Informational factors in identifying environmental sounds in natural auditory scenes

Robert Leech
Division of Neuroscience and Mental Health, Imperial College, London W12 0NN, United Kingdom

Brian Gygi
East Bay Institute for Research and Education, Martinez, California 94553

Jennifer Aydelott and Frederic Dick
School of Psychology, Birkbeck, University of London, London WC1E 7HX, United Kingdom

(Received 15 September 2008; revised 2 August 2009; accepted 24 August 2009)

In a non-linguistic analog of the "cocktail-party" scenario, informational and contextual factors were found to affect the recognition of everyday environmental sounds embedded in naturalistic auditory scenes. Short environmental sound targets were presented in a dichotic background scene composed of either a single stereo background scene or a composite background scene created by playing different background scenes to the different ears. The side of presentation, time of onset, and number of target sounds were varied across trials to increase the uncertainty for the participant. Half the sounds were contextually congruent with the background sound (i.e., consistent with the meaningful real-world sound environment represented in the auditory scene) and half were incongruent. The presence of a single competing background scene decreased identification accuracy, suggesting an informational masking effect. In tandem, there was a contextual pop-out effect, with contextually incongruent sounds identified more accurately. However, when targets were incongruent with the real-world context of the background scene, informational masking was reduced. Acoustic analyses suggested that this contextual pop-out effect was driven by a mixture of perceptual differences between the target and background, as well as by higher-level cognitive factors. These findings indicate that identification of environmental sounds in naturalistic backgrounds is an active process that requires integrating perceptual, attentional, and cognitive resources. © *2009 Acoustical Society of America.* [DOI: 10.1121/1.3238160]

PACS number(s): 43.66.Dc, 43.66.Rq [RYL] Pages: 3147–3155

I. INTRODUCTION

In research on language comprehension, it has long been recognized that we can isolate an individual voice among many competing talkers, e.g., the famous "cocktail-party scenario" (Broadbent, 1958; Cherry, 1953). Listeners can attend to an individual voice in a multitalker environment by focusing on its perceptual characteristics (i.e., the distinctive acoustic properties associated with differences in vocal quality, pitch, and vocal tract length; Brungart, 2001; Darwin and Hukin, 2000; Bregman, 1990; Brokx and Nootebohm, 1982) and its spatial location (i.e., differences in the timing and intensity of the acoustic signal as it reaches the different ears; Cherry, 1953; Darwin and Hukin, 2000; Drennan *et al.*, 2003; Freyman *et al.*, 2001; Hawley *et al.*, 2004). While listeners are able to use these different cues to attend to a given speaker, the presence of competing speech makes language comprehension more difficult by masking and/or diverting attention from the target signal (e.g., Brungart and Simpson, 2002), and by introducing conflicting meaningful content (e.g., Moll *et al.*, 2001; Leech *et al.*, 2007).

Perhaps surprisingly, research on listeners' comprehension of similarly complex everyday non-linguistic auditory scenes (such as identifying the sound of a car start on a busy street) has not been the focus of such intensive research (but cf. Bregman, 1990; Ballas and Mullins, 1991; Gygi and Shafiro, 2007). This is despite the fact that environmental sounds are important and highly frequent auditory events encountered in everyday life (Ballas, 1993). Most studies into environmental sound processing have assumed ideal listening conditions, that is, identifying a single sound without accompanying distracting noise (e.g., Gygi *et al.*, 2004, 2007; Saygin *et al.*, 2005). To our knowledge, only a few studies (Ballas and Mullins, 1991; Oh and Lutfi, 1999; Kidd *et al.*, 2007; Gygi and Shafiro, 2007) have investigated environmental sound processing in the presence of noise or other distractors. This is despite the fact that ecologically important environmental sounds that need attending to (such as a cell phone ring or far-off police siren) can often occur in the midst of multiple auditory scenes that can also be spatially segregated to greater or lesser degrees. For instance, when a pedestrian walks along a city street lined with shops, traffic and construction sounds may predominate in the ear directly exposed to the street, whereas a blend of sounds (music, conversation, and air conditioning) may be more dominant in the ear facing away from the street.

The effects of such competing signals on sound processing have been addressed in speech perception research on so-called energetic and informational masking of auditory targets. Such studies have shown that interference from additional sources of acoustic information is greater than would be predicted by filter models of the auditory periphery (see

Durlach *et al.*, 2003). Informational masking has often been investigated by measuring the masking effect on speech identification of having a simultaneously presented competing speaker in the unattended ear (Brungart and Simpson, 2002; Brungart *et al.*, 2005; Wightman and Kistler, 2005). In these studies, the presence of competing speech in both the contralateral and ipsilateral ears makes it more challenging to detect a target word than purely ipsilateral masking, particularly when the target word is presented at low signal-to-noise levels (Brungart and Simpson, 2002). Such informational masking effects are also observed for non-linguistic stimuli, for instance, spatially segregated, simultaneously presented multitone masking of pure tones (Kidd *et al.*, 2003; Wightman *et al.*, 2003). Interestingly, such masking effects can also be affected by the relative familiarity of the masker. In a study of pure-tone detection, Oh and Lutfi (1999) observed that when spectrally sampled environmental sounds were used as maskers for pure tones, those environmental sounds rated as familiar evoked considerably reduced masking of the tone target relative to unfamiliar sounds.

Target sound detection is affected not only by other distracting sounds presented simultaneously, but by the characteristics of the *ongoing* context that precedes and accompanies the target. Even with simple artificial sounds such as pure tones, frequency-swept, or frequency modulated tones, the degree of *acoustic* overlap between a target sound and sequentially (as well as simultaneously) presented background sounds can affect detection, such that decreasing similarity leads to decreased detection thresholds and increased detection accuracy (Durlach *et al.*, 2003; Cusack and Carlyon, 2003). Similarity of non-target stimuli can also degrade performance in some linguistic tasks. For example, in phoneme monitoring, target phonemes are more salient and easier to detect if they are perceptually incongruent with phonemes that were contained in the *preceding* word in the speech stream (Stemberger *et al.*, 1985). (It is worth noting that phonemes are more like meaningful sounds than a non-speech tone in that their accurate perception requires a many-to-one mapping from acoustics to phonemic category.) Such results suggest that listeners are building up perceptual expectations based on what they have heard previously.

In addition to such perceptually driven "lead-up" effects on sound detection, contextual effects linked to higher-level cognitive processes such as associative memory and semantic processing have also been shown to have effects on sound target detection. Ballas and Mullins (1991) found that the meaningful, real-world contextual information in a sequence of environmental sounds could bias the subsequent identification of an acoustically ambiguous target sound that was embedded in that sequence, such that the preceding sound context could lead listeners to incorrectly identify the ambiguous sound as one that fits with the context (e.g., the sound of a burning fuse misidentified as food frying when preceded by sounds associated with food preparation). Further, Gygi and Shafiro (2007) reported a preliminary study showing that unambiguous environmental sound targets are easier to detect and identify when they occur in an ongoing auditory scene where they are contextually incongruent (e.g., a doorbell presented in the context of a barn scene). Thus, a pop-out effect emerges for targets that are incompatible with the meaningful background in which they appear.

THE PRESENT STUDY

The current experiment builds on this initial work (in particular, Gygi and Shafiro, 2007) and introduces a novel experimental paradigm designed to investigate how detection and identification of short, naturalistic environmental sound targets presented in one ear or the other are affected by the auditory background scene in which the target is heard. Here, participants listened to a series of background auditory scenes, such as a seashore, a restaurant, and so forth. Some trials featured only a single stereophonic background scene, whereas others featured dual monaural scenes, e.g., a seashore scene was played to one ear, and a restaurant scene was simultaneously played to the other. Before listening to each scene, participants were shown one to three photographs of objects, and were played the short sounds associated with those objects (e.g., a glass breaking or a cow mooing). These short sounds—presented at different signal intensities—were embedded in one channel of the background scene that was presented next. The participant's task was to listen to the scene, and as soon as she/he heard one of the target short sounds, to push the button under the picture corresponding to that sound.

This paradigm requires the participant to perform a number of tasks that vary in their degree of difficulty depending on the characteristics of the background and target stimuli. In all conditions, the listener must simultaneously attend to two continuous auditory streams while maintaining one or more target sounds in memory. The listener must then isolate the target(s) from the background sounds (*detection*), and match each individual target to its corresponding picture (*identification*). Target detection depends on the successful perceptual separation of the target and background signals, which is influenced by the level of energetic and informational masking imposed on the target by the background. In multiple competing backgrounds, the total number of elements in the auditory scene is substantially increased, making detection more difficult by introducing a greater number of potential targets. Further, as noted above, previous studies of speech identification (Brungart and Simpson, 2002) and pure-tone detection (Kidd *et al.*, 2003; Wightman *et al.*, 2003) indicate that the presence of multiple competing backgrounds increases informational masking effects relative to a single background, resulting in decreased accuracy, particularly at low signal-to-noise ratios (SNRs). It was predicted that similar results would emerge for naturalistic environmental sounds.

In addition, the findings of Gygi and Shafiro (2007) suggest that the detection and identification of environmental sounds can be influenced by their compatibility with the meaningful context represented by the background auditory scene. Multiple factors may contribute to the contextual pop-out effect observed for incongruent targets. One possibility is that sounds from the same real-world auditory environment tend to share perceptual characteristics, such that targets may be acoustically more similar to compatible backgrounds than

incompatible backgrounds. This account predicts that the more accurate detection of incongruent targets can be accounted for in terms of the acoustic properties of the stimuli, specifically the degree of overlap between targets and backgrounds.

Another possibility is that the meaningful background activates associated real-world knowledge in semantic memory, allowing the listener to generate expectancies about the sound events that are likely to occur within a given auditory scene. Incongruent targets would violate these expectancies, which might serve to enhance their salience relative to the competing background sounds, thereby contributing to the pop-out effect. In evaluating this interpretation, however, it is important to consider the separate tasks of detection and identification, and how these may be affected by context-generated expectancies. More easily detected target stimuli may also be more readily identified, in which case both detection and identification accuracy should be greater for incongruent targets. Alternatively, stimuli that violate expectancies may be easier to detect, but their incompatibility with active memory representations may interfere with identification.

In both of the above accounts, the presence of multiple competing backgrounds would be expected to reduce or eliminate the pop-out effect, by increasing the complexity of the acoustic signal so that incongruent targets no longer stand out in terms of their perceptual characteristics, and/or by disrupting the activation of a set of memory representations that are consistent with a single, unified auditory scene.

These possibilities were investigated in the present study by manipulating various aspects of the experimental paradigm.

(1) The effect of relative signal intensity on the detection and identification of short, naturally occurring sound targets by other natural background scenes was assessed by manipulating target/background SNR at four different levels: −6, −3, 0, and +3 dB.
(2) Informational masking of sound targets was assessed by comparing performance on sound targets presented in one channel of a stereophonic single background (e.g., the sound of a dog barking embedded in the left channel of a stereo recording of a beach scene) versus a target presented in one channel of dual monaural scenes (e.g., the dog bark embedded in the beach scene, which is played to the left ear, with a traffic scene played to the right ear).
(3) As observed above, previous studies (e.g., Brungart and Simpson, 2002) have shown that informational masking may be enhanced when targets are presented at lower SNRs. This possibility is investigated here by testing for an interaction between SNR and number of background scenes, where increased informational masking should be observed when targets are presented at lower SNR, but not at higher SNR.
(4) Following on from Gygi and Shafiro (2007), contextual influences were assessed by comparing performance on sound targets that were embedded in background scenes that were contextually congruent or incongruent with the target (e.g., a cow moo in a farm versus factory background). In order to tease apart more expectancy-driven contextual effects from those due to acoustic similarities between target and background sounds, both contextual and acoustical predictors were entered into the regression model used to analyze the data.
(5) Finally, contextual "pop-out" effects might be lessened or extinguished by additional informational masking. This can be shown by testing for the interaction between the number of background scenes and contextual congruence of the target and background.

II. METHOD

A. Participants

48 adults participated in the experiment (25 females, mean age 35.5 years, range, 19–46 years). 26 of these participants were recruited specifically for the experiment and a further 22 took part in the study as part of a larger battery of auditory and language skills. All participants reported normal hearing and were native British English speakers.

B. Stimuli

Stimuli were 23 single environmental sounds, which were used as targets and 10 natural auditory scenes, used as backgrounds. Each target sound had a color image (250 × 250 pixels) associated with it, which corresponded to the real-world object being represented. All stimuli (backgrounds and targets) were meaningful, naturally occurring sounds, e.g., a "barn sound" background or a "dog bark" target sound. The target and background sounds were selected because they were reliably identified with appropriate verbal labels in previous studies: The background sounds were taken from Gygi and Shafiro (2007) and the target sounds from Cummings *et al.* (2006). The sounds were chosen to represent a wide range of sound categories (e.g., animals, vocal sounds, machines, actions, nature, etc). Target and background sounds were taken as edited in the previous studies and resampled to 22 050 Hz. Background auditory scenes varied in length from 8.00 to 14.91 s, and were scaled to have the same average intensity using PRAAT (Boersma and Weenink, 2009).

III. EXPERIMENTAL DESIGN

This was a 3-within-subjects factorial design. Within-subjects factors were background type [single background/dual background], target/background SNR [low (−6 dB, −3 dB) and high (0 dB, +3 dB)], and target/background contextual congruence [congruent/incongruent].

There were 40 trials in the experiment. In each trial, either a single stereophonic background sound was played in the left and right channels, or two distinct mono background sounds were presented in separate channels, e.g., a casino scene in the left channel, and a beach scene in the right channel. In the case of two different background scenes, the longer scene's duration was edited to match the duration of the shorter scene. One, two, or three short mono target sounds (e.g., "car horn," "door knock," and "flute") were

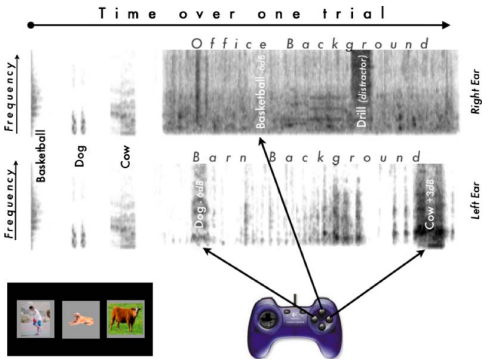

FIG. 1. (Color online) Schematic of the experiment design and presentation.

mixed into one channel of the background sound (see Fig. 1). An example of a congruent sound would be the car horn target sound mixed with the left channel of the "racing cars'" background sound. Multiple target sounds could appear in the same or different channels; e.g., two targets could appear in the left channel, and one in the right. Each target sound was either contextually congruent or incongruent with the background—for instance, the "sheep baaing" target sound was considered congruent with the "barn" background sound, but incongruent with the "office" background. Binary contextual congruence was assessed subjectively by the authors; a full list of target sounds and background scenes along with can be viewed in the supplementary materials (see EPAPS). In addition, acoustic measures of the similarity between each target and its ipsilateral background were calculated (see Sec. III B). Each target occurred twice in a congruent background context and twice in an incongruent background.

The 23 sound targets each occurred at four SNRs (−6, −3, 0, and 3 dB), with SNR calculated relative to the background sound (using rms, scaled in PRAAT) in the same auditory channel over the time window the target was being presented. (Choice of SNR levels was based on results from pilot work and previous studies; SNRs of −6 and −3 dB were classed as "low SNR," and 0 and +3 dB as "high SNR.") There were 92 target sounds presented over the 40 trials of the experiment, with 8 trials containing one target, 12 containing two targets, and 20 containing three targets. In addition, each target sound was also presented once as a contextually incongruent distractor, where the sound was played at 0 dB SNR without any associated picture. There was at most one incongruent distractor sound in any trial.

Target sound assignment was counterbalanced across background and contextual congruence conditions, with each target sound appearing once in each of the dual/congruent, dual/incongruent, single/congruent, and single/incongruent conditions. For each participant, the same number of targets was presented to each ear in each background and target contextual congruence.

Target-to-background signal-to-noise was counterbalanced within individual participants, and an additional level of counterbalancing in SNR levels *across* participants was employed to eliminate potential item-specific SNR effects. Participants were assigned to 1 of 4 different stimulus sets (2 sets were presented to 13 participants each, and 2 sets to 11 participants each), with each set varying in terms of specific target/background SNRs such that across participants, each target sound in each possible combination of contextual and background conditions appeared in every SNR. Similarly, for each trial, the location of the target pictures on the screen (left, middle, or right) was counterbalanced across participants. Finally, individual pairings of target side, background condition, and target contextual congruence were counterbalanced by reversing headphone side (left ear→right headphone) in half of the participants.

A. Procedure

The experiment was presented on an Apple G4 14-in. iBook laptop using the Psychophysics toolbox (Brainard, 1997) in MATLAB. Sounds were presented through Sennheiser HD 25-1 headphones, and participant responses were collected using a Logitech Precision gamepad. Participants held the gamepad while sitting in front of the laptop placed on a desk. Before starting the experiment, participants watched a 2-min-long QuickTime movie, which explained the task and provided an initial two practice trials. Before the main test session participants had three practice trials, which provided visual feedback. After these practice trials, participants received no further feedback beyond general encouragement.

Each trial had two parts: a target/picture familiarization phase and the main target detection phase. During the familiarization phase, three gray boxes appeared on the left, middle, and right of the screen. Depending on the number of target sounds, one to three target-associated pictures would appear in the boxes; pictures remained on the screen until the end of the trial (see Fig. 1). After a 500 ms silence, the target sound corresponding to the left-most picture was presented. After another 500 ms gap, the target sound corresponding to the next picture to the right was presented in sequence. There was 1.0 s of silence between the offset of the last target and the beginning of the main detection phase.

Participants heard the stereo background sound with between one and three mono target sounds inserted onto either the right or left channel. An incongruent distractor sound was present on half of all trials. Participants responded with one of three buttons on the gamepad, which corresponded to one of the three target boxes. For instance, if the participant heard the target represented by the picture on the left, the participant pressed the left button as soon as they heard the sound. Participants were not told whether the target sounds would appear in the left or right channel. Responses were recorded throughout the entire trial, and counted as "correct" only if participants pushed the button corresponding to the target within a 2-s time window starting 300 ms after target onset.

B. Acoustic measures

As noted above, congruence between the ipsilateral background and the target sound could arise from either acoustic similarity between the sounds and/or via real-world associative or semantic knowledge. To establish whether

contextual congruence between background and target was driven by acoustical properties of the sounds alone, a range of spectral, envelope, and periodicity similarity measures was computed. The acoustic measures were the same as those used in Gygi *et al.*, 2007 to characterize variation in environmental sounds. All measures were calculated using MATLAB (Mathworks, Natick, MA).

The predictive value of the acoustic measures for detection and identification of a sound target and for the degree of masking with ipsilateral background sounds was assessed by comparing each acoustic measure for each sound with mean identification accuracy. This was conducted in two ways: first with univariate models with each acoustic variable predicting response accuracy, and second with a stepwise multiple regression model that also included the binary experimenter-defined contextual congruence and the number of distinct background sounds (dual or single).

In addition, for each of the 38 acoustic variables, a measure of the similarity between each occurrence of target and ipsilateral background sounds was calculated as follows: (1) For each target sound, its time-matched segment of ipsilateral background sound was extracted; and (2) for each acoustic measurement, the absolute difference between the measurements for the ipsilateral background and target sounds was calculated. This resulted in an approximate similarity distance for each target/background pairing, which was subsequently entered into stepwise multiple regression models.

The acoustic measures were as follows.

(a) *Envelope measures.* (1) Long term rms/pause-corrected rms (an index of the amount of silence), (2) number of peaks (peak is 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, and (5) burst duration/total duration (a measure of the "roughness" of the envelope).

(b) *Autocorrelation statistics.* (1) Number of peaks, (2) maximum, (3) mean peak, and (4) standard deviation (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.

(c) *Correlogram-based pitch measures (from* Slaney, 1995*).* (1) Mean pitch, (2) median pitch, (3) SD of pitch,(4) maximum pitch, (5) mean pitch salience, and (6) maximum 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.

(d) *Moments of the spectrum.* (1) Mean (centroid), (2) SD, (3) skew, and (4) kurtosis.

(e) rms energy in octave-wide frequency bands from 63 to 16 000 Hz.

(f) *Spectral shift in time measures.* (1) Centroid mean, (2) centroid SD,(3) mean centroid velocity,(4) SD centroid velocity, and (5) maximum 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.

(g) *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.

(h) *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 (1) the height and (2) frequency of the maximum point in the modulation spectrum, as well as the (3) number, (4) mean, and (5) variance of bursts in modulation spectrum (using the burst algorithm described above).

(i) *Spectral flux statistics.* Spectral flux is another measure of the change in the spectrum over time. It is the running correlation of spectra in short (50-ms) time windows. The (1) mean, (2) SD, and (3) maximum value of the spectral flux were used in this analysis.

IV. RESULTS

Background condition, contextual congruence, and SNR were entered into a full-factorial analysis of variance (ANOVA) model using SPSS V16. In order to make model interpretation more transparent, all factors were coded as binary variables by collapsing target/background SNR into lower (−6 dB, −3 dB) and higher (−0 dB, +3 dB) SNR conditions. Preliminary analyses on sublevels of SNR demonstrated no significant or marginal effects.

Accuracy was used as the dependent measure, as opposed to signal-detection statistics such as A' or D', because the probability of misses and false alarms was not stationary over multi-target trials. In order to ensure that participants were not simply responding haphazardly, the responses to the incongruent distractor targets were first analyzed. Participants incorrectly responded to 6.5% of the distractors, but the response rate to targets was considerably higher (87.3% and 86.2%).

To simplify the reporting of results, all significant main effects and interactions are listed in Table I, and are described below and displayed in Figs. 2 and 3.

SNR. As expected, participants were more accurate in identifying targets when they were presented at higher SNR (0 and +3 dB: 88.7% correct, s.e. 1.6%) relative to lower SNR [−3 and −6 dB: 78.8% correct, standard error (s.e.) 1.6%].

TABLE I. Significant effects from ANOVA analysis with accuracy.

Background type (single versus dual)	$F(1,47)=29.556$	$p=<0.001$	$\eta^2=0.386$
Contextual congruence	$F(1,47)=34.059$	$p=<0.001$	$\eta^2=0.42$
Target/background SNR	$F(1,47)=77.116$	$p=<0.001$	$\eta^2=0.621$
Background type × congruence	$F(1,47)=7.233$	$p=<0.01$	$\eta^2=0.133$
Background type × SNR	$F(1,47)=12.728$	$p=p<0.001$	$\eta^2=0.213$

Background type (dual versus single background). Targets heard in the single background condition were more accurately identified (86.8% correct, s.e. 1.3%) than in the dual background condition (80.8% correct, s.e. 1.8%).

Target/background contextual congruence. Participants were also more accurate in identifying targets that were contextually incongruent with the background scene (86.5%, s.e. 1.5%) than congruent targets (81%, s.e. 1.5%). There was no significant interaction between contextual congruence and SNR.

SNR × background type (Fig. 2). When targets were presented at an intensity equal to or greater than that of the ongoing background scene, participants' performance differed little over single and dual backgrounds. However, when target/background SNR was lower, participants were significantly more accurate at identifying targets in the single rather than the dual background condition.

Target/background contextual congruency × background type (Fig. 3). In a single (stereophonic) background, participants were significantly more accurate at identifying targets when they were contextually incongruent with the background. However, this benefit of contextual incongruency (the pop-out effect) became statistically insignificant when participants had to attend to two different background scenes.

A. Acoustic analyses of target and background sounds

Mean performance for each ipsilateral background sound segment and for each target sound (averaged across participants) was first compared with the acoustic measurements using univariate regression. This allowed an assessment of the general acoustic aspects of the masking background sound or the target sound that were implicated in identification accuracy. For the target sounds, there were significant positive relationships between identification and the following acoustic measures: number of peaks and the range of the autocorrelation peaks (a measure of the spread of periodic information across the frequency spectrum). There were also negative relationships with the skewness of the spectrum and the frequency of the maximum burst in the modulation spectrum.

For the background sounds, analyses revealed significant ($p<0.05$) positive relationships between (target) identification accuracy with spectral skew, spectral kurtosis and mean spectral flux, and the mean and SD of the peaks of the autocorrelation (measures of the amount of periodicity and its spread across frequency ranges). There were negative relationships with measures of spectral centroid (i.e., both windowed spectral mean centroid and the first moment of the spectrum).

Finally, to account for overlapping variance between variables, a mixed stepwise multiple regression (with a liberal $p<0.1$ threshold) was calculated starting with all ipsilateral background and target acoustic measures, as well as the binary contextual congruence and dual/single background conditions. The resulting model (all $p<0.05$) found positive relationships between target identification accuracy and the maximum spectral velocity of the background, the maximum burst of the modulation spectrum, the number of peaks in the target, and the range of peaks of the autocorrelation of the target sounds. Negative relationships were

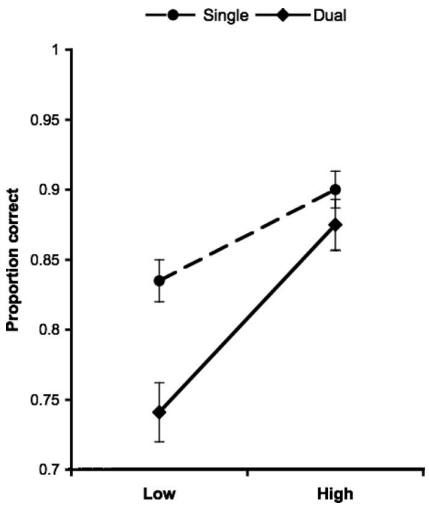

FIG. 2. SNR × background type.

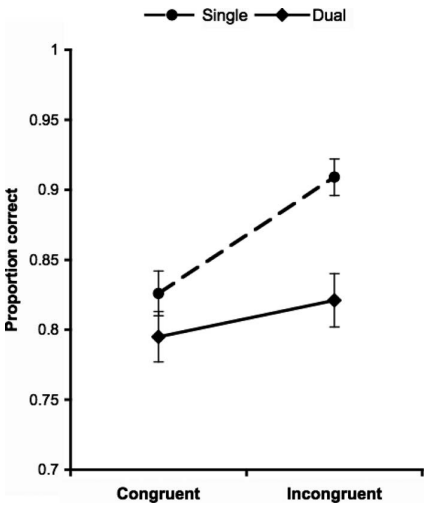

FIG. 3. Target/background contextual congruency × background type.

found between background centroid, SD of the spectral velocity of the background, and the frequency of the maximum burst in the modulation spectrum of the targets. In addition, experimenter-defined contextual congruence and dual/single background conditions were significant predictors of identification accuracy alongside the acoustic variables in the final model.

B. Acoustic similarity and contextual congruence

It is possible that the subjective assessment of contextual congruency may serve as a proxy measure for underlying acoustical similarity between target sound and background. Thus, to determine whether the experimenter-defined binary contextual congruency measures between target and background might be related to acoustic similarity or other factors, logistic and stepwise regression models were constructed using the battery of 38 acoustic similarity measures as predictor variables and contextual congruence as the binary dependent variable for each of the 92 different target/background pairings. The logistic regression model was nonsignificant [chi-square $(36)=25.00$, $p=0.916$]. In the stepwise regression model, no acoustic similarity variable was entered at a liberal $p<0.1$ level. It was also possible that latent acoustic variables underlying the 38 acoustic measures might relate to the contextual congruence measure. To test this, all 38 principal components of the acoustic variables were entered into a stepwise model with contextual congruence as the dependent variable. Again, the stepwise model entered no predictor variable into the model ($p>0.1$). These results suggested that subjective contextual congruence between target and background was not transparently related to acoustical factors.

Turning to the experimental results, a mixed stepwise regression assessed whether the accuracy difference between contextually congruent and incongruent targets could be accounted for by metrics of acoustic similarity alone, or whether they could be accounted for by a combination of acoustic similarity and the binary contextual variable. Here, the experimenter-defined congruency factor (congruent or incongruent), 38 acoustic similarity measures, and single/dual background condition were passed to the model as regressor terms ($p<0.1$). The stepwise procedure converged on a model that included the contextual congruence term ($p<0.05$) and the dual versus single background condition term ($p<0.05$). In addition, the final model included three acoustic measures significant at the $p<0.05$ level: the SD and maximum value of the pitch, and the range of peaks of the autocorrelation statistics. Two further acoustic measures were marginally significant predictors: the maximum value of the spectral velocity, and the mean spectral flux. These results suggest that the contextual congruence and acoustic similarity of target and background account for different variance in participants' accuracy scores.

V. SUMMARY AND DISCUSSION

In summary, the experimental results were as follows: (1) Overall, reducing the SNR of natural sound targets decreased participants' accuracy in detecting and identifying those targets; (2) participants were less accurate in detecting and identifying sounds when they had to attend to two different auditory scenes (presented to two ears) as opposed to a single stereophonic scene; (3) this greater accuracy in one versus two scenes emerged only at low SNRs; (4) participants were more accurate in detecting and identifying targets that were contextually incongruent with the background; (5) acoustic similarity and congruency with the real-world auditory scene accounted for different aspects of the variance in accuracy; and (6) accuracy was only affected by target/background congruence for single stereophonic scenes.

These findings support the general predictions outlined in the Introduction. As expected, based on previous findings, increased contralateral masking in the multiple background condition disrupted target detection and identification overall and reduced detection accuracy at low SNRs. In addition, a contextual pop-out effect was observed for incongruent targets in single backgrounds in the detection accuracy measure. This effect was eliminated when the target stimuli were presented in multiple backgrounds.

The present study explored a number of possible explanations for the contextual pop-out effect. One account is that participants are better able to detect and identify contextually incongruent targets simply because they are more acoustically dissimilar to the background scene, relative to contextually congruent targets. To evaluate this claim, detailed acoustic analyses of the target and background sounds were conducted, to determine which acoustic properties of the stimuli predicted performance in the task, and to establish whether the pattern of findings could be accounted for in terms of the acoustic overlap between backgrounds and targets. The results of the acoustic analyses indicate that accuracy depended on a mixture of spectral and temporal qualities of both the target and background sounds. For the background sounds, increased spectral (mean spectral flux and maximum spectral velocity) and temporal fluctuations (maximum modulation spectrum) led to improved performance. Similarly for target sounds, accuracy increased with envelope measures of fluctuation (the number of amplitude peaks) These findings suggest that temporal peaks in the target sound envelope, as well as spectral and temporal dips in the masking background sounds, served to aid target performance. This is consistent with findings from speech recognition (e.g., Moore *et al.*, 1999) in which listeners have lower speech reception thresholds for masking sounds with spectral and temporal dips. In the current study, detection and identification were also more accurate for periodic background sounds and for backgrounds and target sounds with periodic information spread over a broader range of frequencies, which is roughly consistent with the reduced masking for harmonic background sounds observed in previous speech recognition studies (Treurniet and Boucher, 2001), Gygi *et al.* (2004) also found that greater periodicity in environmental sounds predicted higher accuracy in identifying these sounds when they were vocoded. Greater periodicity was also associated with faster response times to identifying environmental sounds (Ballas, 1993). Finally, general spectral properties (the first and third moments of the spectrum) of the background and target sounds were predictive of success-

ful target detection and identification. Background sounds with higher spectral centroids were harder to detect and identify, while highly skewed background sounds were easier, perhaps because the spectral energy associated with these sounds is not evenly distributed across high and low frequencies, providing less consistent masking across the frequency range. Mirroring this finding, target sounds with less skewed spectra were easier to detect and identify. (Spectral skewness also modulated reaction times for environmental sound identification in Ballas, 1993.)

Measures of target/background similarity revealed that the contextual pop-out effect is only partially explained by acoustic factors. Acoustic variables that contributed significantly to the pop-out effect included correlogram-based pitch measures and peak autocorrelation statistics reflecting the degree of overlap in the spectro-temporal characteristics of the stimuli and the distribution of periodic elements across different frequencies. However, acoustic similarity was not the sole predictor of the congruence of a target stimulus with a background scene, or the emergence of the behavioral pop-out effect. This finding suggests that higher-level cognitive factors may play an important role in the detection and identification of meaningful sounds in complex auditory scenes.

One possibility explored in the present study is that the meaningful sound environments suggested by real-world auditory scenes activate associated memory representations, resulting in the generation of expectancies, such that unexpected sounds are more salient and therefore better detected. A subjective measure of the semantic compatibility of the target sounds with the meaningful context represented by the background proved a significant predictor of detection performance, providing support for this interpretation. The conflict between the expectancies generated by the context and the presentation of an incongruent target might have been expected to have a detrimental effect on identification performance, as has been observed in other domains. In spoken language processing, for example, a biasing sentence context facilitates recognition of words that are compatible with the contextual meaning, and inhibits recognition of incongruent words (Stanovich and West, 1983; Aydelott and Bates, 2004). However, in the present study, incongruent target sounds were *more* accurately identified than congruent targets, suggesting that it is ease of detection, rather than compatibility with the context, that facilitates target identification in this paradigm.

The specific demands of the detection task may account for this finding, as monitoring a continuous auditory scene for a single event requires the listener to suppress responses to potential false alarms. Thus, congruent sounds are more likely to be interpreted as elements of the background scene, and therefore to escape detection. A paradigm in which the background scene served to predict a specific event (e.g., the sound of tires screeching followed by the sound of a car crash) might better show the effects of contextual buildup on target identification (cf. Ballas and Mullins, 1991). In addition, the present study only relies on a limited set of target and background sounds, and thus the generalizability of the results would be enhanced by using a more comprehensive set of sounds that sample the full range of naturally occurring auditory events.

As noted above, the detection of tone, noise, and speech targets is vulnerable to informational masking—and this study extends this finding to the detection and identification of environmental sounds in more naturalistic settings. As in previous studies (e.g., Brungart and Simpson, 2002), this informational masking effect is only apparent when targets are presented at low SNRs within the ipsilateral background. Potentially more noteworthy, however, is the finding that the introduction of different contralateral acoustic information disproportionately reduces identification accuracy for contextually incongruent targets. Indeed, the main effect of background scene is primarily driven by the drop in accuracy for incongruent targets in the presence of a dual background. This supports the claim that the congruency effect is being driven by more attentional or cognitive factors, in that additional information from the second background scene is completely spatially segregated from the target and its incongruent background. Thus, it is informational rather than energetic masking that disrupts contextual pop-out.

As mentioned above, the contralateral masking result is very similar to that reported in several cocktail-party-inspired studies of speech processing, such as the coordinate response measure paradigm (Kidd *et al.*, 2007; Brungart and Simpson, 2002). Brungart and Simpson (2002) and Brungart *et al.* (2005) observed that contralateral informational masking of speech is greater the more speech-like the masking signal, indicating that the masking effect occurs when the listener must segregate simultaneously presented auditory stimuli with similar perceptual characteristics. These results suggest a set of follow-ups to the present study: If indeed there are similar informational masking effects occurring in these non-linguistic scenes, then one should see that reversed or combined environmental sound scenes presented contralaterally to the target should induce significantly more informational masking than artificial noise distractors, such as white noise convolved with the amplitude envelope of an environmental sound scene.

Another possibility is that the contralateral masking effect in the present study is at least partially driven by the increased attentional demand imposed by the presentation of an additional background scene. Having to monitor two distinct, complex, and changing auditory streams may interfere with listeners' ability to generate expectancies (cf. Aydelott and Bates, 2004), thereby contributing to the elimination of the pop-out effect. A similar result has been reported in the phoneme monitoring literature. The phoneme monitoring task resembles the paradigm used in the present study, in that it requires listeners to respond whenever they hear a particular phoneme occurring within a speech stream. Of particular interest is the finding that listeners' detection accuracy is poorer when they are required to perform a secondary task (Martin, 1977; Treisman and Squire, 1974), suggesting that the generation of expectancies based on the information provided by the auditory context is disrupted by increased attentional load, even in the absence of a competing acoustic signal. By manipulating the predictability of the target in the

present paradigm (e.g., by holding constant the temporal or spatial location of the target and the number of targets), it may be possible to disambiguate whether the contralateral masking is driven by the difficulty of segregating and monitoring distinct auditory streams in the face of conflicting information, or other processes of distraction such as stimulus uncertainty or overwhelming attentional load.

In conclusion, the present study demonstrates that real-world contextual information from the auditory scene influences the accurate detection and identification of natural environmental sounds. The acoustic properties of the target and background sounds offer only a partial account of the context effect; the extent to which the target sound is predictable within the meaningful auditory environment represented by the background also plays a significant role. The emergence of context-driven effects on target identification depends on the perception of a single, unified auditory environment, and is therefore highly vulnerable to contralateral masking by a competing background scene. The respective contributions of expectancy, auditory segregation, and attentional demand to the observed pattern of results remain topics for future research.

Aydelott, J., and Bates, E. (**2004**). "Effects of acoustic distortion and semantic context on lexical access," Lang. Cognit. Processes **19**, 29–56.

Ballas, J. A. (**1993**). "Common factors in the identification of an assortment of brief everyday sounds," J. Exp. Psychol. Hum. Percept. Perform. **19**, 250–267.

Ballas, J. A., and Mullins, T. (**1991**). "Effects of context on the identification of everyday sounds," Hum. Perform. **4**, 199–219.

Boersma, P., and Weenink, D. (**2009**). "Praat: Doing phonetics by computer," Version.5.1.18 (Computer program), http://www.praat.org (Last accessed Oct. 15, 2009).

Brainard, D. H. (**1997**). "The psychophysics toolbox," Spatial Vis. **10**, 433–436.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT, Cambridge, MA).

Broadbent, D. E. (**1958**). *Perception and Communication* (Pergamon, Oxford).

Brokx, J. P. L., and Nootebohm, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.

Brungart, D. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Brungart, D., and Simpson, B. (**2002**). "Within-ear and across-ear interference in a cocktail-party listening task," J. Acoust. Soc. Am. **112**, 2985–2995.

Brungart, D., Simpson, B., Darwin, C., Arbogast, T. L., and Kidd, G. (**2005**). "Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task," J. Acoust. Soc. Am. **117**, 292–304.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.

Cummings, A., Ceponiene, R., Koyama, A., Saygin, A. P., Townsend, J., and Dick, F. (**2006**). "Auditory semantic networks for words and natural sounds," Brain Res. **1115**, 92–107.

Cusack, R., and Carlyon, R. P. (**2003**). "Perceptual asymmetries in audition," J. Exp. Psychol. Hum. Percept. Perform. **29**, 713–725.

Darwin, C., and Hukin, R. (**2000**). "Effectiveness of spatial cues, prosody, and talker characteristics in selective attention," J. Acoust. Soc. Am. **107**, 970–977.

Drennan, W. R., Gatehouse, S., and Lever, C. (**2003**). "Perceptual segregation of competing speech sounds: The role of spatial location," J. Acoust. Soc. Am. **114**, 2178–2189.

Durlach, N., Mason, C., Shinn-Cunningham, B., Arbogast, T., Colburn, H., and Kidd, G. (**2003**). "Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity," J. Acoust. Soc. Am. **114**, 368–379.

EPAPS Document No. E-JASMAN-126-044911 for a list of the background and target sounds heard by participants and the different conditions these sounds were heard in (e.g., signal to noise ratios). For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.

Gygi, B., Kidd, G. R., and Watson, C. S. (**2004**). "Spectral-temporal factors in the identification of environmental sounds," J. Acoust. Soc. Am. **115**, 1252–65.

Gygi, B., Kidd, G. R., and Watson, C. S. (**2007**). "Similarity and categorization of environmental sounds," Percept. Psychophys. **69**, 839–55.

Gygi, B., and Shafiro, V. (**2007**). "Effect of auditory context on the identification of environmental sounds," in Proceedings of the 19th International Congress of Acoustics, Madrid, Spain.

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (**2004**). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer," J. Acoust. Soc. Am. **115**, 833–843.

Houtgast, T., and Steeneken, H. J. M. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, 1069–1077.

Kidd, G., Jr., Mason, C. R., and Richards, V. M. (**2003**). "Multiple bursts, multiple looks, and stream coherence in the release from informational masking," J. Acoust. Soc. Am. **114**, 2835–2845.

Kidd, G. R., Watson, C. S., and Gygi, B. (**2007**). "Individual differences in auditory abilities," J. Acoust. Soc. Am. **122**, 418–435.

Leech, R., Aydelott, J., Symons, G., Carnevale, J., and Dick, F. (**2007**). "The effect of semantic and attentional distractors on syntactic processing in typical development and adulthood," Dev. Sci. **10**, 794–813.

Martin, M. (**1977**). "Reading while listening: A linear model of selective attention," J. Verbal Learn. Verbal Behav. **16**, 453–463.

Moll, K., Cardillo, E., and Aydelott Utman, J. (**2001**). "Effects of competing speech on sentence-word priming: Semantic, perceptual, and attentional factors," in *Cognitive Science*, edited by J. D. Moore and K. Stenning (Lawrence Erlbaum Associates, Edinburgh), pp. 651–656.

Moore, B. C. J., Peters, R. W., and Stone, M. A. (**1999**). "Benefits of linear amplification and multichannel compression for speech comprehension in backgrounds with spectral and temporal dips," J. Acoust. Soc. Am. **105**, 400–411.

Oh, E. L., and Lutfi, R. A. (**1999**). "Informational masking by everyday sounds," J. Acoust. Soc. Am. **106**, 3521–3528.

Saygin, A. P., Dick, F., and Bates, E. (**2005**). "An on-line task for contrasting auditory processing in the verbal and nonverbal domains and norms for younger and older adults," Behavior Research Methods **37**, 99–110.

Slaney, M. (**1994**). "Auditory toolbox: A Matlab toolbox for auditory modeling work," Apple Computer Technical Report No. 45, Apple Computer Inc., Cupertino, CA.

Stanovich, K. E., and West, R. F. (**1983**). "On priming by a sentence context," J. Exp. Psychol. Gen. **112**, 1–36.

Stemberger, J. P., Elman, J. L., and Haden, P. (**1985**). "Interference between phonemes during phoneme monitoring: Evidence for an interactive activation model of speech perception," J. Exp. Psychol. Hum. Percept. Perform. **11**, 475–489.

Treisman, A., and Squire, R. (**1974**) "Listening to speech at two levels at once," Q. J. Exp. Psychol. **26**, 82–97.

Treurniet, W. C., and Boucher, D. R. (**2001**). "A masking level difference due to harmonicity," J. Acoust. Soc. Am. **109**, 306–320

Wightman, F. L., Callahan, M. R., Lutfi, R. A., Kistler, D. J., and Oh, E. (**2003**). "Children's detection of pure-tone signals: Informational masking with contralateral maskers," J. Acoust. Soc. Am. **113**, 3297–3305.

Wightman, F. L., and Kistler, D. J. (**2005**). "Informational masking of speech in children: Effects of ipsilateral and contralateral distracters," J. Acoust. Soc. Am. **118**, 3164–3176.