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.

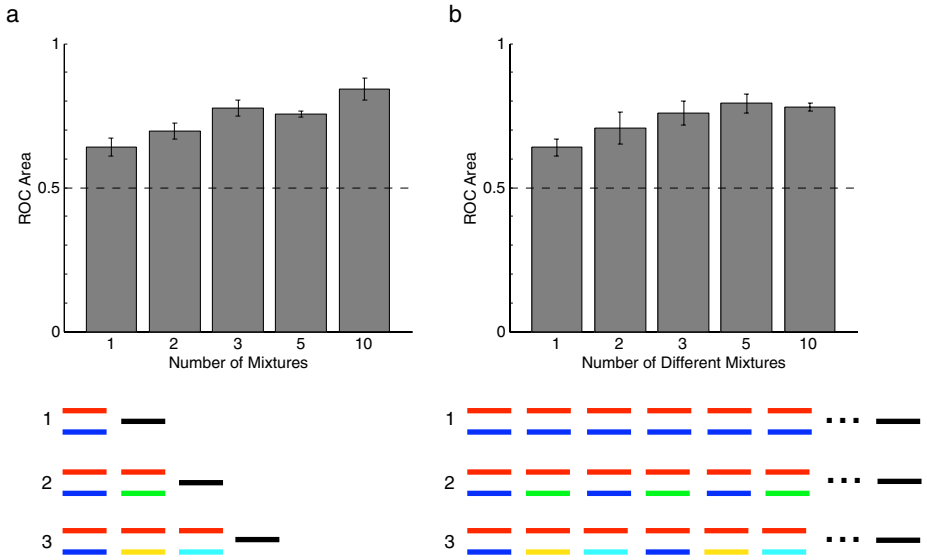

Fig. S1. Results and stimulus configurations for Experiments 2c and 2d. Schematics for conditions with 5 and 10 mixtures are omitted. (*A*) Different numbers of mixtures were presented. (*B*) Ten mixtures were presented in all conditions, and the number of different mixtures was varied.

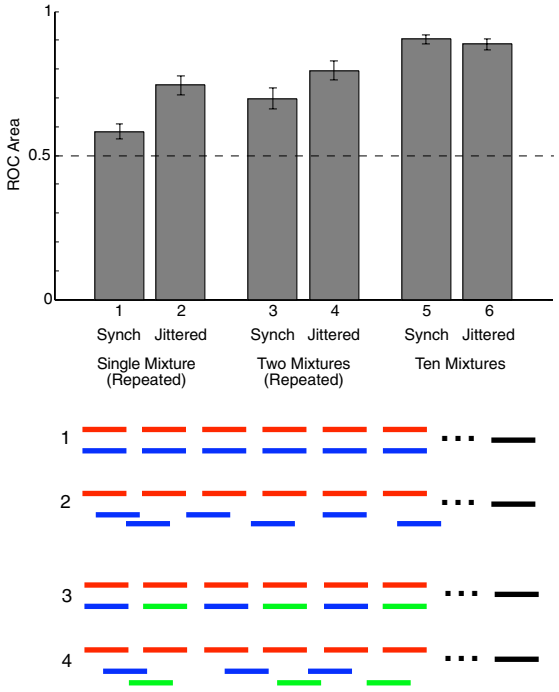

Fig. S2. Stimulus configurations and results of Experiment 5, on the effect of temporal jitter.

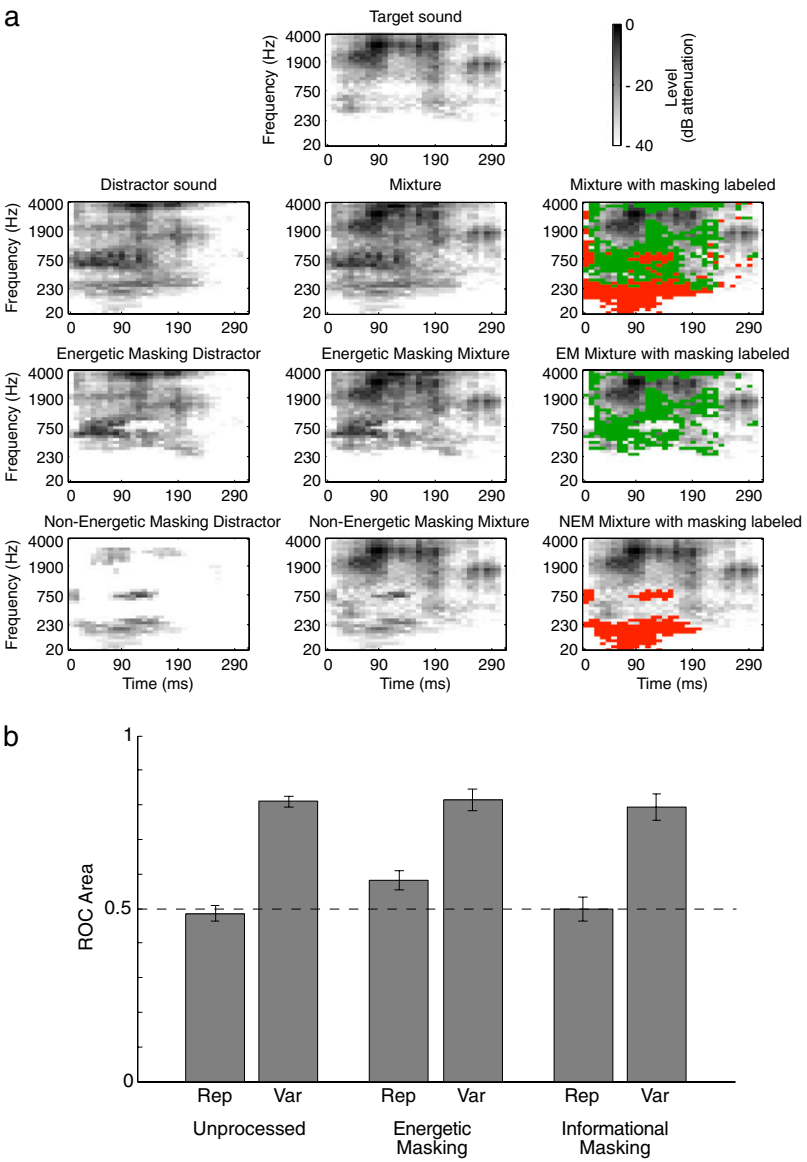

Fig. S3. Source repetition and masking (Experiment 6). (*A*) (*Upper*) Spectrogram showing an example target sound. (*Lower*) Three versions of an example distractor sound (*Left*), each distractor mixed with the target (*Center*), and each mixture with the energetically masked (green) and informationally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