Discovering acoustic structure of novel sounds

Christian E. Stilp, Michael Kiefte, and Keith R. Kluender

Citation: The Journal of the Acoustical Society of America **143**, 2460 (2018); doi: 10.1121/1.5031018
View online: https://doi.org/10.1121/1.5031018
View Table of Contents: http://asa.scitation.org/toc/jas/143/4
Published by the Acoustical Society of America

Discovering acoustic structure of novel sounds[a)]

Christian E. Stilp,[1,b)] Michael Kiefte,[2] and Keith R. Kluender[3]
[1]*Department of Psychological and Brain Sciences, University of Louisville, 317 Life Sciences Building, Louisville, Kentucky 40292, USA*
[2]*School of Communication Sciences and Disorders, Dalhousie University, Halifax, Nova Scotia, Canada*
[3]*Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, Indiana 47907, USA*

(Received 13 July 2017; revised 24 March 2018; accepted 26 March 2018; published online 27 April 2018)

Natural sounds have substantial acoustic structure (predictability, nonrandomness) in their spectral and temporal compositions. Listeners are expected to exploit this structure to distinguish simultaneous sound sources; however, previous studies confounded acoustic structure and listening experience. Here, sensitivity to acoustic structure in novel sounds was measured in discrimination and identification tasks. Complementary signal-processing strategies independently varied relative acoustic entropy (the inverse of acoustic structure) across frequency or time. In one condition, instantaneous frequency of low-pass-filtered 300-ms random noise was rescaled to 5 kHz bandwidth and resynthesized. In another condition, the instantaneous frequency of a short gated 5-kHz noise was resampled up to 300 ms. In both cases, entropy relative to full bandwidth or full duration was a fraction of that in 300-ms noise sampled at 10 kHz. Discrimination of sounds improved with less relative entropy. Listeners identified a probe sound as a target sound (1%, 3.2%, or 10% relative entropy) that repeated amidst distractor sounds (1%, 10%, or 100% relative entropy) at 0 dB SNR. Performance depended on *differences* in relative entropy between targets and background. Lower-relative-entropy targets were better identified against higher-relative-entropy distractors than lower-relative-entropy distractors; higher-relative-entropy targets were better identified amidst lower-relative-entropy distractors. Results were consistent across signal-processing strategies. © *2018 Acoustical Society of America*.
https://doi.org/10.1121/1.5031018

[AKCL] Pages: 2460–2473

I. INTRODUCTION

In everyday listening, the waveform that reaches the ear is typically a combination of multiple overlapping sounds. A challenge for auditory perception is to distinguish sounds from different sources amidst competing sounds. There is an extended literature concerning auditory scene analysis, source segregation, and stream segregation (see Bregman, 1990, for a review) with substantial allied research concerning informational and energetic masking (e.g., Pollack, 1975; Lutfi, 1990; Brungart, 2001; Durlach *et al.*, 2003a, b; Kidd *et al.*, 2008). Broadly stated, listeners use myriad acoustic properties to separate and segregate simultaneous sounds including spectral composition, temporal characteristics, modulation properties, and perceived location; it is beyond the scope of this report to provide comprehensive review. Instead, the present focus is upon more global stimulus characteristics that promote segregation, including repetition (Andreou *et al.*, 2011; McDermott *et al.*, 2011; Masutomi *et al.*, 2016), familiarity (Newman and Evers, 2007; Johnsrude *et al.*, 2013), similarity (Kidd *et al.*, 2002; Durlach *et al.*, 2003b), and predictability (Bendixen, 2014).

Here, innovative signal processing strategies are employed to assess how listeners use predictability (inverse of entropy) to distinguish, discover, and identify new sounds with quantifiably different degrees of acoustic structure. Natural sounds are far from random, exhibiting substantial structure (predictability, redundancy, or simply nonrandomness) in their spectral and temporal compositions. Through evolution and experience, sensory systems become attuned to structure in natural stimuli (Attneave, 1954; Barlow, 1961). In turn, sensory processing becomes more efficient (Bell and Sejnowski, 1997; Lewicki, 2002; Smith and Lewicki, 2006). Stilp and colleagues (Stilp *et al.*, 2010; Stilp and Kluender, 2011, 2012, 2016; Kluender *et al.*, 2013) have directly demonstrated efficient coding of stimulus structure for auditory perception.

True random (white) noise of infinite bandwidth and duration constitutes maximum entropy and the minimum of acoustic structure. When noise is restricted to a finite bandwidth or duration (particularly narrow bandwidth or brief duration), it deviates from randomness and contains detectable idiosyncrasies in frequency and/or in time when repeated. When time-limited segments of white noise are concatenated, listeners can tap along with the "frozen" noise segment, suggesting that very little structure needs to be present to recognize and distinguish sounds (Guttman and Julesz, 1963; Kaernbach, 1992, 1993; see also Agus *et al.*, 2010; Agus and Pressnitzer, 2013; Andrillon *et al.*, 2015). However, investigators have been unable to identify consistent short-term spectral features used to recognize a given sample of frozen noise (Kaernbach, 1993; Agus *et al.*, 2010). Listeners in these studies likely used some acoustic properties to detect repetitions, even if subtle and variable across listeners (Agus and Pressnitzer, 2013; Andrillon *et al.*, 2015). It has been observed that listeners are adept at

[a)]Portions of this work were presented at the 159th Meeting of the Acoustical Society of America, Baltimore, MD, USA, April 2010.
[b)]Electronic mail: christian.stilp@louisville.edu

 0001-4966/2018/143(4)/2460/14/$30.00

exploiting spectrotemporal idiosyncrasies when perceiving short-duration noise. For example, discrimination of bandpass-filtered (100–3300 Hz) frozen noise peaks at 40-ms duration before degrading at longer durations (Goossens *et al.*, 2008).

Observers can use stimulus structure to learn to identify novel visual objects within cluttered scenes. Brady and Kersten (2003) generated "digital embryos," novel three-dimensional (3-D) shapes that simulated some aspects of embryological development. These "embryos" possessed 3-D spatial structure dictated by constraints imposed by an algorithm designed to mimic developmental progression. The target embryo was presented repeatedly against successive backgrounds of embryos generated in a similar fashion. From trial to trial, background embryos, patterns of illumination, and target embryo position changed, but orientation and illumination of the target embryo remained constant. Observers learned to combine information from multiple viewings of different cluttered scenes to recognize the target embryo. Brady and Kersten termed this "bootstrapped learning," through which observers learned to recognize initially ambiguous target stimuli through structure and repetition.

Returning to audition, studies concerning perception of simultaneous sounds have generally used one of four types of sounds: clicks, pure tones, speech, and white noise (Fig. 1). White noise represents a lower limit of acoustic structure because it possesses little discernible redundancy or predictability in frequency or in time. Clicks and tones represent upper limits of acoustic structure due to their extreme sparseness in time and frequency, respectively. Speech sounds possess intermediate degrees of acoustic structure because they possess more structure than noise but are less spectrotemporally sparse than clicks or tones. From this perspective, many investigations of simultaneous sound perception measured detection of a high-structure target sound (e.g., tones, speech) amidst low-structure background sounds (e.g., noise). Other studies investigated listeners' ability to detect higher-structure targets from among similarly higher-structure backgrounds [e.g., detecting tones amidst other tones (Wegel and Lane, 1924), multitone complexes (Neff and Green, 1987) or multitone sequences (Watson *et al.*, 1975; 1976); perceiving speech amidst competing talker(s) (Cherry, 1953; Summerfield and Assmann, 1989; Simpson and Cooke, 2005)]. Rarely has a lower-structure target sound been presented amidst a higher-structure background [e.g., perceiving noise bands amidst pure tones in what Greenwood (1961) termed a "reverse experiment" for deriving masked audiograms].

Natural sounds have structure due to physical constraints upon the sources that produce them. So, it may seem attractive to posit that listeners simply exploit this structure to segregate sound sources. However, there are two shortcomings to this suggestion. First, it is necessary to quantify acoustic structure. While one can construct an ordinal array of sounds that vary in acoustic structure as in Fig. 1, this falls short of providing a metric that is required to make quantitative comparisons and test hypotheses. Second, it is necessary to gain control over listening experience. For example, listeners have incomparable experience perceiving speech, and are highly practiced at speech perception in adverse listening conditions. By contrast, listeners have little experience outside the laboratory with clicks and tones employed by researchers.

The present experiments examined listeners' discovery of acoustic structure in novel complex sounds. Complementary signal-processing strategies independently varied relative acoustic entropy (the inverse of acoustic structure) across frequency or time. A restricted low-frequency band of random noise was either stretched up to 5 kHz, or a short gated interval of random noise was resampled up to 300-ms duration. In both cases, entropy relative to full bandwidth or full duration was a fraction of that in a 300-ms noise sampled at 10 kHz.

Two main hypotheses were evaluated. First, it was hypothesized that listeners would be sensitive to acoustic structure in novel complex sounds. Discrimination of sounds with more acoustic structure (less relative entropy) was predicted to be superior to discrimination of sounds with less acoustic structure (more relative entropy). This prediction was tested by measuring discrimination of novel sounds in quiet.

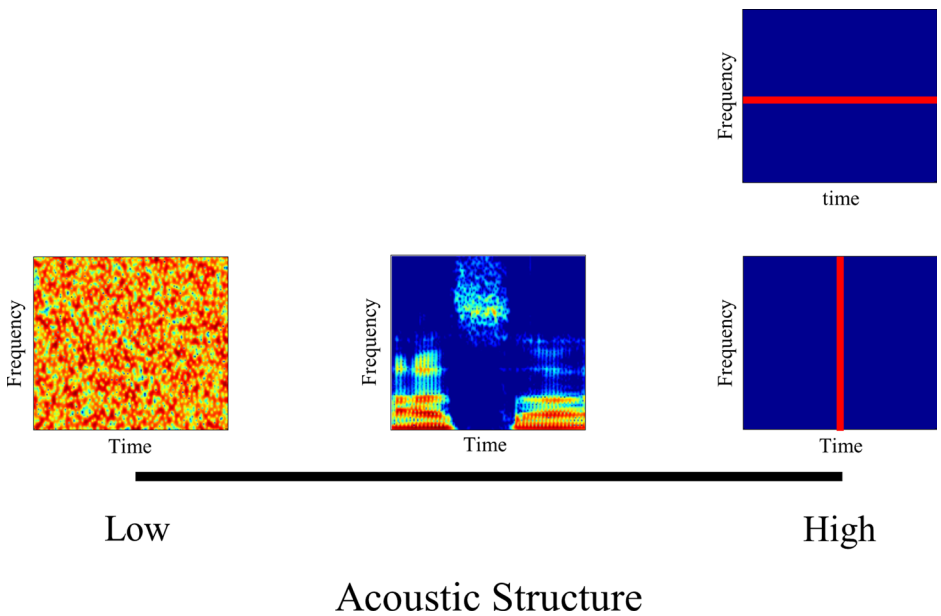

FIG. 1. (Color online) Spectrograms of stimuli commonly used in auditory masking and streaming experiments arranged along a conceptual dimension of acoustic structure. Random noise can be considered a relative minimum of acoustic structure (high unpredictability), while pure tones and clicks can be considered a relative maximum of acoustic structure (low unpredictability). Speech sounds possess acoustic structure that is intermediate to these extremes.

Second, it was hypothesized that listeners would use acoustic structure to extract novel complex sounds from background sounds. Acoustic structure was manipulated in target and background sounds separately. Performance was predicted to vary as a function of the similarity in relative acoustic entropy between target and background sounds. Similar amounts of acoustic structure in target and background sounds were predicted to *impede* target recognition; whereas disparate degrees of acoustic structure in target and background were predicted to *facilitate* target recognition. This pattern of results would be consistent with investigations where target/masker similarity impeded target sound detection when much simpler sounds were employed (Barker and Cooke, 1999; Brungart, 2001; Kidd *et al.*, 2002; Durlach *et al.*, 2003b; Leibold and Neff, 2007). This prediction was tested using the embedded repetition paradigm of McDermott *et al.* (2011) through which listeners reported whether a probe sound matched the target sound that repeated amidst a series of background sounds.

II. EXPERIMENT 1

A. Method

1. Listeners

Seventy-nine undergraduate students participated in exchange for course credit (20 in the control study; 28 in experiment 1a; 31 in experiment 1b). All reported no known hearing impairment, and none participated in more than one experiment.

2. Stimuli

Analogous to Brady and Kersten (2003), random stimuli were generated with spectrotemporal structure but which are otherwise completely unfamiliar to listeners. Stimuli were generated through one of two signal-processing strategies: instantaneous frequency dilation (FD; experiment 1a) or time dilation (TD; experiment 1b). Each process began with generation of a 300-ms segment of white noise sampled at 10 kHz before total entropy was parametrically reduced relative to this original noise stimulus. This produced novel sounds with varying degrees of acoustic structure. All of the signal processing to produce these "acoustic embryos" was performed in MATLAB.[1]

a. Instantaneous frequency dilation (FD). In FD, the noise sample was low-pass filtered with a cutoff frequency determined by the desired relative entropy. In order to produce the low-pass-filtered noise, a segment of noise with duration $300q$ ms (where q is the desired final entropy, ranging from 0.01 to 1) was resampled so that the final duration was 300 ms. A least-squares linear-phase low-pass finite impulse response (FIR) filter as implemented in the resample routine in MATLAB with filter order $20/q$ was utilized. For example, to generate a stimulus with 10% entropy relative to the original 5000-Hz-bandwidth 300-ms stimulus, a 30-ms segment of white noise was generated and resampled by a factor of 10:1 (making the filter order 200), resulting in a low-pass-filtered stimulus with cutoff frequency at 500 Hz. The FD procedure effectively resampled the frequency axis of the spectrum of this noise sample by scaling the instantaneous frequency[2] of the Hilbert transform to $\pm\pi$. The instantaneous frequency (ω_t) is defined as the first difference of the instantaneous phase

$$\omega_t = \frac{d}{dt}\varphi(t). \quad (1)$$

The Hilbert transform was determined from the low-pass-filtered segment, and the first difference of unwrapped phase was calculated to determine instantaneous frequency.[3] The instantaneous frequency was then scaled by the inverse of the target entropy (e.g., for 10% entropy, the instantaneous frequency was scaled by a factor of 10) and used to modulate a constant-amplitude pure tone. This effectively expanded the bandwidth of the signal to 5 kHz. This produced a novel sound with a roughly flat long-term distribution of energy from 0 to 5 kHz, while maintaining the same relative entropy as the low-pass-filtered signal. As seen in Fig. 2, at low levels of relative entropy, the stimulus consists of random impulses and slowly varying sweeps that span the full bandwidth with the overall spectrotemporal density related to relative entropy. FD was used to produce stimuli at nine relative entropy levels that were equally spaced logarithmically (10^0 to 10^2 in $10^{0.25}$-steps; see Fig. 2 and Mm. 1–Mm. 3 for examples). These relative entropy levels are listed with the corresponding low-pass-filter cutoff frequencies in Table I.

Mm. 1. Audio of 5 frequency-dilated white-noise embryos at 1% relative entropy. This is a file of type "wav" (69 KB).

Mm. 2. Audio of 5 frequency-dilated white-noise embryos at 3.2% relative entropy. This is a file of type "wav" (69 KB).

Mm. 3. Audio of 5 frequency-dilated white-noise embryos at 10% relative entropy. This is a file of type "wav" (69 KB).

b. Time dilation (TD). In TD, the time axis was effectively resampled. Instead of low-pass filtering the 300-ms segment of noise, noise was sampled for only the duration determined by the target relative entropy level. For example, to achieve 10% of the entropy of the original noise sample, a 30-ms interval of noise was used (i.e., 10% of 300 ms). The instantaneous frequency was then upsampled by the inverse of the target entropy (e.g., for 10% entropy, the instantaneous frequency was scaled by a factor of 10). The noise sample was then resynthesized with no amplitude modulation of the complex signal, similar to the FD stimuli. Durations of noise to be dilated followed the same percentages as used in FD (10^0 to 10^2 in $10^{0.25}$-steps; see Table I). Spectrotemporal variability increased as a function of relative entropy (see Fig. 2 and Mm. 4–Mm. 6 for examples). As shown in Fig. 2, this procedure resulted in modulated tones that vary around a center frequency of approximately 1600 Hz with the range of deviations related to relative entropy.

Mm. 4. Audio of 5 time-dilated white-noise embryos at 1% relative entropy. This is a file of type "wav" (69 KB).

Mm. 5. Audio of 5 time-dilated white-noise embryos at 3.2% relative entropy. This is a file of type "wav" (69 KB).

Mm. 6. Audio of 5 time-dilated white-noise embryos at 10% relative entropy. This is a file of type "wav" (69 KB).

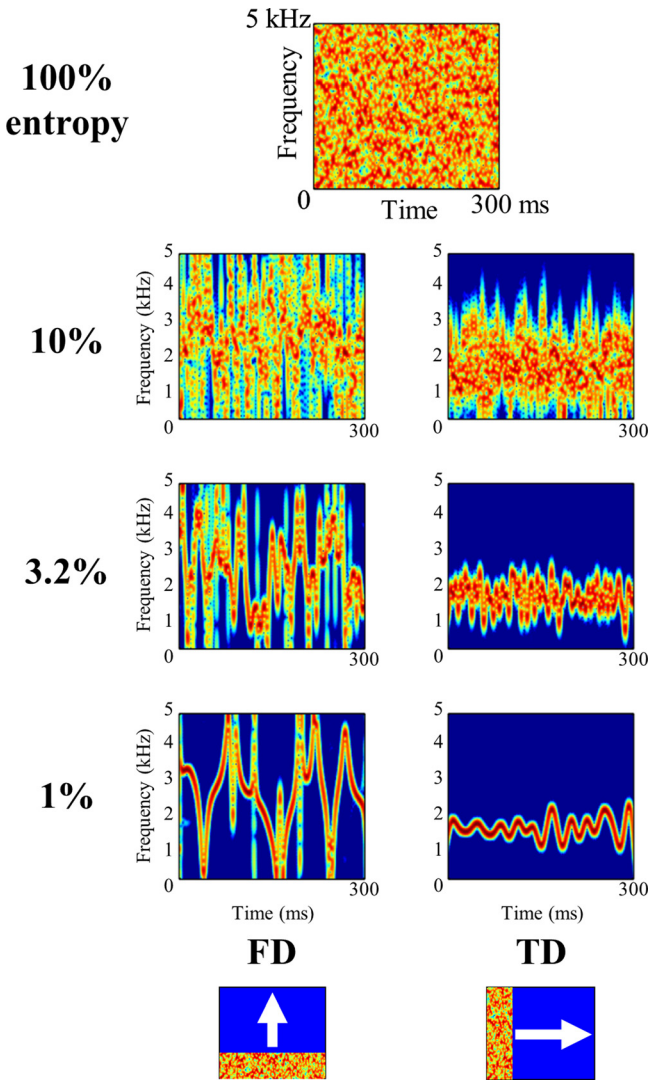

FIG. 2. (Color online) Stimulus generation procedures for experiment 1. The top row depicts the spectrogram of an undilated 300-ms, 5-kHz bandwidth noise sample. The left column depicts examples of instantaneous frequency dilation (FD) at the labeled entropy percentages (relative to the undilated noise at top). In FD, the frequency axis is truncated via low-pass filtering and then stretched to the full 5-kHz bandwidth. The right column depicts examples of time dilation (TD), where stimulus duration is truncated and then stretched to the full 300-ms duration.

There are multiple ways of producing reduced-entropy stimuli. For example, the frequency modulation of a tone could be manipulated directly or hand-edited spectrograms could be resynthesized. However, by taking a noise sample as a starting point, we have circumvented some questions as to how the final stimulus frequency should be modulated. By expanding the distribution of instantaneous frequencies in a low-pass-filtered noise, stimulus relative entropy has been manipulated while still controlling for the overall bandwidth and duration. In addition, the randomly generated stimuli were at least somewhat reminiscent of naturally occurring sounds in their spectrotemporal fluctuations while still remaining completely unfamiliar.

In both conditions, entropy is defined relative to the bandwidth/duration of the original stimulus. Auditory processing will alter the amount of information through masking and other types of nonlinear processing that occurs in the hearing mechanism. There is no standardized psychometric scale for relative entropy, so we have chosen an arbitrary, but constant, reference instead.

TABLE I. Stimulus parameters. The first column depicts the equal logarithmic spacing of entropy levels, calculated as 10 raised to the listed exponent. The second column lists percentages of relative entropy compared to that in a 5000-Hz bandwidth, 300-ms, undilated sample of random noise. The third column lists low-pass-filter cutoff frequencies used in instantaneous frequency dilation. The fourth column lists durations of the original noise sample used in time dilation.

Exponent (10^x)	Percent entropy	Low-pass cutoff (Hz)	Duration (ms)
0	1.00%	50.00	3.00
0.25	1.78%	88.91	5.33
0.50	3.16%	158.11	9.49
0.75	5.62%	281.17	16.87
1.00	10.00%	500.00	30.00
1.25	17.78%	889.14	53.35
1.50	31.62%	1581.14	94.87
1.75	56.23%	2811.71	168.70
2.00	100.00%	5000.00	300.00

3. Procedure

a. Control study. A control study was conducted to assess baseline discriminability of novel sounds absent competing background sounds. Forty 300-ms FD stimuli were generated at each of the nine levels of relative entropy listed in Table I (360 FD stimuli total). At each level of relative entropy, stimuli were arranged in 20 two-alternative forced-choice (AXB) discrimination trials with 250-ms ISIs. No sounds were repeated across trials. The same procedure was followed for 300-ms TD stimuli (360 stimuli total).

FD and TD trials were divided into separate blocks of 180 trials. Listeners were instructed to report whether the first or third sound in the triad differed from the second (standard) sound; response boxes were labeled accordingly. Feedback was provided on each trial by illuminating a light over the correct response button after the response had been recorded. Block order was counterbalanced across listeners, and each block lasted approximately 15 min. In all experiments, trial sequences were upsampled to 48 828 Hz, D/A converted (TDT RP2, Tucker-Davis Technologies, Alachua, FL), and presented diotically at 72 dB sound pressure level (SPL) (A) via circumaural headphones (Beyer-Dynamic DT-150, Beyerdynamic Inc., Farmingdale, NY) in sound-isolating booths (Acoustic Systems, Inc., Austin, TX).

b. Experiments 1a and 1b. The main experiment tested three levels of relative entropy in target sounds (1%, 3.2%, and 10%) and three levels of relative entropy in distractor sounds (1%, 10%, and 100%). Levels of relative entropy were fully crossed, producing nine experimental conditions. For each condition, 30 target stimuli and 1080 distractor stimuli were generated: 9990 stimuli total. Thus, 9990 FD stimuli (experiment 1a) and 9990 TD stimuli (experiment 1b) were generated, with no sound being heard on more than one trial.

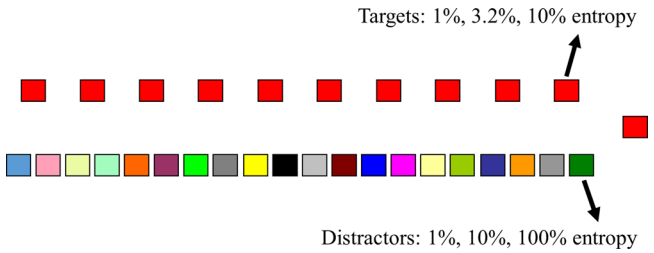

FIG. 3. (Color online) Sample trial sequence in experiment 1. Squares in the top row represent the repeating target sound, which had either 1%, 3.2%, or 10% of the entropy of random noise. Squares in the bottom row represent 20 different distractor sounds, all of which had either 1%, 10%, or 100% of the entropy of random noise. The final square was the probe sound, to which listeners reported whether or not this matched the repeating target sound. Here, target and probe sounds matched.

Target and distractor sounds were combined to form trial sequences following McDermott *et al.* (2011; Fig. 3). In each trial, one target sound was repeated 10 times with 400-ms ISIs, and 20 different distractor sounds were concatenated with 50-ms ISIs. Target and distractor sequences were combined at 0 dB SNR and the midpoint of each target sound was centered over the ISI between two distractor sounds.[4] Following the final distractor sound, a 150-ms silent interval preceded the 300-ms probe sound. The probe sound either matched the repeating target sound ("same" trial) or was a new sound at the same level of relative acoustic entropy ("different" trial). In all, total trial duration was 7400 ms.

Listeners were instructed to report whether the probe sound was the same as the sound heard repeating (the target sound). Listeners were also provided with a schematic figure resembling Fig. 3 depicting the "same" and "different" trial structures. Following McDermott *et al.* (2011), responses were collected on response boxes labeled "Sure Yes," "Yes," "No," and "Sure No," with a light above each button. Listeners were instructed to use the full range of responses according to their level of confidence. Following each response, the light corresponding to the correct answer (Sure Yes for a same trial, Sure No for a different trial) would turn on briefly before extinguishing as the next trial began.

Listeners first completed 18 practice trials, one same trial and one different trial in each of the nine experimental conditions. Trials were arranged to increase by predicted difficulty, beginning with the largest mismatches between target relative entropy and distractor relative entropy, followed by progressively smaller differences in relative entropy. The order of the same trial and different trial in each condition was randomized. All listeners heard the same practice trials in the same order, including the same randomization of same/different trials.

Next, listeners completed two test sessions of 81 test trials. Each test session contained nine trials (six different trials, three same trials) in each of the nine experimental conditions. Trial orders were fully randomized, with each listener hearing a different randomization. No trials were repeated across sessions, and no sound was heard on more than one trial. Feedback was provided on each trial. Listeners were allowed to take a brief break between these two longer sessions. The entire experiment lasted approximately 45 min.

B. Results

1. Control study

Results from the control study are plotted in Fig. 4. To compare this with analyses of experimental results below, percent correct discrimination in the AXB trials was converted into d' scores following Macmillan and Creelman (2004). Results were analyzed in a 2 (signal processing strategy: FD, TD) by 9 (relative entropy level) repeated measures analysis of variance (ANOVA). Performance differed across the two signal-processing strategies ($F_{1,19} = 9.19$, $p < 0.01$, $\eta_p^2 = 0.33$), with better performance for FD embryos [mean $d' = 1.90$, standard error (s.e.) $= 0.21$] than TD embryos (mean $d' = 1.37$, s.e. $= 0.19$). Performance also differed as a function of relative acoustic entropy ($F_{8,152} = 29.79$, $p < 0.001$, $\eta_p^2 = 0.61$). On the whole, discrimination was better for noise samples with greater acoustic structure (i.e., less relative entropy; see Fig. 4). More telling is the significant interaction between signal-processing strategy and relative acoustic entropy ($F_{8,152} = 8.60$, $p < 0.001$, $\eta_p^2 = 0.31$). Regression slopes were derived for each listener's discrimination of FD and TD stimuli separately, then contrasted in a paired-samples t-test. The inverse relationship between relative acoustic entropy and discrimination was significantly stronger for FD stimuli (mean slope $= -0.51$, s.e. $= 0.06$) than for TD stimuli (mean slope $= -0.18$, s.e. $= 0.03$) ($t_{19} = 6.12$, $p < 0.001$). Finally, the control study revealed that levels of relative entropy presented in target stimuli in experiment 1 (1%, 3.2%, 10%) were all discriminable in isolation, as indicated by a mean d' across all listeners of at least 1 (see Fig. 4).

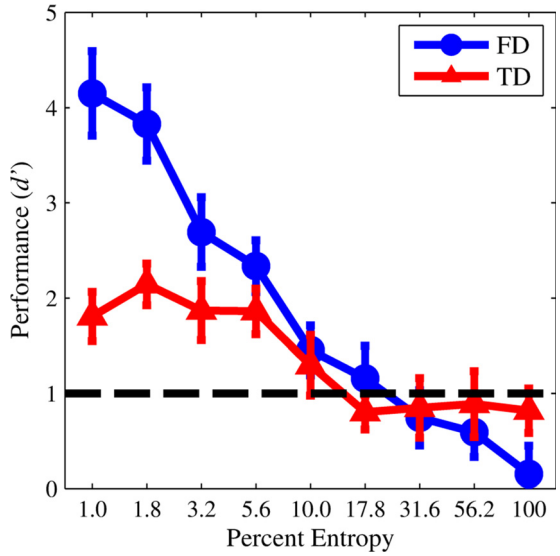

FIG. 4. (Color online) Results from the control study preceding experiment 1. Discriminability (d') is plotted as a function of percent entropy relative to a 300-ms, 5-kHz bandwidth segment of white noise. Performance was measured in percent correct then converted to d' according to Macmillan and Creelman (2004); 100% correct discrimination translated to $d' = 6.29$. Circles depict discrimination of frequency-dilated (FD) embryos, and triangles depict discrimination of time-dilated (TD) embryos. Error bars represent standard error of the mean.

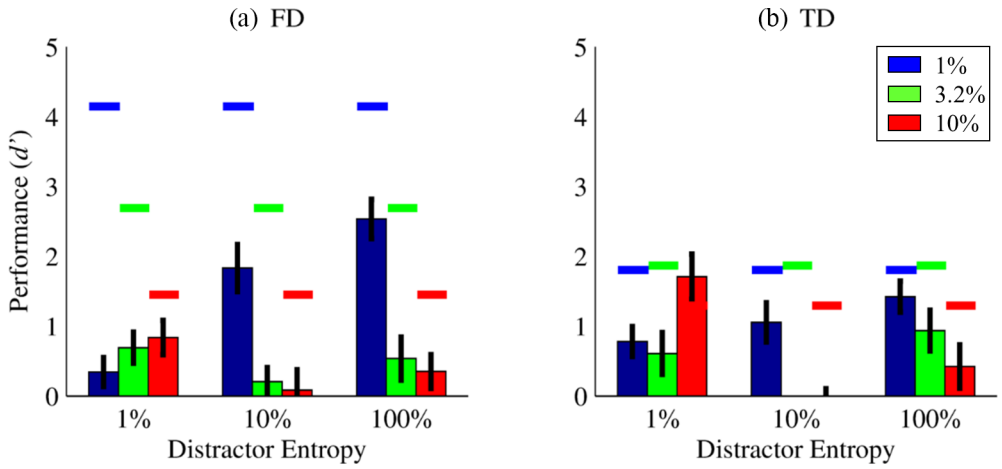

FIG. 5. (Color online) Results from experiment 1. Discriminability (d') is plotted as a function of distractor relative entropy (1%, 10%, 100%). Bars depict different levels of target relative entropy (1%, 3.2%, 10%). Percentages are relative to a 300-ms, 5-kHz-bandwidth segment of white noise. Error bars represent standard error of the mean. (a) Results from experiment 1a, with frequency-dilated white noise embryos. Solid horizontal lines depict discriminability of FD embryos in quiet from the control study. (b) Results from experiment 1b, with time-dilated white noise embryos. Solid lines depict discriminability of TD embryos in quiet from the control study.

2. Experiment 1a: FD

Following the methods of McDermott *et al.* (2011), "Sure Yes" and "Yes" responses were pooled into one category of "same" responses, and "Sure No" and "No" responses were pooled into one category of "different" responses. Given that some conditions were predicted to be difficult (small to no mismatches in target and distractor relative entropy) and others easier (greater mismatches in target and distractor relative entropy), a performance criterion was implemented. Listeners were required to achieve an average d' of at least one across their five best conditions (greater than one-half of the experiment) in order for their data to be included in analyses. This criterion was chosen to retain listeners who demonstrate some level of proficiency in this task and to avoid skewing results in any particular direction, as each listener's five best conditions were chosen blindly.

Twenty-eight listeners participated in experiment 1a. Five listeners failed to meet the performance criterion, so their data were excluded from further analysis. Results from the remaining 23 listeners are shown in Fig. 5(a) and Table II. Results were analyzed in a 3 (target relative entropy: 1%, 3.2%, 10%) by 3 (distractor entropy: 1%, 10%, 100%) repeated measures ANOVA. Performance varied significantly as a function of target relative entropy ($F_{2,44} = 16.63$, $p < 0.001$, $\eta_p^2 = 0.43$). Follow-up paired-sample t-tests with Bonferroni correction for multiple comparisons ($\alpha = 0.017$) indicated that detection of 1% entropy targets (mean $d' = 1.57$, s.e. $= 0.14$) was superior to that of 3.2% entropy targets (mean $= 0.48$, s.e. $= 0.13$; $t_{22} = 7.26$, $p < 0.001$) and 10% entropy targets (mean $= 0.43$, s.e. $= 0.17$; $t_{22} = 4.31$, $p < 0.001$). These latter conditions did not differ from one another ($t_{22} = 0.22$, *n.s.*). Distractor relative entropy exhibited a nonsignificant trend of better performance at higher levels of entropy ($F_{2,44} = 2.64$, $p = 0.09$).

The interaction between target and distractor entropies was significant ($F_{4,88} = 9.91$, $p < 0.001$, $\eta_p^2 = 0.31$). Rather than assessing every possible pairwise comparison, instances where relative entropy levels were matched versus mismatched across target and distractor sounds were of principal interest. Follow-up one-tailed paired-sample t-tests corrected for multiple comparisons ($\alpha = 0.025$) indicated that target embryos were better recognized when their relative entropy differed from that of background embryos. This was true for detection of 1% entropy targets (mean d' in 1% entropy backgrounds $= 0.34$, mean d' in 10% entropy backgrounds $= 1.83$, $t_{22} = 3.64$, $p < 0.01$) and 10% entropy targets (mean d' in 1% entropy backgrounds $= 0.84$, mean d′ in 10% entropy backgrounds $= 0.08$, $t_{22} = 2.15$, $p < 0.025$).

3. Experiment 1b: TD

Thirty-one listeners participated in experiment 1b. Nine listeners failed to meet the performance criterion, so their data were excluded from further analysis. Results from the remaining 22 listeners are shown in Fig. 5(b) and Table II. Results were analyzed in a 3 (target relative entropy) by

TABLE II. Mean d' values and standard errors (in parentheses) for experiment 1. Columns depict levels of relative entropy in the target embryos, and rows depict levels of relative entropy in the distractor embryos. Percentages are relative to a 300-ms, 5-kHz-bandwidth segment of white noise.

		Frequency dilation (expt. 1a)			Time dilation (expt. 1b)		
		Target relative entropy			Target relative entropy		
Distractor relative entropy		1%	3.2%	10%	1%	3.2%	10%
	1%	0.34 (0.21)	0.69 (0.22)	0.84 (0.25)	0.78 (0.22)	0.61 (0.30)	1.71 (0.32)
	10%	1.83 (0.34)	0.21 (0.20)	0.08 (0.29)	1.06 (0.28)	−0.33 (0.28)	−0.14 (0.25)
	100%	2.54 (0.28)	0.54 (0.31)	0.35 (0.24)	1.42 (0.22)	0.94 (0.29)	0.42 (0.31)

3 (distractor relative entropy) repeated measures ANOVA. Performance again differed as a function of target relative entropy ($F_{2,42} = 3.65$, $p < 0.05$, $\eta_p^2 = 0.15$). Follow-up paired-sample t-tests with Bonferroni correction for multiple comparisons ($\alpha = 0.017$) indicated that detection of 1% entropy targets (mean $d' = 1.09$, s.e. $= 0.13$) was superior to that of 3.2% entropy targets (mean $= 0.41$, s.e. $= 0.19$; $t_{21} = 2.85$, $p < 0.01$), but no other comparisons significantly differed (10% entropy targets: mean $= 0.66$, s.e. $= 0.17$). Performance also varied as a function of distractor relative entropy ($F_{2,42} = 9.17$, $p < 0.001$, $\eta_p^2 = 0.30$). Bonferroni-corrected paired-sample t-tests indicated that performance amidst 10% distractors (mean $d' = 0.20$, s.e. $= 0.17$) was significantly poorer than 1% distractors (mean $= 1.03$, s.e. $= 0.15$; $t_{21} = 4.14$, $p < 0.001$) or 100% distractors (mean $= 0.93$, s.e. $= 0.12$; $t_{21} = 3.34$, $p < 0.01$). Performance did not differ for 1% and 100% entropy distractors ($t_{21} = 0.48$, *n.s.*).

The interaction between target and distractor entropies was significant ($F_{4,84} = 5.34$, $p < 0.001$, $\eta_p^2 = 0.20$). As in experiment 1a, the principal comparisons of interest were when target and distractor relative entropy matched versus mismatched. Bonferroni-corrected one-tailed paired-sample t-tests revealed that 10% entropy target sounds were better recognized when their relative entropy differed from that of background sounds (mean d' in 1% entropy backgrounds $= 1.71$, mean d' in 10% entropy backgrounds $= -0.14$, $t_{21} = 4.42$, $p < 0.001$). Detection of 1% entropy targets followed a broadly similar pattern but did not significantly differ (mean d' in 1% entropy backgrounds $= 0.78$, mean d' in 10% entropy backgrounds $= 1.05$; $t_{21} = 0.77$, *n.s.*).

4. Comparisons across experiments 1a and 1b

Performance was also evaluated as a function of signal processing strategy. Results were analyzed in a 2 (signal processing strategy: FD, TD; between-subjects factor) by 3 (target relative entropy; within-subjects factor) by 3 (distractor relative entropy; within-subjects factor) mixed-design ANOVA. There was no main effect of signal processing strategy ($F_{1,43} = 1.00$, *n.s.*), as overall performance in experiments 1a (mean $= 0.83$, s.e. $= 0.07$) and 1b (mean $= 0.72$, s.e. $= 0.08$) was comparable. The signal-processing-strategy-by-target-relative-entropy interaction ($F_{2,86} = 2.31$, $p < 0.11$) did not reach statistical significance, but the signal-processing-strategy-by-distractor-relative-entropy interaction was significant ($F_{2,86} = 4.22$, $p < 0.05$, $\eta_p^2 = 0.09$). Performance improved in a broadly linear fashion with increasing distractor relative entropy in experiment 1a, but exhibited a parabolic trend in experiment 1b with better performance in 1% and 100% entropy backgrounds and poorer performance in 10% backgrounds. Finally, the three-way interaction ($F_{4,172} = 1.98$, $p < 0.10$) exhibited a modest but nonsignificant effect.

C. Discussion

Acoustic structure (relative entropy) was systematically manipulated through the creation of acoustic embryos. Embryos varied from having the same relative entropy as a 300-ms, 5-kHz bandwidth sample of white noise (100%) down to a relatively minimal amount of entropy (1%). Structure was titrated in the frequency domain (through instantaneous frequency dilation) or the temporal domain (through time dilation). In the control study, performance spanned a wider range for FD embryos than TD embryos, but in both cases discriminability closely followed the amount of acoustic structure with better performance for lower-relative-entropy (higher-structure) embryos.

Listeners' recognition of targets varied as functions of acoustic structure in the target embryos as well as the distractor embryos. Lower-relative-entropy targets were detected relatively well amidst higher-relative-entropy distractors, but were more difficult to identify amidst lower-relative-entropy distractors. Conversely, higher-relative-entropy targets were better detected against lower-relative-entropy backgrounds, but not higher-relative-entropy backgrounds. Thus, *differences* in acoustic structure provide a powerful cue for discovery of acoustic structure. This finding parallels greater informational masking (poorer performance) being reported when target and masker sounds were more similar, and less masking (better performance) observed when target and masker sounds were less similar (Barker and Cooke, 1999; Brungart, 2001; Kidd *et al.*, 2002; Durlach *et al.*, 2003b; Leibold and Neff, 2007).

Although the dimension of relative entropy used in this manner results in a somewhat nonlinear mapping to psychoacoustic dimensions, it is important to note that stimuli were more easily recognized when competing stimuli had *different* levels of relative entropy and one would expect this result to hold whether we use either a physical or psychoacoustic measure of entropy.

Instantaneous frequency dilation and time dilation are complementary manipulations of stimulus relative entropy, but these methods did not produce equivalent discriminability for the given noise parameters (5 kHz bandwidth, 300 ms duration). Frequency-dilation stimuli were discriminated significantly better than TD stimuli in the control experiment, yet overall performance did not significantly differ across experiments 1a and 1b. Comparisons of relative-entropy-dilation methods and results are further discussed in Sec. IV.

Experiment 1 assessed discovery of acoustic structure using the embedded repetition task of McDermott and colleagues (2011). The present study builds on these findings in two important ways. First, McDermott and colleagues tested noise samples that were modified to adhere to environmental statistics (i.e., high correlations across short intervals in frequency or in time, lower correlations for larger separations). Here, acoustic embryos were not obligated to follow environmental statistics, instead varying solely as a function of entropy relative to band-limited white noise. Thus, sounds do not need to reflect environmental statistics in order to be recognized following via repetition. Second, patterns of performance differed widely despite testing the same experimental parameters. One of the conditions tested by McDermott and colleagues featured 10 repetitions of the target sound and 20 different distractor sounds that were asynchronous relative to the target sound. Performance was relatively constant in this condition [mean area under the curve $= 0.818$ (their experiment 3, Fig. 3A, condition 1) and 0.838 (their experiment 3, Fig. 3B, condition 1), which corresponds to d' between 1.28 and 1.39].[5] Experiment 1 of the present report tested these same parameters, but performance varied widely as a function of relative entropy in the

target embryos and distractor embryos (means ranging from 0.08 to 2.54 in experiment 1a, from −0.33 to 1.71 in experiment 1b). This difference is likely due to varying the amounts of acoustic structure in test stimuli. Visual inspection of stimulus spectrograms in McDermott *et al.* (2011) show that acoustic structure in their stimuli (i.e., sparseness in the spectrotemporal domain) was fairly constant, but it was varied extensively here (Fig. 2). Thus, recognition following embedded repetition depends on more than only target repetition and distractor variability, as similarities and differences in acoustic structure between target and distractor sounds plays a substantial role.

Frequency- and time-dilation signal processing methods varied acoustic structure quite widely in experiment 1. It is worth noting that this range of acoustic structure is appreciably wider than what is experienced in the natural acoustic environment. Natural sounds possess structure in their frequency and temporal compositions, making white noise (100% entropy embryos) an unrepresentative upper limit of structure heard in everyday environments. Natural sounds do not include relatively equal amounts of energy across frequency. Voss and Clarke (1975) reported that power spectra of speech and music are well-represented by a simple exponential function,[6] $1/f$ (see also Attias and Schreiner, 1997). In the $1/f$ power spectrum, energy decreases linearly with increasing log-frequency. Indeed, the long-term average spectrum of speech approaches a $1/f$ distribution. To better reflect this structure of natural sounds, experiment 2 used FD and TD processing on pink noise samples rather than the white noise used in experiment 1. The same hypotheses from experiment 1 are tested in experiment 2, but with one important distinction. Measuring entropy relative to unprocessed pink noise instead of unprocessed white noise means that entropy has been lowered overall. Given the results of the control study in experiment 1, this overall lowering of relative acoustic entropy should improve performance in experiment 2.

III. EXPERIMENT 2

A. Method

1. Listeners

Sixty-six undergraduate students were recruited for this experiment (20 in the control study; 24 in experiment 2a; 22 in experiment 2b). All reported no known hearing impairment, and none participated in more than one experiment. All received course credit as compensation for participation.

2. Stimuli

Stimuli for experiment 2 were generated in the same manner as in experiment 1 with one exception. Following generation of each 300-ms noise sample, the waveform was filtered using a finite impulse response filter with 100 coefficients and a constant −3 dB/octave slope to create pink noise. Following generation of pink noise samples, stimuli were processed using the same FD or TD methods as described in experiment 1. Now, percentages of relative entropy are relative to unprocessed pink noise as opposed to the unprocessed white noise tested in experiment 1. Dilation of pink noise via FD resulted in novel sounds with nonuniform distributions of instantaneous frequency from 0 to 5000 Hz, with instantaneous frequency biased more towards lower frequencies (see Mm. 7–Mm. 9 for examples). TD stimuli were centered at approximately 500 Hz with bandwidth increasing as a function of relative entropy (see Fig. 6 and Mm. 9–Mm. 12 for examples).

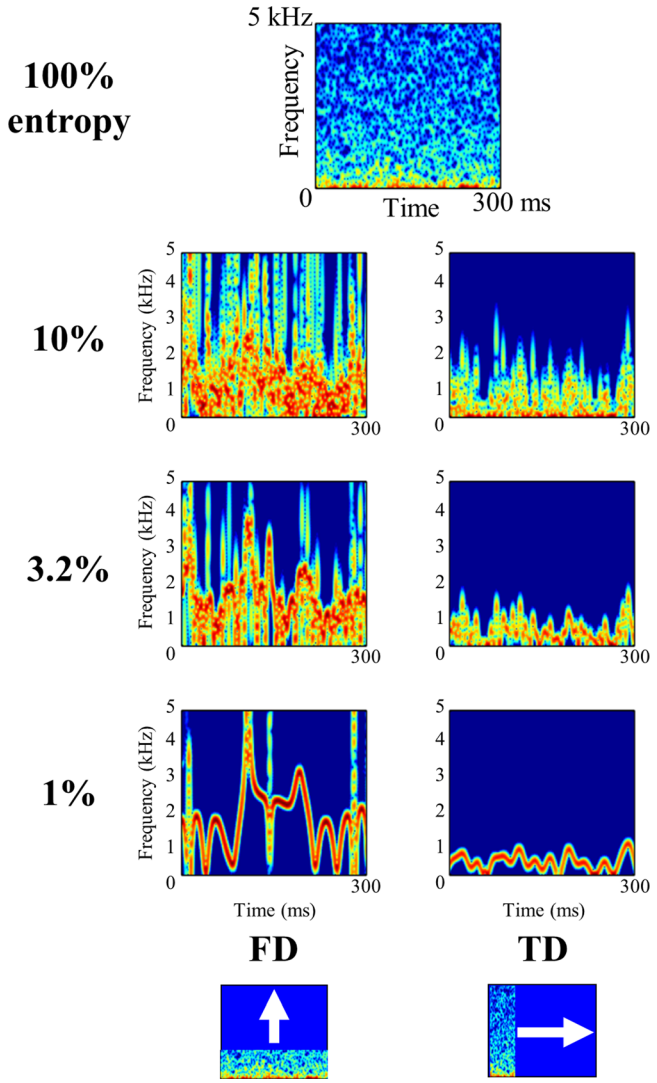

FIG. 6. (Color online) Stimulus generation procedures for experiment 2. The top row depicts the spectrogram of an undilated 300-ms, 5-kHz bandwidth pink noise sample. The left column depicts examples of instantaneous frequency dilation (FD) at the labeled percentages of entropy (relative to the undilated noise at top). As in Fig. 2, in FD, the frequency axis is truncated via low-pass filtering and dilated to the full 5-kHz bandwidth. The right column depicts examples of time dilation (TD), where stimulus duration is truncated and then dilated to the full 300-ms duration.

Mm. 7. Audio of 5 frequency-dilated pink-noise embryos at 1% relative entropy. This is a file of type "wav" (69 KB).

Mm. 8. Audio of 5 frequency-dilated pink-noise embryos at 3.2% relative entropy. This is a file of type "wav" (69 KB).

Mm. 9. Audio of 5 frequency-dilated pink-noise embryos at 10% relative entropy. This is a file of type "wav" (69 KB).

Mm. 10. Audio of 5 time-dilated pink-noise embryos at 1% relative entropy. This is a file of type "wav" (69 KB).

Mm. 11. Audio of 5 time-dilated pink-noise embryos at 3.2% relative entropy. This is a file of type "wav" (69 KB).

Mm. 12. Audio of 5 time-dilated pink-noise embryos at 10% relative entropy. This is a file of type "wav" (69 KB).

Similar to the control study preceding experiment 1, 720 300-ms stimuli were generated (360 FD and 360 TD; 40 stimuli at each of the nine levels of relative entropy for each signal processing strategy) and arranged into two-alternative forced-choice (AXB) discrimination trials. No sounds were heard on more than one trial.

Similar to experiment 1, 30 stimuli were generated for use as target sounds and 1080 stimuli generated for use as distractor sounds for each of nine experimental conditions, generating 9990 FD stimuli (experiment 2a) and 9990 TD stimuli (experiment 2b). Trial structure matched that presented in experiment 1.

3. Procedure

Procedures for control study and the experiment matched those used in experiment 1.

B. Results

1. Control study

Results from the control study for experiment 2 are plotted in Fig. 7. Results were analyzed in a 2 (signal processing strategy) by 9 (percent relative entropy) repeated measures ANOVA. Performance did not significantly differ across the two signal processing strategies (FD mean = 2.66, s.e. = 0.26; TD mean = 2.48, s.e. = 0.28; $F_{1,19} = 1.12$, *n.s.*). Performance again differed substantially as a function of relative entropy ($F_{8,152} = 37.48$, $p < 0.001$, $\eta_p^2 = 0.66$). As in the control study for experiment 1, discrimination was better for noise samples with more acoustic structure (less relative entropy; see Fig. 7). The interaction between signal processing strategy and relative acoustic entropy was also significant ($F_{8,152} = 3.46$, $p < 0.01$, $\eta_p^2 = 0.15$). Regression slopes were derived for each listener's discrimination of FD and TD stimuli separately, then contrasted in a paired-samples *t*-test. The inverse relationship between relative acoustic entropy and discrimination was stronger for FD stimuli (mean slope = −0.44, s.e. = 0.04) than for TD stimuli (mean slope = −0.28, s.e. = 0.04; $t_{19} = 3.17$, $p < 0.01$). The control study confirmed target relative entropy levels tested in experiment 2 were all well discriminated when presented in isolation (discrimination of stimuli at 1%, 3.2%, and 10% all far exceeded $d' = 1$).

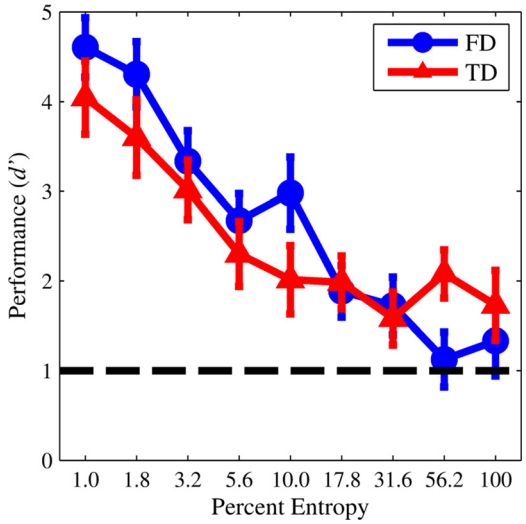

FIG. 7. (Color online) Results from the control study preceding experiment 2. Discriminability (d') is plotted as a function of the percent entropy relative to a 300-ms, 5-kHz-bandwidth segment of pink noise. Performance was measured in percent correct then converted to d' according to Macmillan and Creelman (2004); 100% correct discrimination translated to $d' = 6.29$. Circles depict discrimination of frequency-dilated (FD) embryos, and triangles depict discrimination of time-dilated (TD) embryos. Error bars represent standard error of the mean.

2. Experiment 2a: FD

Two listeners failed to meet the performance criterion, so their data were excluded from further analysis. Results from the remaining 22 listeners are shown in Fig. 8(a) and Table III. Results were analyzed in a 3 (target entropy: 1%, 3.2%, 10%) by 3 (distractor entropy: 1%, 10%, 100%) repeated measures ANOVA. Performance significantly differed as functions of target relative entropy ($F_{2,42} = 6.30$, $p < 0.01$, $\eta_p^2 = 0.23$). Follow-up paired-sample *t*-tests with Bonferroni correction for multiple comparisons ($\alpha = 0.017$) revealed that detection of 1% entropy targets (mean $d' = 1.77$, s.e. = 0.20) was superior to that of 10% entropy targets (mean = 1.06, s.e. = 0.15; $t_{21} = 3.61$, $p < 0.005$), but neither significantly differed from detection of 3.2% entropy targets (mean = 1.39, s.e. = 0.18; $t_{21} < 1.90$, $p > 0.07$). Performance also differed as a function of background relative entropy ($F_{2,42} = 5.19$, $p < 0.01$, $\eta_p^2 = 0.20$). Bonferroni-corrected paired-sample *t*-tests revealed superior performance in 100% entropy backgrounds (mean = 1.79, s.e. = 0.21) compared to 1% entropy backgrounds (mean = 1.14, s.e. = 0.16; $t_{21} = 3.56$, $p < 0.005$), but neither differed from performance in 10% entropy backgrounds (mean = 1.30, s.e. = 0.17; $t_{21} < 2.13$, $p > 0.04$).

The interaction between target and background entropies was significant ($F_{4,84} = 10.12$, $p < 0.001$, $\eta_p^2 = 0.33$). Bonferroni-corrected one-tailed paired-samples *t*-tests revealed that detection of 1% entropy targets was significantly better amidst 10% entropy backgrounds (mean $d' = 2.26$) compared to 1% entropy backgrounds (mean = 0.31; $t_{21} = 5.20$, $p < 0.001$), but no differences were observed for 10% entropy targets (1% entropy background mean = 1.24, 10% entropy background mean = 1.03; $t_{21} = 0.45$, *n.s.*).

3. Experiment 2b: TD

All 22 listeners who participated in experiment 2b met the performance criterion. Results are shown in Fig. 8(b) and Table III. Performance was analyzed in a 3 (target relative entropy) by 3 (distractor relative entropy) repeated measures ANOVA. Performance significantly differed as a function of target relative entropy ($F_{2,42} = 5.64$, $p < 0.01$, $\eta_p^2 = 0.21$). Follow-up paired-sample *t*-tests with Bonferroni correction for multiple comparisons indicated that detection of 1% entropy targets (mean $d' = 1.46$, s.e. = 0.21) was superior to that of 3.2% entropy targets (mean = 0.72, s.e. = 0.20; $t_{21} = 2.79$, $p < 0.011$)

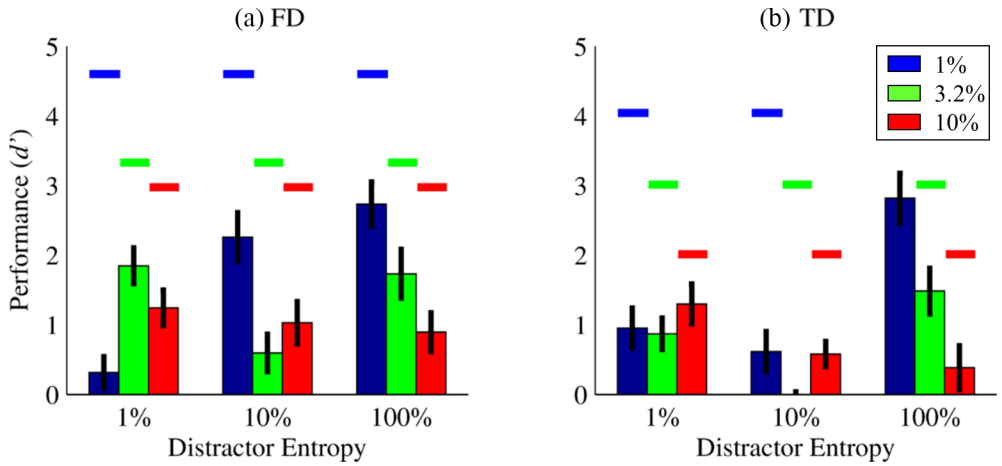

FIG. 8. (Color online) Results from experiment 2. Discriminability (d') is plotted as a function of distractor relative entropy (1%, 10%, 100%). Bars depict different levels of target relative entropy (1%, 3.2%, 10%). Percentages are relative to a 300-ms, 5-kHz-bandwidth segment of pink noise. Error bars represent standard error of the mean. (a) Results from experiment 2a, with frequency-dilated pink noise embryos. Solid horizontal lines depict discriminability of FD embryos in quiet from the control study. (b) Results from experiment 2b, with time-dilated pink noise embryos. Solid lines depict discriminability of TD embryos in quiet from the control study.

and 10% entropy targets (mean = 0.76, s.e. = 0.16; $t_{21} = 2.74$, $p < 0.013$). Performance did not differ across 3.2% and 10% entropy targets ($t_{21} = 0.17$, *n.s.*). Performance also varied as a function of distractor relative entropy ($F_{2,42} = 17.20$, $p < 0.001$, $\eta_p^2 = 0.45$). Bonferroni-corrected paired-sample *t*-tests revealed that performance amidst 10% entropy backgrounds (mean = 0.33, s.e. = 0.14) was significantly poorer than in backgrounds of 1% entropy (mean = 1.04, s.e. = 0.16; $t_{21} = 4.05$, $p < 0.001$) or 100% entropy (mean = 1.56, s.e. = 0.21; $t_{21} = 6.01$, $p < 0.001$). Performance modestly differed in 1% versus 100% entropy backgrounds ($t_{21} = 2.12$, $p = 0.05$).

The interaction between target and distractor entropies was significant ($F_{4,84} = 9.36$, $p < 0.001$, $\eta_p^2 = 0.31$). Follow-up *t*-tests indicate that signal detection was only slightly better when target and background relative entropy levels differed compared to when they were the same, but not sufficiently different to achieve statistical significance. Performance with 1% entropy targets was similar in 1% entropy backgrounds (mean = 0.96) and 10% backgrounds (mean = 0.61, $t_{22} = 0.87$, *n.s.*). Performance with 10% entropy targets approached a statistically significant difference in 1% entropy backgrounds (mean = 1.30) versus 10% backgrounds (mean = 0.58, $t_{22} = 2.10$, $p = 0.05$).

4. Comparisons across experiments 2a and 2b

Performance was also evaluated as a function of signal processing strategy. Results were analyzed in a 2 (signal processing strategy: FD, TD; between-subjects factor) by 3 (target relative entropy; within-subjects factor) by 3 (distractor relative entropy; within-subjects factor) mixed-design ANOVA. There was a significant main effect of signal processing strategy ($F_{1,42} = 5.59$, $p < 0.05$, $\eta_p^2 = 0.12$), with better performance in experiment 2a (FD; mean = 1.41, s.e. = 0.13) than in experiment 2b (TD; mean = 0.98, s.e. = 0.13). The signal-processing-strategy-by-target-relative-entropy interaction was not significant ($F_{2,84} = 0.89$, *n.s.*), but the strategy-by-distractor-relative-entropy interaction was significant ($F_{2,84} = 4.96$, $p < 0.01$, $\eta_p^2 = 0.11$). This was likely due to performance improving with increasing distractor relative entropy for FD stimuli (mean $d' = 1.14$, 1.30, and 1.79 for 1%, 10%, and 100% entropy, respectively) but exhibiting a more quadratic pattern for TD stimuli (mean $d' = 1.04$, 0.33, and 1.56 for 1%, 10%, and 100% entropy, respectively). Finally, the three-way interaction between signal processing strategy, target relative entropy, and distractor relative entropy was significant ($F_{4,168} = 3.23$, $p < 0.05$, $\eta_p^2 = 0.07$). While patterns of performance were highly similar across experiments 2a and 2b for 100% entropy distractor embryos, patterns differed considerably for 1% and 10% entropy distractor embryos (see Fig. 8).

C. Discussion

Experiment 2 used dilated pink noise samples to more closely reflect the range of acoustic structure listeners experience in everyday listening. Across the control and main

TABLE III. Mean d' values and standard errors (in parentheses) for experiment 2. Columns depict levels of relative entropy in the target sounds, and rows depict levels of relative entropy in the distractor sounds. Percentages are relative to a 300-ms, 5-kHz-bandwidth segment of pink noise.

		Frequency dilation (expt. 2a)			Time dilation (expt. 2b)		
		Target relative entropy			Target relative entropy		
Distractor relative entropy		1%	3.2%	10%	1%	3.2%	10%
	1%	0.31 (0.23)	1.85 (0.26)	1.24 (0.26)	0.96 (0.29)	0.87 (0.23)	1.30 (0.29)
	10%	2.26 (0.35)	0.60 (0.27)	1.03 (0.30)	0.61 (0.29)	−0.21 (0.25)	0.58 (0.18)
	100%	2.74 (0.32)	1.73 (0.35)	0.90 (0.28)	2.82 (0.36)	1.49 (0.33)	0.39 (0.31)

experiments, patterns of performance were highly similar to those observed in experiment 1. This replication using pink noise brings the present studies into closer approximation to natural acoustic ecology. However, as stated following experiment 1, adherence to environmental statistics is not a prerequisite for discovering sounds through repetition. Experiment 2 and the experiments of McDermott and colleagues (2011; Masutomi *et al.*, 2016) reported successful segregation for stimuli modeling some aspects of natural sound statistics, but listeners were also successful in experiment 1 when white noise embryos did not adhere to these regularities.

Pink noise has less overall entropy than white noise, so the percentages of relative entropy (1%, 3.2%, 10%, and 100%) reflected smaller absolute amounts of entropy. Given that discrimination improved for stimuli with less relative entropy (Figs. 4, 7), this generated the prediction that pink noise embryos would be better discriminated than white noise embryos. This prediction was assessed by analyzing results from both control studies in a mixed-design, 2 (original noise sample: white, pink; between-subjects factor) by 2 (processing strategy: FD, TD; within-subjects factor) by 9 (percentage acoustic entropy; within-subjects factor) ANOVA. Listeners discriminated pink-noise-dilated stimuli (mean $d' = 2.57$, s.e. $= 0.22$) significantly better than white-noise-dilated stimuli (mean $= 1.63$, s.e. $= 0.22$; $F_{1,38} = 8.95$, $p < 0.01$, $\eta_p^2 = 0.19$). Noise color did not interact with signal processing strategy ($F_{1,38} = 2.09$, *n.s.*) or level of relative acoustic entropy ($F_{1,38} = 2.09$, *n.s.*). The three-way interaction did reach statistical significance ($F_{8,304} = 2.83$, $p < 0.01$, $\eta_p^2 = 0.07$), likely reflecting differences in slopes in performance for TD stimuli across experiments (cf. Figs. 4 and 7).

Recognition of pink noise embryos exhibited similar patterns of performance to the white noise embryos presented in experiment 1. Crossover interactions were again observed, indicating that performance was influenced by similar versus different amounts of acoustic structure in target and background embryos. Similar to the control experiment, overall performance was higher in experiment 2 compared to experiment 1. This was supported by two separate mixed ANOVAs, each evaluating how noise color (white versus pink; between-subjects) interacted with target and distractor relative entropy levels (within-subjects factors) for a given signal processing strategy. For FD stimuli, better performance was observed for pink noise embryos (mean $d' = 1.41$, s.e. $= 0.11$) than white noise embryos (mean $= 0.83$, s.e. $= 0.10$; $F_{1,43} = 15.15$, $p < 0.001$, $\eta_p^2 = 0.26$). No interactions were statistically significant, suggesting patterns of performance did not differ across experiments. For TD stimuli, noise color approached statistical significance in the predicted direction ($F_{1,42} = 3.19$, $p = 0.08$; pink noise: mean $= 0.98$, s.e. $= 0.10$; white noise: mean $= 0.72$, s.e. $= 0.10$). There was a significant three-way interaction between noise color, target relative entropy, and distractor relative entropy ($F_{4,168} = 3.28$, $p < 0.05$, $\eta_p^2 = 0.07$). Compared to experiment 1b, streaming TD embryos resulted in more similar performance levels for 1% entropy distractors, more different performance levels for 100% entropy distractors, and a quadratic shape for 10% entropy distractors [comparing Fig. 8(b) to Fig. 5(b)]. Nevertheless, all results are well-predicted by degree of acoustic structure, and the importance of similarity of relative structure across target and background sounds.

IV. GENERAL DISCUSSION

Natural sounds have structure due to physical constraints upon the sources that produce them. Listeners are sensitive to this structure and exploit it to detect target sounds amidst competing background sounds. The present experiments examined listeners' ability to discover novel auditory objects amongst complex backgrounds as a function of their acoustic structure. Stimuli ("acoustic embryos") were generated by parametrically manipulating the amount of structure in noise in either the frequency domain (through instantaneous frequency dilation, FD) or temporal domain (through time dilation, TD). Listeners' ability to discriminate these acoustic embryos from one another was shown to be dependent upon their relative entropy. Less relative entropy (greater structure) improved discrimination performance. Next, it was demonstrated that listeners discover these acoustic embryos amidst competing sounds sufficiently well to recognize embryos when presented in isolation. Recognition of discovered embryos improved when differences in relative entropy of targets and competing sounds was greater.

The present effort is distinctive relative to many previous efforts to better understand listeners' ability to recognize and discriminate random noise (Guttman and Julesz, 1963; Kaernbach, 1992, 1993; Goossens *et al.*, 2008; Agus *et al.*, 2010; Agus and Pressnitzer, 2013; Andrillon *et al.*, 2015). In previous studies, noise discriminability was exclusively evaluated in terms of noise duration and/or familiarity (repetition). In control experiments here, duration was fixed and no tokens were repeated, and discrimination of dilated white noise embryos exceeded levels reported in previous noise discrimination tasks. For example, Goossens *et al.* (2008) reported d' values between roughly 0.5 and 1.5 for 409-ms noise bands of varying bandwidths. Agus and colleagues (2010) reported detection of repeated 500-ms noise tokens reaching d' of roughly 1.1. Here, d' reached as high as 4.2, with 8 of the 20 listeners achieving perfect discrimination of frequency-dilated white-noise embryos with 1% relative entropy.

Discrimination of pink noise samples was significantly better than discrimination of white noise samples, which is consistent with better discrimination with lesser relative entropy because pink noise has less entropy than white noise. While there are no known reports of pink noise discrimination, as in white noise discrimination experiments, listeners undoubtedly used some short-term spectral properties for discrimination, even if the properties themselves were not necessarily used equally by all listeners (Agus *et al.*, 2013; Andrillon *et al.*, 2015). As the amount of acoustic structure was varied, these features likely varied in both bandwidths and durations (see Figs. 2 and 6). Thus, even small amounts of acoustic structure in noise provided listeners with sufficient evidence to detect differences between noise samples.

In the identification experiments, listeners better recognized embryos in quiet when they were initially discovered amidst competing sounds with lesser or greater relative entropy. This pattern was observed for embryos generated from dilating

white noise samples (experiment 1) and pink noise samples (experiment 2). Consistent with work by McDermott *et al.* (2011) and Masutomi *et al.* (2016), listeners used repetition to segregate novel sounds from complex backgrounds, but here, this ability depended on *differences* in acoustic structure across the two. The importance of differing degrees of structure in targets versus background sounds is consistent with results from informational masking studies where similar signal characteristics in target and masker sounds impeded performance (e.g., Barker and Cooke, 1999; Brungart, 2001; Kidd *et al.*, 2002; Durlach *et al.*, 2003b; Leibold and Neff, 2007). Shinn-Cunningham (2008) has suggested that similarity of perceptual features provides a putative explanation for informational masking that interferes with stream segregation.

Recognition of a target sound amidst competing sounds is often cast in the framework of perceptual masking. Aspects of the present experiments make it difficult to situate results neatly within this framework. On every trial, every acoustic embryo in the background was different. All embryos were randomly selected, and no effort was made to ensure that the spectrum of each background embryo overlapped (energetic masking) or did not overlap with the spectrum of the target embryo. If energetic masking were to significantly affect listeners' abilities to discover structure, one would expect that full (100% relative entropy) white noise embryos in the background would produce the worst performance and 1% relative entropy background embryos (i.e., highly sparse in the spectrotemporal domain) would produce the best performance, but this was not observed. Additionally, a follow-up experiment by McDermott *et al.* (2011) (from which the present experimental paradigm was taken) created distractors that were tailor-made to produce energetic versus informational masking of the target sound. Equivalent results were observed in both conditions, leading them to conclude that "both energetic and informational masking contribute to the difficulty of segmenting our sound mixtures, but that hearing a sound multiple times in distinct mixtures can ameliorate both factors" (p. 2 of their Supporting Information).

Two complementary signal-processing strategies were employed to create acoustic structure from noise: instantaneous frequency dilation (FD) and time dilation (TD). Levels of relative entropy (1%, 3.2%, and 10%) cannot be taken to equate entropy between TD and FD conditions because choice of bandwidth (5 kHz) and duration (300 ms) were chosen on the basis of experimental convenience. Consequently, 10% TD is unlikely to have equivalent relative entropy to 10% FD. In experiment 1, FD embryos were significantly easier to discriminate; however, listeners' ability to discover and recognize TD versus FD embryos was not significantly different. In experiment 2 using pink noise, discrimination was comparable for both signal-processing strategies, but listeners were significantly better discovering and recognizing FD embryos. Thus, differences across signal-processing strategies were not systematic across the present experiments.

The present results bear similarities and differences to perception of speech amidst competing sounds through glimpsing. Glimpsing refers to extraction of local spectrotemporal regions with advantageous signal-to-noise ratios to facilitate speech recognition (Cooke, 2006). Discovering structure of novel sounds might hypothetically benefit from glimpsing. For example, discovering a repeated target sound improves with additional repetitions until asymptote (McDermott *et al.*, 2011). This could be interpreted as accumulating additional glimpses of the target sound against the varying background sounds, because performance is poorer when the same background sound is repeated throughout the trial and all glimpses are the same (McDermott *et al.*, 2011).

Glimpsing cannot explain the present results. For acoustic embryos, ability to discover acoustic structure was broadly better for highest relative entropy (minimum structure) competing sounds as well as dependent upon differences between relative entropy of target versus background. For speech, there appears to be an inverse monotonic relationship between detecting speech sounds and complexity (relative entropy) of the background. Speech recognition performance progressively declines when the background noise changes from a single competing talker to modulated noise to unmodulated noise (e.g., Carhart *et al.*, 1969; Festen and Plomp, 1990; Brungart, 2001). Recognition of consonants in vowel-consonant-vowel contexts decreases as additional talkers are added to multitalker babble (Simpson and Cooke, 2005). To the extent that glimpsing explains performance with speech, glimpsing cannot explain in full the patterns of performance for discovering acoustic structure of novel sounds.

Finally, it is important to note that the present experiments employed two out of many possible methods for manipulating acoustic structure in complex sounds. Varying relative acoustic entropy clearly modulated the degree of spectrotemporal sparsity in these novel stimuli (see Figs. 2 and 6), but instantaneous frequency dilation and time dilation are certainly not the only ways to vary acoustic structure in sounds. There are infinite routes that can be taken to traverse the space between the relative maximum (random noise) and minimum (pure tones, clicks) amounts of acoustic structure in sounds (Fig. 1). The present approaches enjoyed the virtue of simplicity, as no *post hoc* stimulus editing was needed. However, at the same time, directly controlling relative acoustic entropy came at the expense of directly controlling specific acoustic characteristics of these novel sounds. Alternative approaches to this research question might utilize stimuli whose acoustic characteristics are controlled more directly. Equivalently, other efforts might choose a different standard stimulus to which relative levels of entropy are measured (here, random noise with 5 kHz bandwidth and 300 ms duration). These and other possibilities may be combined with the present results in the effort to systematically traverse the stimulus space between maximum- and minimum-structure sounds as depicted in Fig. 1, thereby offering a more global account of perceptual sensitivity to acoustic structure.

In conclusion, studies reported here provide compelling evidence that listeners rapidly discover structure in novel sounds. Sounds with less relative entropy (greater structure) are easier to discriminate from one another despite equivalent bandwidth and duration. Relatively few exposures to structured sounds is required for listeners to discover structure amidst competing sounds sufficiently well to recognize these sounds in quiet. Taken together, results are consistent with hypotheses of efficient coding by sensorineural systems

in ways complementary to a structured natural world (Attneave, 1954; Barlow, 1961; Bell and Sejnowski, 1997; Lewicki, 2002; Smith and Lewicki, 2006; Kluender *et al.*, 2013).

ACKNOWLEDGMENTS

The authors wish to thank Josh McDermott for providing data from McDermott *et al.* (2011). Funding was provided by SSHRC (to M.K.) and NIDCD (to K.R.K.).

[1]See supplementary material at http://dx.doi.org/10.1121/1.5031018 for MATLAB code used to generate these novel sounds.

[2]The instantaneous frequency of a random noise is normally distributed with center frequency at the center of the bandwidth of the original stimulus and a variance proportional to the bandwidth (Broman, 1981). Some instantaneous frequencies are necessarily negative and this is expected. Similarly, some frequencies may be beyond the Nyquist frequency. However, to achieve the desired bandwidth from the low-pass-filtered stimuli, we needed to scale the instantaneous frequencies—negative and positive—which entails widening their statistical distribution. The instantaneous frequencies were scaled by the ratio of the target bandwidth to the bandwidth of the low-pass-filtered stimulus. Following the example of 10% entropy described above, the instantaneous frequencies were scaled by a factor of 10.

[3]Note that it does not matter whether the first difference of the phase is scaled or the phase itself, as the instantaneous frequency will be scaled by an equivalent amount in both cases.

[4]In McDermott *et al.* (2011), performance asymptoted with five repetitions of the target, and peak performance was achieved when no distractors were repeated. Additionally, performance was comparable when targets were centered over distractor-pair ISIs (as described above), concurrent with every other distractor sound, or concurrent with presentation of every distractor. Thus, trial parameters tested in the present experiments produced the highest observed performance in McDermott *et al.* (2011).

[5]Values were calculated by converting the area under the curve to a *z*-score using the qnorm function in R and multiplying by $\sqrt{2}$.

[6]$1/f$ noise is sometimes expressed as $1/f^{\alpha}$, with the value of the exponent α equaling 1 or a small variation thereof depending upon the stimulus set (van der Schaaf and van Hateren, 1996). Varying this exponent generates a family of $1/f^{\alpha}$ curves, all of which are linear in log-frequency space but with different (negative) slopes. Pink noise belongs to this $1/f^{\alpha}$ family but with the distinguishing characteristic of equal power in each octave.

Agus, T. R., and Pressnitzer, D. (**2013**). "The detection of repetitions in noise before and after perceptual learning," J. Acoust. Soc. Am. **134**(1), 464–474.

Agus, T. R., Thorpe, S. J., and Pressnitzer, D. (**2010**). "Rapid formation of robust auditory memories: Insights from noise," Neuron **66**(4), 610–618.

Andreou, L.-V., Kashino, M., and Chait, M. (**2011**). "The role of temporal regularity in auditory segregation," Hear. Res. **280**(1-2), 228–235.

Andrillon, T., Kouider, S., Agus, T., and Pressnitzer, D. (**2015**). "Perceptual learning of acoustic noise generates memory-evoked potentials," Curr. Biol. **25**(21), 2823–2829.

Attias, H., and Schreiner, C. E. (**1997**). "Temporal low-order statistics of natural sounds," Adv. Neural Inf. Process. Syst. **9**, 27–33.

Attneave, F. (**1954**). "Some informational aspects of visual perception," Psychol. Rev. **61**(3), 183–193.

Barker, J., and Cooke, M. (**1999**). "Is the sine-wave speech cocktail party worth attending?," Speech Commun. **27**(3-4), 159–174.

Barlow, H. B. (**1961**). "Possible principles underlying the transformation of sensory messages," in *Sensory Communication* (MIT Press Scholarship Online, Cambridge, MA), pp. 217–234.

Bell, A. J., and Sejnowski, T. J. (**1997**). "The 'independent components' of natural scenes are edge filters," Vision Res. **37**(23), 3327–3338.

Bendixen, A. (**2014**). "Predictability effects in auditory scene analysis: A review," Front. Neurosci. **8**, 1–16.

Brady, M. J., and Kersten, D. (**2003**). "Bootstrapped learning of novel objects," J. Vision **3**(6), 413–422.

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

Broman, H. (**1981**). "The instantaneous frequency of a Gaussian signal: The one-dimensional density function," IEEE Trans. Acoust., Speech, Signal Process. **29**(1), 108–111.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**(3), 1101–1109.

Carhart, R., Tillman, T. W., and Greetis, E. S. (**1969**). "Perceptual masking in multiple sound backgrounds," J. Acoust. Soc. Am. **45**(3), 694–703.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears,"J. Acoust. Soc. Am. **25**(5), 975–979.

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**(3), 1562–1573.

Durlach, N., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. (**2003a**). "Note on informational masking (L).," J. Acoust. Soc. Am. **113**(6), 2984–2988.

Durlach, N., Mason, C. R., Shinn-Cunningham, B., Arbogast, T. L., Colburn, H. S., and Kidd, G. (**2003b**). "Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity," J. Acoust. Soc. Am. **114**(1), 368–380.

Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**(4), 1725–1736.

Goossens, T., van de Par, S., and Kohlrausch, A. (**2008**). "On the ability to discriminate Gaussian-noise tokens or random tone-burst complexes," J. Acoust. Soc. Am. **124**(4), 2251–2262.

Greenwood, D. D. (**1961**). "Auditory masking and the critical band," J. Acoust. Soc. Am. **33**(4), 484–502.

Guttman, N., and Julesz, B. (**1963**). "Lower limits of auditory periodicity analysis," J. Acoust. Soc. Am. **35**, 610.

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., and Carlyon, R. P. (**2013**). "Swinging at a cocktail party voice familiarity aids speech perception in the presence of a competing voice," Psychol. Sci. **24**(10), 1–10.

Kaernbach, C. (**1992**). "On the consistency of tapping to repeated noise," J. Acoust. Soc. Am. **92**(2), 788–793.

Kaernbach, C. (**1993**). "Temporal and spectral basis of the features perceived in repeated noise," J. Acoust. Soc. Am. **94**(1), 91–97.

Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (**2008**). "Informational masking," in *Auditory Perception of Sound Sources*, edited by W. A. Yost, A. N. Popper, and R. R. Fay (Springer, New York), pp. 143–189.

Kidd, G., Jr., Mason, C. R., and Arbogast, T. L. (**2002**). "Similarity, uncertainty, and masking in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. **111**(3), 1367–1376.

Kluender, K. R., Stilp, C. E., and Kiefte, M. (**2013**). "Perception of vowel sounds within a biologically realistic model of efficient coding," in *Vowel Inherent Spectral Change*, edited by G. S. Morrison and P. F. Assmann (Springer, Berlin), pp. 117–151.

Leibold, L. J., and Neff, D. L. (**2007**). "Effects of masker-spectral variability and masker fringes in children and adults," J. Acoust. Soc. Am. **121**(6), 3666–3676.

Lewicki, M. S. (**2002**). "Efficient coding of natural sounds," Nat. Neurosci. **5**(4), 356–363.

Lutfi, R. A. (**1990**). "How much masking is informational masking?," J. Acoust. Soc. Am. **88**(6), 2607–2611.

Macmillan, N. A., and Creelman, C. D. (**2004**). *Detection Theory: A User's Guide*, 2nd ed. (Lawrence Erlbaum Associates, Mahwah, NJ).

Masutomi, K., Barascud, N., Kashino, M., McDermott, J. H., and Chait, M. (**2016**). "Sound segregation via embedded repetition is robust to inattention," J. Exp. Psychol., Human Percept. Perform. **42**(3), 386–400.

McDermott, J. H., Wrobleski, D., and Oxenham, A. J. (**2011**). "Recovering sound sources from embedded repetition," Proc. Natl. Acad. Sci. **108**(3), 1188–1193.

Neff, D. L., and Green, D. M. (**1987**). "Masking produced by spectral uncertainty with multicomponent maskers," Percept. Psychophys. **41**(5), 409–415.

Newman, R. S., and Evers, S. (**2007**). "The effect of talker familiarity on stream segregation," J. Phonet. **35**(1), 85–103.

Pollack, I. (**1975**). "Identification of random auditory waveforms," J. Acoust. Soc. Am. **58**(6), 1262–1271.

Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cognit. Sci. **12**(5), 182–186.

Simpson, S. A., and Cooke, M. (**2005**). "Consonant identification in N-talker babble is a nonmonotonic function of N," J. Acoust. Soc. Am. **118**(5), 2775–2778.
Smith, E., and Lewicki, M. S. (**2006**). "Efficient auditory coding," Nature **439**(7079), 978–982.
Stilp, C. E., and Kluender, K. R. (**2011**). "Non-isomorphism in efficient coding of complex sound properties," J. Acoust. Soc. Am. **130**(5), EL352–EL357.
Stilp, C. E., and Kluender, K. R. (**2012**). "Efficient coding and statistically optimal weighting of covariance among acoustic attributes in novel sounds," PLoS One **7**(1), e30845.
Stilp, C. E., and Kluender, K. R. (**2016**). "Stimulus statistics change sounds from near-indiscriminable to hyperdiscriminable," PLoS One **11**(8), e0161001.
Stilp, C. E., Rogers, T. T., and Kluender, K. R. (**2010**). "Rapid efficient coding of correlated complex acoustic properties," Proc. Natl. Acad. Sci. U.S.A. **107**(50), 21914–21919.
Summerfield, Q., and Assmann, P. F. (**1989**). "Auditory enhancement and the perception of concurrent vowels," Percept. Psychophys. **45**(6), 529–536.
van der Schaaf, A., and van Hateren, J. H. (**1996**). "Modelling the power spectra of natural images: Statistics and information," Vision Res. **36**(17), 2759–2770.
Voss, R. F., and Clarke, J. (**1975**). "1-F-noise in music and speech," Nature **258**(5533), 317–318.
Watson, C. S., Kelly, W. J., and Wroton, H. W. (**1976**). "Factors in the discrimination of tonal patterns. II. Selective attention and learning under various levels of stimulus uncertainty," J. Acoust. Soc. **60**, 1176–1186.
Watson, C. S., Wroton, H. W., Kelly, W. J., and Benbassat, C. A. (**1975**). "Factors in the discrimination of tonal patterns. I. Component frequency, temporal position, and silent intervals," J. Acoust. Soc. Am. **57**, 1175–1185.
Wegel, R. L. F., and Lane, C. E. (**1924**). "The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear," Phys. Rev. **23**(2), 266–285.