Recovering sound sources from embedded repetition

Josh H. McDermott[a,1], David Wrobleski[b], and Andrew J. Oxenham[b]

[a]Center for Neural Science, New York University, New York, NY 10003 and [b]Department of Psychology, University of Minnesota, Minneapolis, MN 55455

Edited by Edward Adelson, Massachusetts Institute of Technology, Cambridge, MA, and approved December 1, 2010 (received for review April 8, 2010)

Cocktail parties and other natural auditory environments present organisms with mixtures of sounds. Segregating individual sound sources is thought to require prior knowledge of source properties, yet these presumably cannot be learned unless the sources are segregated first. Here we show that the auditory system can bootstrap its way around this problem by identifying sound sources as repeating patterns embedded in the acoustic input. Due to the presence of competing sounds, source repetition is not explicit in the input to the ear, but it produces temporal regularities that listeners detect and use for segregation. We used a simple generative model to synthesize novel sounds with naturalistic properties. We found that such sounds could be segregated and identified if they occurred more than once across different mixtures, even when the same sounds were impossible to segregate in single mixtures. Sensitivity to the repetition of sound sources can permit their recovery in the absence of other segregation cues or prior knowledge of sounds, and could help solve the cocktail party problem.

auditory scene analysis | cocktail party problem | generative models of sound | natural sound statistics | sound segregation

Auditory scenes generally contain multiple sources, the sounds from which add together to produce a mixed signal that enters the ears. In most behavioral contexts, however, it is the sources, not the mixture, that are of interest. This is often termed the "cocktail party problem"—organisms must infer individual sound sources from ambiguous mixtures of sounds (1–7).

Recovering individual sound sources from an auditory scene requires assumptions, or priors, about what sources are like (8). For instance, listeners implicitly assume that frequency components that are regularly spaced (9, 10), begin and end simultaneously (11), or have similar distributions of binaural spatial cues (12) belong to the same sound. Listeners also use knowledge of specific familiar sound classes, filling in masked syllable segments in ways that are consistent with known speech acoustics (13).

Priors on sounds are thus used by the auditory system and must somehow be acquired; yet natural environments rarely feature isolated sound sources from which they could be readily learned. Organisms face a "chicken and egg" problem—sound sources must be separated from mixtures for their properties to be learned, but to separate sources from mixtures, listeners need to know something about their characteristics to begin with.

It is possible that priors are at least partially built into the auditory system by evolution, or that listeners can learn them from occasionally hearing sound sources in isolation. In this paper we consider an alternate, complementary, solution—that listeners might detect sources as repeating spectro-temporal patterns embedded in the acoustic input. Both individual sound sources and their mixtures produce combinations of acoustic features, but because mixtures result from multiple independent sources, the feature configurations that they produce are unlikely to occur repeatedly with consistency. Repetition is thus a signature of individual sources. The repetition of a sound source is generally not explicit in the signal that enters the ear, due to the corruption of a source's acoustic signature by other sounds. However, repeating sources induce temporal regularities in the mixed auditory input, which we suggest are detected and used by the auditory system to recover sound sources.

To explore this idea, we studied the conditions under which listeners could identify novel sound sources that they only ever heard in mixtures with other sounds. We developed a method to synthesize novel sounds that shared some of the correlation structure of natural sounds (14–16) but that lacked strong grouping cues, and presented listeners with mixtures of these sounds. Listeners were generally unable to identify the sounds composing a single such mixture, but when presented with multiple mixtures of a particular target sound with various others, they heard the target repeating across mixtures and could reliably identify it. Even two presentations of the target yielded a significant benefit.

Our results indicate that listeners detect latent repeating spectro-temporal structure within sound mixtures and from this can identify individual sound sources. Sound source repetition thus serves as a powerful cue that can "bootstrap" performance in situations in which other bottom-up cues and top-down knowledge are unavailable, and as such may play an important role in auditory scene analysis.

Results

Generative Model for Sounds. To test whether source repetition might by itself be sufficient for sound segregation, it was important both to use novel sounds, so that familiarity would not enable segregation, and to minimize the presence of bottom-up grouping cues in our test stimuli. However, we wanted our results to have real-world relevance, and thus to use stimuli with some similarity to natural sounds. We met these goals by modeling the time-frequency decomposition (spectrogram) of a sound as a Gaussian-distributed random variable with correlations that resembled those in natural sounds.

We first generated spectrograms for sets of spoken words (Fig. 1*A*) and animal vocalizations (Fig. 1*B*). Such spectrograms generally share a simple property: the energy at nearby points tends to be similar (14–16). This is evident when the correlation between pairs of spectrogram cells is plotted as a function of their time and frequency offset (Fig. 1 *C* and *D*). For both classes of natural sounds, correlations are high for small offsets and decline with separation in time or frequency, whereas for noise signals they are absent. Such results follow from the common finding that natural modulation spectra (related to correlation functions via the Fourier transform) peak at low modulation frequencies (14–16) and thus exhibit correlations over moderate time/frequency scales.

We used correlation functions similar to those of natural sound sets (Fig. 1 *C* and *D*) to generate a covariance matrix, each element of which was the covariance between two spectrogram cells. Spectrograms were drawn from the resulting Gaussian distribution and applied to samples of white noise, yielding novel sounds (Fig. 1 *E* and *F*). Related stimuli result from constraining the modulation spectrum of noise (16); our spectrogram-domain method had advantages in implementing our task (*SI Materials and Methods*). Although our stimuli shared important statistical properties of real sounds, they lacked the grouping cues provided by abrupt temporal onsets and harmonic spectral structure, both of which are important for sound segregation (1, 2) but which are not captured by second-order correlations.

Author contributions: J.H.M. and A.J.O. designed research; J.H.M. and D.W. performed research; J.H.M. analyzed data; and J.H.M. and A.J.O. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: jhm@cns.nyu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1004765108/-/DCSupplemental.

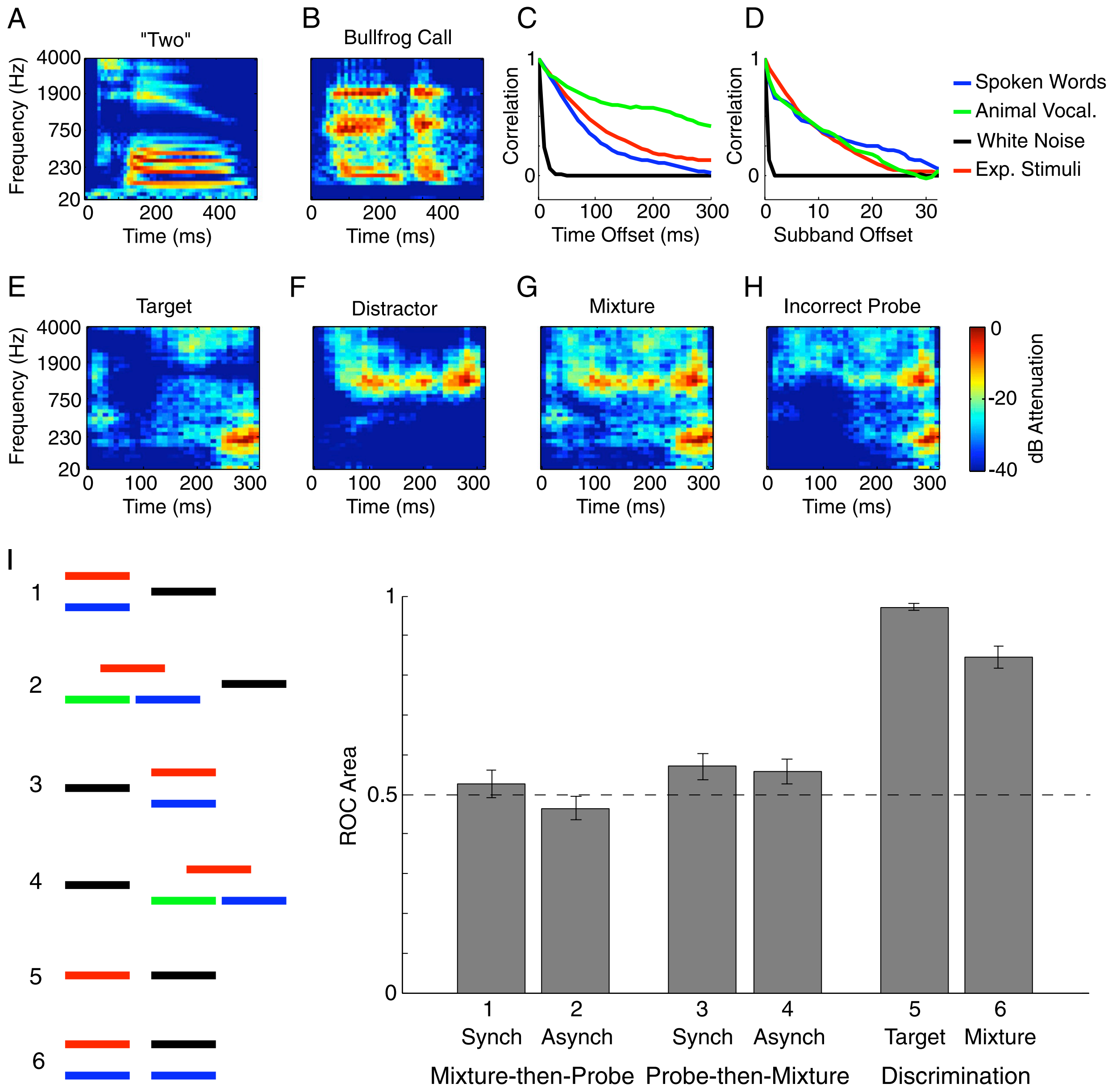

Fig. 1. Stimulus generation and results of Experiment 1. (*A* and *B*) Time-frequency decomposition of a spoken word and a bullfrog vocalization. (*C* and *D*) Correlation between nearby time-frequency cells as a function of their temporal (*C*) and spectral (*D*) separation. (*E* and *F*) Two spectrograms generated by our model. (*G*) Spectrogram of the mixture of the sounds from *E* and *F*. (*H*) Spectrogram of an incorrect probe sound, generated to be physically consistent with the mixture in *G*. (*I*) Results and stimulus configurations from Experiment 1. Line segments represent sounds; sounds presented simultaneously are drawn as vertically displaced. Distinct sounds are indicated by different colors. Red segments represent target sounds, and black segments represent probe sounds. Error bars denote SEs. The dashed line represents the chance performance level.

Performance-Based Measure of Sound Segregation. We assessed sound segregation by presenting mixtures of sounds (Fig. 1*G*) followed by a probe sound. Listeners judged whether the probe had been present in the mixture(s). The probe was either one of the sounds in the mixture(s), termed the "target" sound, or another sound with statistics similar to the target (Fig. 1*H*). In the latter case, the probe was constrained to be physically consistent with the mixture (such that—like the target—it never had more energy than the mixture). Each target was presented only once per experiment, so that subjects could not learn the targets from the probes.

Following the probe presentation, subjects selected one of four responses ("sure no," "no," "yes," or "sure yes") to indicate whether they thought the probe was one of the sounds in the mixture. These responses were used to generate a receiver operating characteristic (ROC) curve. The area beneath the curve was our performance measure (17); chance and perfect performance corresponded to areas of 0.5 and 1, respectively. All of the effects reported here are evident in the stimulus examples available at http://www.cns.nyu.edu/~jhm/source_repetition.

Experiment 1: Sound Segregation with Single Mixtures. We began by presenting subjects with single mixtures of two sounds (Fig. 1*I*). Sound segregation should permit a listener to judge whether a subsequent probe sound was one of the sounds in the mixture. However, performance was generally at chance levels, even after considerable practice [condition 1: $t(9) = 0.64$, $P = 0.54$]. Performance remained close to chance when we included a third sound and made the sounds asynchronous [condition 2: $t(9) = -0.65$, $P = 0.53$]. Asynchrony should enhance the bottom-up grouping cue provided by onset differences between sources (1, 2, 11);

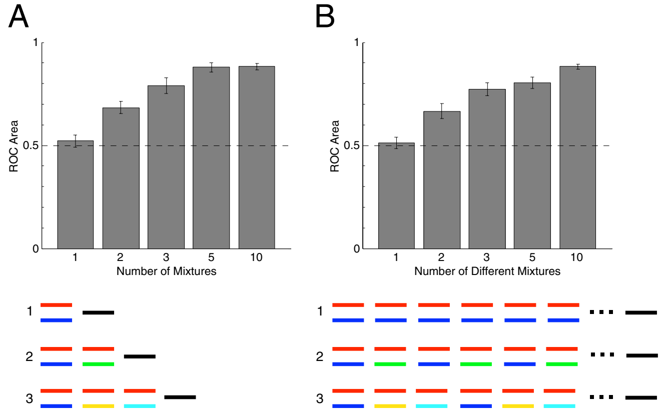

Fig. 2. Effect of multiple mixtures on sound source recovery. (*A*) Different numbers of mixtures were presented. (*B*) Ten mixtures were presented in all conditions, and the number of different mixtures was varied. Conventions here and elsewhere are as in Fig. 1*I*. Red segments represent target probes, black segments represent incorrect probes, and different colors represent different sounds. Schematics for conditions with 5 and 10 mixtures are omitted.

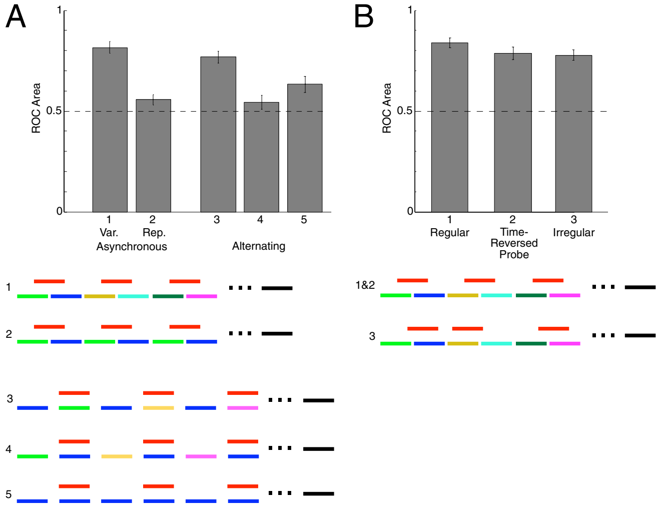

Fig. 3. Stimuli and results of Experiment 3. (*A*) Effect of mixture variability persists with asynchronous and alternating presentation. Conditions 3 and 4 differ in the pairing of the target with variable (condition 3) or repeated (condition 4) distractors. (*B*) Subjects can perform task even when incorrect probes are time-reversed versions of the target sound, or when the target sound is presented irregularly.

the lack of effect suggests that any onsets in our stimuli were too weak to support segregation. We also tried presenting the probe sound before the mixture, so that subjects knew what sound to listen for, but performance was still not significantly different from chance [condition 3, synchronous: $t(9) = 2.23$, $P = 0.053$; condition 4, asynchronous: $t(9) = 1.8$, $P = 0.1$], although there was a small effect of hearing the probe first [$F(1,9) = 7.33$, $P = 0.02$].

The poor performance was not due to an inability to discriminate different synthetic sounds; the correct and incorrect probe sounds were easily distinguished when presented in isolation [condition 5: $t(9) = 56.1$, $P < 10^{-12}$]. Moreover, when the target and incorrect probe sounds for a particular mixture were each mixed with the same unrelated second sound, the resulting mixtures themselves were discriminable [condition 6: $t(9) = 12.5$, $P < 10^{-6}$]. Thus, chance performance in the sound segregation task was not due to limits on encoding of the mixtures (as it would be if the stimulus differences needed to perform our task were completely masked by the other sound in the mixture). Rather, performance was evidently limited by the inability to segregate the mixture into two sounds. The subjective experience of listening to the mixtures was consistent with this conclusion. The mixtures usually sounded like a single sound that was qualitatively different from the target sound.

These results indicate that our stimuli met our principal objectives. Despite having some naturalistic structure, they lacked the grouping cues needed to segregate them from a mixture. This made them well suited to our primary goal of testing whether sound structure could be extracted from multiple occurrences of a target sound.

Experiment 2: Sound Segregation with Multiple Mixtures. To test whether listeners could benefit from sound source repetition across mixtures, we presented target sounds repeatedly, each time mixed with a different "distractor" sound. Despite the difficulty of segregating single mixtures, a target presented more than once in succession was usually heard repeating through the mixtures, and listeners rapidly developed an impression of it. In Experiment 2a we quantified this benefit, varying the number of mixtures and measuring how well subjects could discriminate correct from incorrect target probes. Performance was again at chance levels with a single mixture, but improved as subjects heard more mixtures (Fig. 2*A*). Performance was significantly improved even with two mixtures [$t(9) = 3.66$, $P = 0.005$] and appeared to asymptote with about five mixtures.

To rule out the possibility that the improvement with multiple mixtures was due merely to repeated exposure to the target, in Experiment 2b we held the number of mixtures constant at 10, but varied how many different mixtures occurred in the sequence. In the single-mixture condition, subjects heard the same mixture 10 times. The 10-mixture condition was the same as in Experiment 2a. The other conditions repeatedly presented two, three, or five mixtures in a fixed order over the course of the sequence, with each mixture containing the target sound.

Performance again steadily increased with the number of different mixtures (Fig. 2*B*), even though the target was always presented the same number of times. The ability to hear the target sound thus appears to depend on the number of different mixtures that a listener hears, not on the total number of target presentations. An ANOVA comparing the two experiments showed a main effect of the number of different mixtures [$F(4,36) = 115.35$, $P < 0.0001$], but no effect of experiment type [$F(1,9) = 0.73$; $P = 0.42$] and no interaction [$F(4,36) = 0.59$, $P = 0.67$]. See *SI Results* for additional controls.

As with the single mixtures of Experiment 1, the sounds composing the single repeated mixtures tended to blend together and rarely bore close resemblance to the target sound. This is consistent with the idea that listeners detect repeating sound structure and attribute it to individual sources; when the same mixture repeats, it is heard as a source, and the target structure is no more apparent than when it is heard only once.

Experiment 3a: Asynchronous Mixtures. Experiment 2 featured synchronously presented sounds, but distinct sources in real-world scenes are generally asynchronous. Experiment 3a confirmed that the benefit of multiple distinct mixtures persisted when the target and distractors were temporally offset to better resemble natural conditions (Fig. 3*A*, *Left*, condition 1 vs. condition 2). As before, a single repeated mixture yielded near-chance performance, but presenting different mixtures in succession enabled discrimination of the target sound [$F(1,7) = 116.87$, $P < 0.0001$]. The effect of multiple mixtures in this case swamps that of any grouping cue provided by the asynchrony (consistent with the weak onsets in our sounds), and is not specific to synchronously presented sounds.

The effect was also evident when the target sound was presented with every other distractor in a sequence (Fig. 3*A*, *Right*, conditions 3–5). When the distractors that co-occurred with the target varied (condition 3), performance was well above chance, even though the distractors that alternated with the target repeated ($P = 0.004$, sign test). But when the distractor sequence

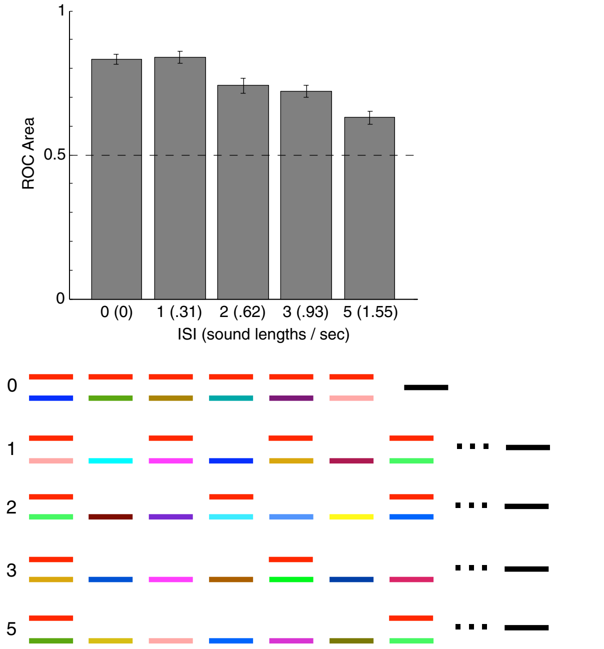

Fig. 4. Effect of interstimulus interval. In all conditions, the target sounds (shown in red) were presented six times. Condition 0 is identical to the variable mixture conditions of Experiment 2 except for the number of target presentations.

was phase-shifted by a target length, so that the repeating distractors co-occurred with the target (condition 4), the target was generally unidentifiable. When every distractor repeated (condition 5), performance tended to be intermediate between the other two conditions (significantly worse than the variable condition and better than the repeated condition, $P = 0.008$ and 0.06, respectively, sign test; also better than condition 1, Experiment 2b, $P = 0.06$). This configuration is reminiscent of some used in studies of pure tone streaming (18). In this condition, the repetition of the distractor may compete with that of the mixture.

Experiment 3b: Spectrotemporal Structure and Irregular Presentation. To test whether listeners extracted the temporal structure of sounds in addition to their spectral content, in Experiment 3b we presented variable mixtures but used a time-reversed version of the target sound for the incorrect probe (that thus had the same power spectrum as the target but differed in temporal structure). As shown in Fig. 3*B*, performance remained high when distinguishing between the correct and the time-reversed probes, although there was a slight advantage with our standard incorrect probes [$F(2,18) = 4.03$, $P = 0.04$]. Listeners thus derived a spectrotemporal profile for the target sound and did not merely encode the average spectrum of the mixture sequence. Performance also remained high when the targets were presented at irregular temporal intervals (Fig. 3*B*), indicating that periodically occurring acoustic structure was not necessary for the effect ($P < 0.0001$ for both conditions, two-tailed t test).

Experiment 4: Temporal Integration. If the benefit of multiple mixtures on sound segregation reflects the extraction of repeating structure from the auditory input, it should be constrained by the short-term storage capacity of the auditory system; to recognize that a structure repeats, the input must be stored over the repetition time. We examined the effect of target spacing on subjects' ability to extract the target from a mixture sequence, holding the number of target presentations fixed at six but varying how frequently the targets occurred (Fig. 4). Performance was unaffected by short delays but declined steadily thereafter [$F(4,24) = 22.98$, $P < 0.0001$]. The results are consistent with an integration process that tracks acoustic structure using an auditory memory buffer, although they leave open the question of whether time delays or the intervening acoustic input are driving the effect. Either way, it appears that when the storage capacity of the integration process is exceeded, repetition becomes difficult to track.

Computational Schemes for Extracting Embedded Repetition. It is easy to envision simple computational schemes in which the structure of a repeating source could be extracted from mixtures. As a proof of concept, Fig. 5 illustrates one such approach. A target estimate is initialized to the first segment of the mixture sequence and over time is refined through an averaging process that is time-locked to peaks in the cross-correlation of the target estimate and the spectrogram (*SI Methods*). The correlation peaks reveal the delay at which the signal contains the target, and the averaging (taking the pointwise minimum of the previous target estimate and the current spectrogram segment) combines information across mixtures. Although the estimation process is constrained by the averaging window (*SI Methods*), it does not require knowledge of the target duration, repetition pattern, or other characteristics.

Fig. 5 shows a spectrogram of a sequence of mixtures of a target sound with various others (*A*), followed by spectrograms of a sequence of target estimates derived for this mixture sequence (*B*), graphs showing the cross-correlation between each successive target estimate and the next 700-ms block of the spectrogram (*C*), and a spectrogram of the true target (*D*). The correlation peaks occur at the onset of the target in the mixture, and the estimation process converges on the true target after several iterations (see also *SI Results*, Experiment 6).

Discussion

The recovery of individual sound sources from mixtures of multiple sounds is a central challenge of hearing. Our results suggest one solution: a sound source can be recovered if it occurs more than once and is not always mixed with the same other sounds. This is true even in cases where other grouping cues are impoverished to the point that a single instance of the source is unsegmentable. The auditory system evidently detects repeating spectro-temporal structure embedded in mixtures, and interprets this structure as a sound source. Repetition of sound sources is not explicit in the input to the ear, because the source waveform is generally corrupted at each presentation by other sounds. Source repetition can nonetheless be detected by integrating information over time. Listeners in our experiments were able to form detailed impressions of sound sources that they only ever heard in mixtures, and thus were able to recover this latent structure.

Source repetition can be viewed as another acoustic grouping cue, but it is distinct from other cues in one important respect—its use does not require prior knowledge of sound characteristics. Other grouping cues are rooted in particular properties of natural sounds, be they statistical regularities that hold for broad sets of sounds (e.g., the "bottom-up" cues of common onset or harmonicity) or attributes specific to individual sounds or sound classes (e.g., the "top-down" cues of speech acoustics). Such properties serve as cues because they characterize the particular sorts of sounds found in the world. Knowledge of these sound properties thus must first be internalized by the auditory system from the environment, either over the course of evolution or by learning during an organism's development. Repetition, in contrast, requires only the assumption that sound sources maintain some consistency over time. Our finding that repetition alone can support segregation suggests that it can bootstrap the auditory system in situations where characteristics of sound sources are not yet known, be it early in development or in unfamiliar auditory environments.

The practical utility of this phenomenon for sound segregation obviously depends on the presence of repeating sounds. Not all sounds occur repetitively, but repetition is nonetheless common to natural auditory environments. Examples include the sounds of rhythmic motor behaviors (e.g., walking, running, scratching, clapping) and repetitive physical processes (e.g., branches swaying, water trickling). It is also striking that many animal

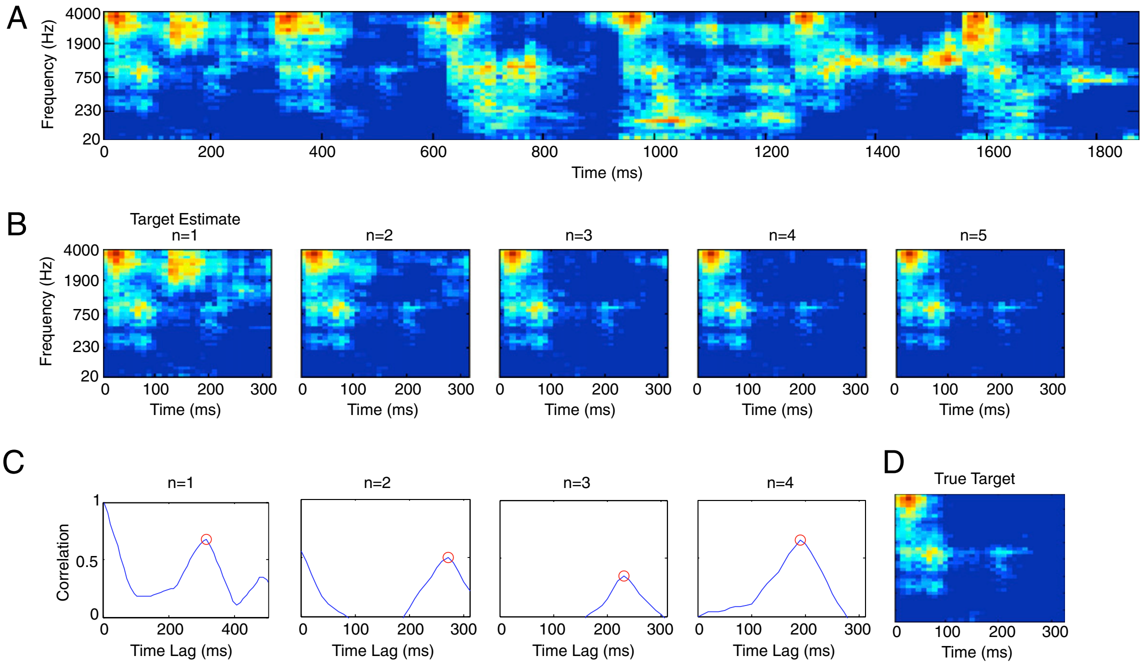

Fig. 5. A candidate computational scheme to extract a repeating target sound from mixtures. (*A*) Spectrogram of a sequence of mixtures of one target sound with various distractors. (*B*) Spectrograms of target sound estimates after each iteration of the algorithm. Only the first 300 ms is shown for ease of comparison with *D*. (*C*) Cross-correlation of target estimate with the next block of the input spectrogram from *A*, as a function of the time shift applied to the spectrogram block. The red circle denotes the peak of the correlation function as found by a peak-picking algorithm. (*D*) Spectrogram of the true target sound. Note the resemblance to the target estimate after five iterations, shown directly above.

vocalizations consist of repetitions of a single call (19), and as such would benefit from repetition-based segregation. Although the targets in our experiments repeated exactly, we found informally that moderate variation in the exemplars had little effect on the ability to hear the target repeating. This is not surprising from a computational standpoint; if the repeating sounds produce a peak in the correlation function, as they will when their variation is not excessive, then an algorithm like that of Fig. 5 will recover their central tendency. "Fingerprinting" techniques for detecting repeating patterns (20) are an alternative model for repetition detection, and these are particularly tolerant of variability. It thus seems likely that source repetition could play an important role in everyday hearing.

The effect of repetition can be viewed as an extension of Bregman's "old-plus-new" idea (1), whereby frequencies added to a spectrum are segregated from those that are continuously present. Our effects involve continuity only at an abstract level, because our stimuli had dynamic spectra and were often separated by short gaps (Figs. 3 and 4). Our results thus implicate a mechanism that can extract dynamic spectrotemporal structure (e.g., as in Fig. 5) distinct from the spectral subtraction mechanisms often posited (1). The upshot of this is that repetition can drive the segregation of complex, quasi-realistic sounds from mixtures.

The effects described in this paper are examples of "streaming" (1, 18, 21, 22), in that the repeating targets segregate from the distractors over time. Perhaps because we presented temporally overlapping sounds, our effects differ in some respects from the well-known case of alternating tones that segregate when repeated. We found that sounds segregated only when one of the sounds varied, not when both were repeated. Our findings bear a closer resemblance to the classic finding that repeating tones are easier to detect when accompanying masker tones vary from one presentation to the next (23–25). Those effects are conceptually similar to ours, but the acoustics are considerably different, as are the conditions under which the effects hold. For instance, the tone effects depend on spectral separation between the target tone and the masker, perhaps relying on spectral separation as a bottom-up segregation cue, and are adversely affected by even brief gaps between tones (25). These differences from our phenomena raise the possibility of distinct mechanisms; the tone effects seem closely related to Bregman's old-plus-new phenomena, and could have a similar explanation. There is also some conceptual similarity between our results and demonstrations that infants and adults can learn repeating patterns in streams of phonemes (26). This latter case seems likely to represent a distinct phenomenon, given that the patterns are acquired over longer time scales and usually are not consciously accessible.

Our study highlights the experimental use of generative models of sound. Studies of the cocktail party problem have traditionally used unnatural synthetic stimuli (9, 27, 28) or familiar real-world sounds such as speech (3, 10, 12, 29). Generative models have the advantage of producing novel stimuli that lack the confounding effects of familiarity but that share properties of natural sounds. The statistics captured by our model are but a small subset of those characterizing the full distribution of natural sounds, but they nonetheless have two important consequences. First, stimuli with naturalistic modulation are sparse in the time-frequency domain, and thus they do not uniformly mask one another (8). Detection of repetition likely requires some degree of sparsity in the sensory input, because otherwise there would be little to gain from hearing sounds in multiple mixtures; most sounds would mask one another over most of their extent. Second, natural statistics allowed the generation of many stimuli that did not all sound the same. Presumably because the auditory system is tuned to the properties of natural sounds (30–33), in this case spectro-temporal modulation (34), naturalistic stimuli are better discriminated than unnatural stimuli (35). Different samples of white noise, for instance, sound much less distinct than do different samples from our model, which likely would make the task of discriminating targets prohibitively difficult.

Consistent with these notions, pilot experiments with alternative correlation functions indicated that the phenomena do not depend sensitively on their exact shape, but that large deviations from natural correlations do render the stimuli less discriminable and less sparse, to the point that the task becomes impossible. For instance, we found that the task could not be performed when the stimuli were different samples of white noise. Although repetition of individual samples of white noise is sometimes noticeable (36, 37), their perceptual similarity and spectrotemporal uniformity apparently precludes this when samples are embedded in mixtures. It thus was important to use a naturalistic sound model. Sparsity is likely crucial to the phenomenon, and the discriminability of natural stimuli facilitated the experimental task.

The utility of source repetition could extend to vision and olfaction, which also confront scene analysis problems. Organisms receive multiple overlapping objects or odors as sensory input, and repetition might enable the recovery of individual objects or odors without prior knowledge of their characteristics. The problems are not analogous in all of their details (e.g., odors are not defined by their temporal structure, and visual objects do not combine linearly when forming an image, due to occlusion; ref. 7), but the same general principle may apply: a particular mixture of sources (objects or odors) is unlikely to occur repeatedly, such that repeating patterns in the input are diagnostic of single sources. Repeating patterns should induce input correlations that could guide temporal integration and reveal single objects or odors, just as we found with sound.

The cocktail party problem has been believed to be solved via the combination of grouping cues derived from statistical regularities of natural sounds, and knowledge of specific sounds or sound classes. Using a simple generative model to produce novel sounds, we found that sound source repetition provides a third source of information with which to parse sound mixtures, one that the auditory system can use even when other segregation cues are unavailable, and which could perhaps be used to learn other grouping cues. The auditory system seems attuned to repetition, and can use it to succeed in conditions that would otherwise be insurmountable.

Materials and Methods

Sound analysis and synthesis used spectrograms specifying the logarithm of the rms amplitude in a set of time-frequency windows. Spectrograms were generated by first passing a signal through an auditory filter bank, then passing each filter output through a set of time windows. The rms level of the windowed signal yielded the value of a spectrogram cell. Adjacent filters and time windows overlapped by 50%.

Correlations between pairs of spectrogram cells were measured for the initial 500-ms segment of each natural sound. These correlations were averaged across pairs of cells with the same time or frequency offset to yield temporal and spectral correlation functions for each stimulus set, as displayed in Fig. 1 *C* and *D*.

Synthetic stimuli with similar correlations were created by modeling the spectrogram as a multivariate Gaussian variable, specified by a mean spectrogram, and a covariance matrix containing the covariance between every pair of spectrogram cells. The mean of each spectrogram cell was set proportional to the corresponding filter bandwidth. The covariance matrix was generated from exponentially decaying correlation functions that approximated the shape of correlation functions for natural sounds. For each pair of cells, the covariance was the product of the corresponding temporal and spectral correlations and a constant variance.

To generate sounds, a time-frequency decomposition was generated for a sample of white noise. The signal in each window was scaled to set its log-amplitude to that of the corresponding cell in a spectrogram sampled from our generating distribution. The results were passed through the filter bank again (as in other analysis and synthesis decompositions; ref. 38) and summed to generate a sound signal. Because adjacent filters and time windows overlapped and thus interfered with each other when amplitudes were altered, the spectrogram of the resulting sound generally differed from the sampled spectrogram from which the sound was generated. However, these differences were subtle, and the intended correlation structure remained present in the sounds, as can be seen in the correlations measured in the synthetic sounds (Fig. 1 *C* and *D*).

Methods are described in more detail in *SI Methods*.

ACKNOWLEDGMENTS. We thank Bart Anderson, Stephen David, Monty Escabi, Heather Read, and Nathan Witthoft for comments on the manuscript, and Andrew Schwartz and Barbara Shinn-Cunningham for useful discussions. Thus work was supported by National Institutes of Health Grant R01 DC 07657.

1. Bregman AS (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).
2. Darwin CJ, Carlyon RP (1995) Auditory grouping. *The Handbook of Perception and Cognition*, ed Moore BCJ (Academic, New York), Vol 6.
3. Bronkhorst AW (2000) The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acustica* 86:117–128.
4. Narayan R, et al. (2007) Cortical interference effects in the cocktail party problem. *Nat Neurosci* 10:1601–1607.
5. Bee MA, Micheyl C (2008) The cocktail party problem: What is it? How can it be solved? And why should animal behaviorists study it? *J Comp Psychol* 122:235–251.
6. Elhilali M, Shamma SA (2008) A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. *J Acoust Soc Am* 124:3751–3771.
7. McDermott JH (2009) The cocktail party problem. *Curr Biol* 19:R1024–R1027.
8. Ellis DPW (2006) Model-based scene analysis. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, eds Wang D, Brown GJ (Wiley, Hoboken, NJ), pp 115–146.
9. Roberts B, Brunstrom JM (1998) Perceptual segregation and pitch shifts of mistuned components in harmonic complexes and in regular inharmonic complexes. *J Acoust Soc Am* 104:2326–2338.
10. de Cheveigne A, Kawahara H, Tsuzaki M, Aikawa K (1997) Concurrent vowel identification, I: Effects of relative amplitude and F0 difference. *J Acoust Soc Am* 101:2839–2847.
11. Darwin CJ, Ciocca V (1992) Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component. *J Acoust Soc Am* 91:3381–3390.
12. Best V, Ozmeral E, Gallun FJ, Sen K, Shinn-Cunningham BG (2005) Spatial unmasking of birdsong in human listeners: Energetic and informational factors. *J Acoust Soc Am* 118:3766–3773.
13. Warren RM (1970) Perceptual restoration of missing speech sounds. *Science* 167:392–393.
14. Voss RF, Clarke J (1975) "1/f noise" in music and speech. *Nature* 258:317–318.
15. Attias H, Schreiner CE (1997) Temporal low-order statistics of natural sounds. *Advances in Neural Information Processing*, eds Mozer M, Jordan M, Petsche T (MIT Press, Cambridge, MA), Vol 9.
16. Singh NC, Theunissen FE (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. *J Acoust Soc Am* 114:3394–3411.
17. MacMillan NA, Creelman CD (1991) *Detection Theory: A User's Guide* (Cambridge Univ Press, New York).
18. Bregman AS, Pinker S (1978) Auditory streaming and the building of timbre. *Can J Psychol* 32:19–31.
19. Wiley RH, Richards DG (1978) Physical constraints on acoustic communication in the atmosphere: Implications for the evolution of animal vocalizations. *Behav Ecol Sociobiol* 3:69–94.
20. Cotton C, Ellis D (2009) Finding similar acoustic events using matching pursuit and locality-sensitive hashing. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York* (Institute of Electrical and Electronics Engineers, New York).
21. Moore BCJ, Gockel H (2002) Factors influencing sequential stream segregation. *Acta Acustica* 88:320–332.
22. Snyder JS, Alain C (2007) Toward a neurophysiological theory of auditory stream segregation. *Psychol Bull* 133:780–799.
23. Kidd G, Jr., Mason CR, Deliwala PS, Woods WS, Colburn HS (1994) Reducing informational masking by sound segregation. *J Acoust Soc Am* 95:3475–3480.
24. Kidd G, Jr., Mason CR, Richards VM (2003) Multiple bursts, multiple looks, and stream coherence in the release from informational masking. *J Acoust Soc Am* 114:2835–2845.
25. Micheyl C, Shamma SA, Oxenham AJ (2007) Hearing out repeating elements in randomly varying multitone sequences: a case of streaming?. *Hearing: From Sensory Processing to Perception*, eds Kollmeier B, et al. (Springer, Berlin).
26. Saffran JR, Aslin RN, Newport EL (1996) Statistical learning by 8-month-old infants. *Science* 274:1926–1928.
27. Carlyon RP (1991) Discriminating between coherent and incoherent frequency modulation of complex tones. *J Acoust Soc Am* 89:329–340.
28. McDermott JH, Oxenham AJ (2008) Spectral completion of partially masked sounds. *Proc Natl Acad Sci USA* 105:5939–5944.
29. Culling JF, Darwin CJ (1993) Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0. *J Acoust Soc Am* 93:3454–3467.
30. Smith EC, Lewicki MS (2006) Efficient auditory coding. *Nature* 439:978–982.
31. Nelken I, Rotman Y, Bar Yosef O (1999) Responses of auditory-cortex neurons to structural features of natural sounds. *Nature* 397:154–157.
32. Garcia-Lazaro JA, Ahmed B, Schnupp JW (2006) Tuning to natural stimulus dynamics in primary auditory cortex. *Curr Biol* 16:264–271.
33. Escabí MA, Miller LM, Read HL, Schreiner CE (2003) Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *J Neurosci* 23:11489–11504.
34. Chi T, Gao Y, Guyton MC, Ru P, Shamma SA (1999) Spectro-temporal modulation transfer functions and speech intelligibility. *J Acoust Soc Am* 106:2719–2732.
35. Woolley SM, Fremouw TE, Hsu A, Theunissen FE (2005) Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nat Neurosci* 8:1371–1379.
36. Kaernbach C (2004) The memory of noise. *Exp Psychol* 51:240–248.
37. Agus TR, Thorpe SJ, Pressnitzer D (2010) Rapid formation of robust auditory memories: Insights from noise. *Neuron* 66:610–618.
38. Crochiere RE, Webber SA, Flanagan JL (1976) Digital coding of speech in sub-bands. *Bell Syst Tech J* 55:1069–1085.

Supporting Information

McDermott et al. 10.1073/pnas.1004765108

SI Results

Experiments 2c and 2d. Our method required using probe sounds that were distinct from the target half of the time. In the single-mixture conditions of Experiments 1, 2a, and 2b, these "incorrect" probes were constrained to be physically consistent with the mixture; they could be no higher in level than the mixture at any point in the spectrogram. In the multiple-mixture conditions, the probes were constrained to be physically consistent with one of the mixtures in the sequence, selected at random. In principle, subjects might have been basing their performance in the multiple-mixture conditions not on perceiving the segregated target sound, but rather by noticing when the probes were physically inconsistent with some of the mixtures on such trials (e.g., by noting that the probe contained frequencies that some of the mixtures did not).

To help exclude this possibility, we repeated Experiments 2a and 2b using incorrect probes that were constrained only to be acoustically similar to the targets. Incorrect probes were generated by fixing a time slice (1/8 of the sound's duration) to be equal to the targets, drawing conditional samples (*SI Materials and Methods*), and keeping only those samples whose spectrogram, expressed in dB (relative to the maximum time-frequency cell) and clipped at –40 dB, had a correlation coefficient of 0.8–0.9 with that of the target sound. Thus, the incorrect probes were no less physically consistent with the single-mixture conditions on average than with the multiple-mixture conditions; in both cases, they could have more energy than the mixtures at certain spectrogram locations. If noticing these inconsistencies was the basis for the subjects' performance, then the single- and multiple-mixture conditions should produce similar results.

We also used distractor sounds that were customized for each target, such that each distractor masked part of the target according to at least one of the criteria used for the distractors in Experiment 6. This was done to rule out the possibility that some of the mixtures might sound sufficiently similar to the target such that subjects could merely match the probe sound to individual mixtures. In all other respects, the methods were similar to those used in Experiments 2a and 2b. Eight of the original 10 subjects participated.

As shown in Fig. S1, we obtained similar results using this alternative method. Performance again improved with the number of different mixtures heard, indicating that subjects were not simply noticing properties of individual mixtures relative to the probe sound. As in Experiments 2a and 2b, we found a main effect of the number of different mixtures [$F(4,28) = 15.0$, $P < 0.0001$], but no effect of experiment type [$F(1,7) = 0.22$, $P = 0.65$] and no interaction [$F(4,28) = 0.95$, $P = 0.45$]. The main difference between the results of Experiments 2a and 2b and Experiments 2c and 2d was the the latter experiments' better performance in the single-mixture conditions. This difference indicates that listeners can achieve greater-than-chance performance with single mixtures by monitoring something akin to physical consistency (e.g., whether the probe sound contains frequencies that the mixture does not), but that the benefit of multiple mixtures exceeds this small effect.

Experiment 5: Temporal Jitter. If mixture variability is indeed the key to recovering a sound source, then it should be possible to enhance performance for a single repeated mixture by varying the time offset between target and distractor. We ran an experiment with the one-, two-, and 10-mixture conditions of Experiment 2b, with the distractor sounds either synchronous with the target sounds (as in Experiment 2b) or jittered randomly in time by up to 120 ms in either direction. As shown in Fig. S2, varying the timing of the distractors relative to the targets improved performance for the one-mixture [$t(9) = 5.34$, $P < 0.0001$] and two-mixture conditions [$t(9) = 3.09$, $P = 0.01$], but not for the 10-mixture condition [$t(9) = 0.87$, $P = 0.4$, paired t test]. This difference produced an interaction between synchrony and mixture number [$F(2,8) = 28.26$, $P < 0.0001$]; there were also significant main effects of both factors, as is apparent from the results graph. Temporal variability thus aids segregation when the sounds in the mixtures do not themselves vary much, but is not of benefit otherwise.

Experiment 6: Grouping Ambiguities vs. Energetic Masking. Many studies have considered sound segregation to be hindered by two distinct factors, commonly termed "energetic" and "informational" masking (1–10). Given that exposure to a sound in multiple distinct mixtures apparently can help an observer overcome both factors, we explored whether this was the case for human listeners.

When multiple sound sources each have energy at approximately the same point in frequency and time, they "energetically" mask each other. The sound higher in energy dominates, and the energy of the other sources at that point is not physically evident (Fig. S3*A*, second row, far right, green-labeled cells). However, even if a target sound is not energetically masked, the presence of another sound source can impair its identification. A mixture of sounds contains acoustic energy scattered over frequency and time, some parts of which belong together and some of which do not (Fig. S3*A*, second row, far right; red-labeled cells belong to the distractor rather than to the target). If this energy is improperly grouped, then the target will be misheard. This effect has come to be known as "informational" masking, because the source of the impairment is not physical spectro-temporal overlap, but rather an ambiguity of grouping (1–10).

Hearing a target sound mixed successively with different distractor sounds could help overcome both types of masking, (energetic masking because features that are physically obscured in one mixture are unlikely to be obscured in the next, and informational masking because the features belonging to a particular sound will tend to occur repeatedly in a fixed configuration, signaling that they belong together). By tracking feature configurations over time, the auditory system could build up a representation of the sound that is robust to both factors. The computational scheme outlined in Fig. 5 provides one example of how this might occur. The distractors occasionally obscure features of the target (energetically masking it). The distractors also tend to have energy in places where the target does not, and in a single mixture it is unclear how the energy should be grouped. The time-locked averaging mechanism proposed in the main text averages out both effects.

To test whether listeners can use multiple mixtures in this way, we first generated a set of customized distractor sounds for each target sound, each of which both energetically and nonenergetically masked the target to a significant extent. We did this by generating many potential distractors and selecting those that had energy in some of the places where the target cell did not and also that exceeded the target sound in amplitude in some of the places where its energy was above a threshold value (*SI Methods*). We then isolated the energetic and nonenergetic components of masking by thresholding the distractor stimuli in the time-frequency domain (11). To eliminate nonenergetic masking, we set the distractors to 0 at spectrogram locations in which the target sound had minimal energy (< −40 dB for the maximum level across cells). The resulting sounds had energy only in places where the target did, and as such could only energetically mask the target (Fig. S3*A*, third row). To minimize energetic masking but preserve nonenergetic masking, we made the complimentary manipulation, setting the distractors for each target sound to 0 in places where the target was above the threshold and the distractor was sufficiently high to have a chance of masking it (*SI Methods*; Fig. S3*A*).

We measured subjects' ability to perceive the target in sequences of mixtures with these three types of distractors. As shown in Fig. S3*B*, for all three distractor types, subjects remained close to chance after hearing a repeating single mixture, but were far above chance when presented with multiple different mixtures, producing a main effect of mixture variability [$F(1,7) = 137.49$, $P < 0.0001$] and no interaction with distractor type [$F(2,14) = 1.95$, $P = 0.18$]. These results indicate 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. This finding is consistent with the computational scheme outlined in the main text, which overcomes energetic and informational masking with the same simple averaging mechanism.

SI Materials and Methods

Subjects. Ten subjects (four females; average age, 26 ± 4 y) participated. All had pure-tone thresholds of 20 dB hearing level or less at octave frequencies between 250 and 8,000 Hz, and none reported any history of hearing disorders. The same subjects were used throughout, but in Experiments 2b, 2c, 3a, 3b, and 6, only 8 of the 10 subjects were available, and in Experiment 4, only 7 of the 10 subjects were available.

Sound Analysis and Synthesis. A set of 39 filters equally spaced on an ERB_N scale (12) spanning 20–4,000 Hz, with half-cosine frequency responses was used for sound analysis and synthesis. The time windows were raised cosines, 20 ms in width.

Because we wanted to synthesize sounds with the properties of individual natural sound sources rather than mixtures of sources, it was important to analyze recordings of isolated sounds. Spectrogram correlations were measured for 350 English words spoken by two speakers, one male and one female, and 30 animal vocalizations taken from sound effects CDs. Each sound clip was edited to remove any silence at the beginning and end. Correlations between pairs of spectrogram cells at either the same frequency or the same time point were measured for the initial 500-ms segment of each natural sound. These correlations were then averaged across pairs of cells with the same offset, yielding temporal correlation functions at each frequency and spectral correlation functions at each time point. The shape of these correlation functions was fairly consistent across frequency and time, as in previous reports (13), so we averaged them to yield single temporal and spectral correlation functions for each stimulus set, as displayed in Fig. 1 *C* and *D*. There were some differences in these functions across the sets of sounds, but all were clearly distinct from the correlations of white noise (Fig. 1 *C* and *D*). We found qualitatively similar correlation functions with alternative sets of sounds, such as excerpts of sentences, or sounds made by inanimate objects (e.g., impact sounds) – correlations generally fell slowly and smoothly with increasing time or frequency offsets, although the rate of decay varied depending on the specific sound set analyzed.

The correlation functions used to generate the covariance matrix of our generating distribution had decay constants of −0.075 per filter and −0.065 per time window. We imposed separable correlations in time and frequency; although there are some deviations from this in natural sound sets (14), these are slight. The mean of each spectrogram cell in the generating distribution was set such that the stimuli would have a flat spectrum on average. This deviated from the average spectra of natural sounds, but it ensured that the high frequencies were audible and not easily masked by simultaneous low frequencies. Onset and offset ramps (10-ms half-Hanning windows) were applied to all synthetic sounds.

Generation of Incorrect Probes. In half of the trials, the probe was different from the target. Our challenge was to generate these "incorrect" probes such that performance would depend primarily on sound segregation rather than on other factors. Simply using another sample from our generating distribution proved to be inadequate, because such a sound often had more energy at some time-frequency location than the mixture of the target and a distractor, and could be judged on this basis. We found it necessary to choose incorrect probes that were both statistically comparable to the target sounds and physically consistent with the mixture in question.

We adopted the following procedure. At a randomly selected time slice (equal to 1/8 of the sound's duration, or 4 of 32 time windows), the incorrect probe was set equal to the mixture (because the target was typically equal to the mixture in some places; see Fig. 1 for an example). A conditional sample was then drawn from the Gaussian generating distribution (15) to yield a new sound with the covariance structure of the target sounds. This sample was then set equal to the mixture at all points in the spectrogram where it exceeded the mixture level, to ensure that the incorrect probe was physically consistent with the mixture. The resulting spectrogram was then rejected if it differed from the mixture by less than an average of 7 dB, to ensure that the incorrect probe was not more similar to the mixture than was the target.

Procedural Details. Sounds were played out by a LynxStudio Lynx22 24-bit D/A converter at a sampling rate of 48 kHz, and were presented diotically over Sennheiser HD580 headphones at a sound pressure level of 72 dB. Incorrect probes were scaled by the same factor as the corresponding target so as to remain physically consistent with the mixture.

Subjects were instructed to use all four responses approximately equally often. In all experiments, subjects completed two blocks containing 20 trials per condition.

From pilot versions of the experiments, it became apparent that hearing the target sound was essentially impossible in conditions with a single mixture. To help maintain motivation, feedback was given in only 75% of all trials in all conditions. Pilot versions that eliminated feedback on all trials or provided it on all trials yielded similar results, so this choice appears to not have been critical.

Trial Structure. Each trial was initiated by pressing a key. In Experiment 1, subjects were presented with a mixture followed by a probe sound (conditions 1 and 2), a probe sound followed by a mixture (conditions 3 and 4), a target sound followed by a probe sound (condition 5), or a mixture followed by another mixture (condition 6). In conditions 1–4, the task was to judge whether the probe sound was one of the sounds in the mixture. In conditions 5 and 6, the task was to judge whether the two sounds were the same or different. In Experiments 2–6, subjects were presented with mixture(s) followed by a probe sound. The task was to judge whether the probe sound was one of the sounds in the mixture(s).

Experiment Structure. In Experiments 1 and 2a, trials for a condition were grouped together because stimulus timing and/or tasks differed across condition; conditions were completed in opposite order in the two blocks, to reduce order effects. In all other experiments, trials were ordered randomly. In Experiment 3a, conditions 1 and 2 were run in separate sessions from conditions 3, 4, and 5. Subjects began by completing a full-length practice session (20 trials per condition) of Experiment 1. Before starting Experiment 2a, subjects also completed a full-length practice session of that experiment.

Experiment 3b: Time-Reversed Targets. Condition 2 used time-reversed versions of the target as the incorrect probes; the task was as in the other experiments. To make this task feasible, we used target sounds that were selected to be asymmetric in time; those included had to have spectrograms with a correlation of <0.2 with their time reversal. We also used these sounds in conditions 1 and 3 of this experiment. Incorrect probes for conditions 1 and 3 were generated as in the other experiments.

Experiment 6: Energetic and Informational Masking. Target sounds were generated by the same process as used in the other experiments, but were rejected if 75% of the cells were not within 40 dB of the maximum spectrogram cell. This was done to facilitate the generation of distractor sounds that energetically masked the targets. Distractor sounds were generated separately for each target and were selected to produce a criterion amount of masking. To be included as a distractor, a sound had to produce a mixture that met the following two conditions in at least 25% of the spectrogram cells: (*i*) the mixture exceeded the target by at least 5 dB and the target was no more than 40 dB below the maximum level across the windows of that target, and (*ii*) the mixture was no more than 40 dB below its maximum level and the target was at least 40 dB below its maximum level. The first condition produced distractors that energetically masked the target. The second condition produced distractors that "informationally" masked the target, because they contained energy where the target did not. These distractors were then thresholded in the time-frequency domain as described in the text. Incorrect probes were generated for each type of distractor using the procedure described above.

The criteria for zeroing a cell in the distractors that minimized energetic masking were that the target energy be no more than 40 dB below its maximum and that the distractor energy be no more than 10 dB below that of the target. These criteria of physical overlap neglect masking over time and between adjacent frequency bins, and thus the resulting distractors surely produced some residual energetic masking. However, they generated far less of it than did the unthresholded distractors, while preserving nonenergetic masking of the target.

Target Estimation Model. The spectrogram of the acoustic input (the mixture sequence) was divided into 700-ms blocks, with 50% overlap between adjacent blocks. The target was estimated with the following series of steps:

(*i*) The target estimate was initialized to the first block.

(*ii*) The cross-correlation of the target estimate with the current block was computed for different time delays.

(*iii*) A peak-picking algorithm (http://billauer.co.il/peakdet.html, with the delta parameter set to 0.05) was used to identify the first large peak in the correlation function (which should indicate the position of the next target occurrence).

(*iv*) The target estimate was updated with the current spectrogram block. The updating process involved taking the pointwise minimum of the target estimate and the current spectrogram block, with the spectrogram block time-shifted by the delay of the peak. The minimum was used because mixing two sounds generally serves to increase the spectrogram energy over that present in either sound alone, such that the target sound is likely to never be more than the minimum of two mixtures containing it (16).

(*v*) Steps *ii*–*iv* were repeated with the next block of the spectrogram.

The block size and overlap constrain the duration of the targets that can be detected. Specifically, to produce a peak in the cross-correlation function, a target must fall within the block. To ensure that targets are not "missed," the amount by which blocks overlap must exceed the target length, so that if a target falls on the boundary of a block, then the next block is guaranteed to contain it. In our simulations, we chose the block size to roughly match the analysis window suggested by the results of Experiment 4. We arbitrarily set the overlap to 50%, to ensure detection of the 300-ms experimental stimuli. The overlap could be easily extended to permit the detection of longer-duration targets.

The algorithm is reasonably robust. Targets that overlap the block boundary are not erroneously averaged, because they do not produce a correlation peak; the peak-picking algorithm detects only peaks with lower values on either side. The algorithm uses only the first peak in the correlation function for a block, such that if multiple examples of the target fall within an analysis block, only the first one triggers the averaging process, and the rest are left for the next block. If a particular target exemplar falls within two successive blocks, there is no effect of it being counted twice, because the pointwise minimum operation does not change the target estimate in this case.

Nonetheless, the scheme is clearly oversimplified. For instance, listeners can sometimes extract a target source from mixtures in the presence of other repeating sounds (e.g., Experiment 3a, condition 3), indicating that multiple templates may be used simultaneously. The algorithm that we implemented also does not address what should be done in the event that a peak is not detected in an analysis block, as when the target spacing exceeds the block length, conditions under which human perception suffers (Experiment 4). Moreover, the algorithm works only to the extent that the correlation peaks identified correspond to the target position in the signal. If a peak corresponding to something other than the target onset is chosen (as can sometimes occur if random variation in the sound structure produces a peak), then errors can be introduced in the target estimate. Some of these errors simply reflect suboptimal peak-picking. It is likely that the brain has more robust algorithms than we do, and we would not expect our model to match the performance of human listeners. However, it is also notable that human subjects do not perform at ceiling in our task, and that targets are easier to hear in some mixture sequences than in others. It would be interesting to explore whether any of this variability could be explained by variation in the model's performance due to the clarity of correlation peaks in different mixture sequences. That said, the model is intended mainly as a proof of concept that latent repeating structure could be extracted with a relatively simple, bottom-up mechanism. We make no claims that it is near optimal, or that it can match human performance.

1. Watson CS (1987) Uncertainty, informational masking and the capacity of immediate auditory memory. *Auditory Processing of Complex Sounds*, eds Yost WA, Watson CS (Erlbaum, Hillsdale, NJ), pp 267–277.
2. Leek MR, Brown ME, Dorman MF (1991) Informational masking and auditory attention. *Percept Psychophys* 50:205–214.
3. Lutfi RA (1992) Informational processing of complex sound, III: Interference. *J Acoust Soc Am* 91:3391–3401.
4. Neff DL (1995) Signal properties that reduce masking by simultaneous, random-frequency maskers. *J Acoust Soc Am* 98:1909–1920.
5. Wright BA, Saberi K (1999) Strategies used to detect auditory signals in small sets of random maskers. *J Acoust Soc Am* 105:1765–1775.
6. Brungart DS (2001) Informational and energetic masking effects in the perception of two simultaneous talkers. *J Acoust Soc Am* 109:1101–1109.
7. Freyman RL, Balakrishnan U, Helfer KS (2001) Spatial release from informational masking in speech recognition. *J Acoust Soc Am* 109:2112–2122.
8. Arbogast TL, Mason CR, Kidd G Jr. (2002) The effect of spatial separation on informational and energetic masking of speech. *J Acoust Soc Am* 112:2086–2098.

9. Richards VM, Tang Z, Kidd GD Jr. (2002) Informational masking with small set sizes. *J Acoust Soc Am* 111:1359–1366.
10. Durlach NI, et al. (2003) Note on informational masking. *J Acoust Soc Am* 113: 2984–2987.
11. Brungart DS, Chang PS, Simpson BD, Wang D (2006) Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *J Acoust Soc Am* 120:4007–4018.
12. Glasberg BR, Moore BCJ (1990) Derivation of auditory filter shapes from notched-noise data. *Hear Res* 47:103–138.
13. Attias H, Schreiner CE (1997) Temporal low-order statistics of natural sounds. *Advances in Neural Information Processing*, eds Mozer M, Jordan M, Petsche T (MIT Press, Cambridge, MA), Vol 9.
14. Singh NC, Theunissen FE (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. *J Acoust Soc Am* 114:3394–3411.
15. MacKay DJC (1998) Introduction to Gaussian processes. *Neural Networks and Machine Learning*, ed Bishop CM (Springer, Berlin) Vol Vol 168, NATO ASI Series.
16. Ellis DPW (2006) Model-based scene analysis. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, eds Wang D, Brown GJ (Wiley, Hoboken, NJ), pp 115–146.

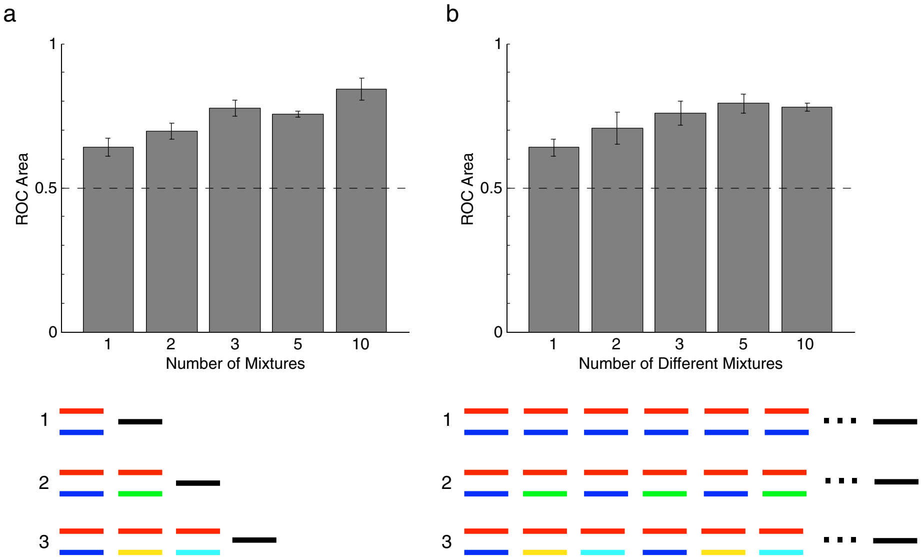

Fig. S1. Results and stimulus configurations for Experiments 2c and 2d. Schematics for conditions with 5 and 10 mixtures are omitted. (*A*) Different numbers of mixtures were presented. (*B*) Ten mixtures were presented in all conditions, and the number of different mixtures was varied.

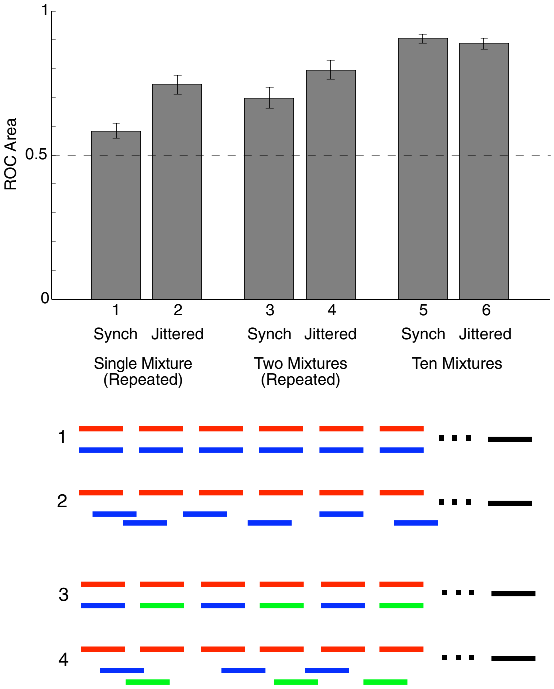

Fig. S2. Stimulus configurations and results of Experiment 5, on the effect of temporal jitter.

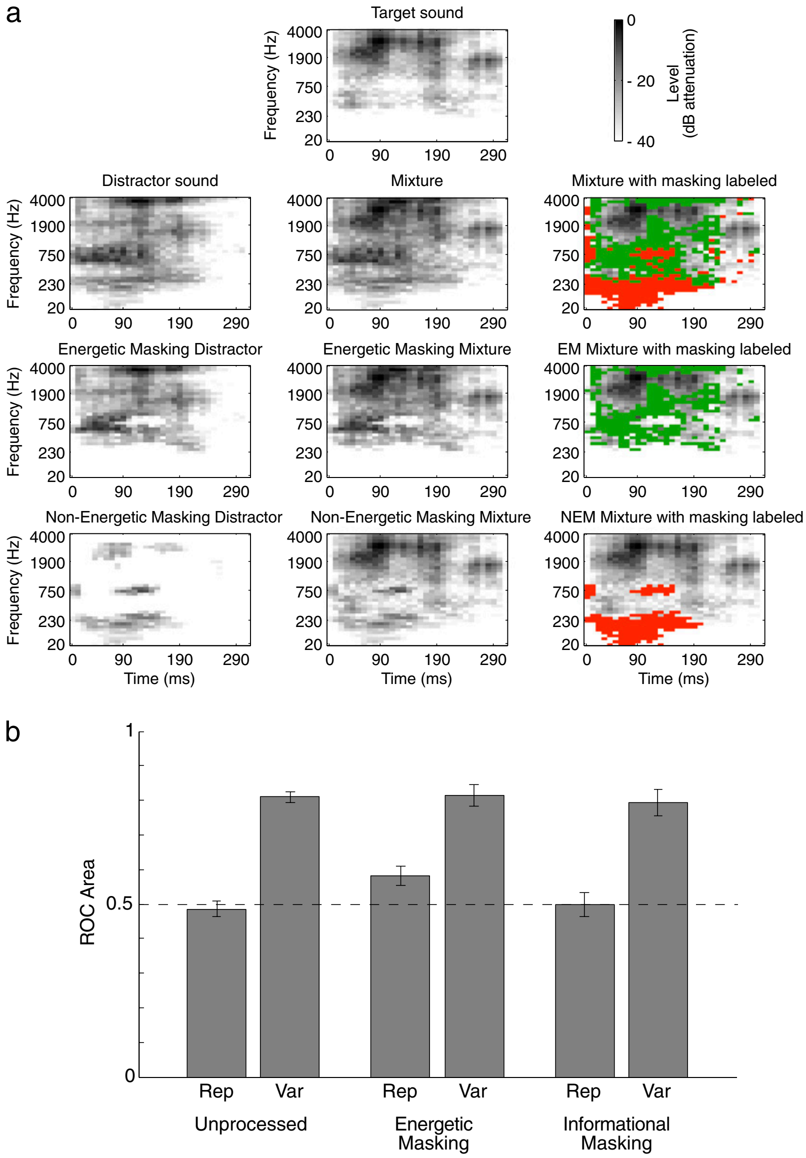

Fig. S3. Source repetition and masking (Experiment 6). (*A*) (*Upper*) Spectrogram showing an example target sound. (*Lower*) Three versions of an example distractor sound (*Left*), each distractor mixed with the target (*Center*), and each mixture with the energetically masked (green) and informationally masking (red) cells labeled (*Right*). (*B*) Results of Experiment 6. Each distractor type was featured in repeating and variable distractor conditions.