.50	*Mean*	*0.74*	*2.95*	*3.52*

The scenes used in the congruent/incongruent stimuli are marked with an asterisk (*).

The scenes that were only used for foil stimuli are marked with a double asterisk (**).