



Ecological origins of perceptual grouping principles in the auditory system

Wiktor Młynarski^{a,b,1} and Josh H. McDermott^{a,b,c,d,1}

^aDepartment of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; ^bCenter for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139; ^cMcGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; and ^dProgram in Speech and Hearing Biosciences and Technology, Harvard University, Boston, MA 02115

Edited by Wilson S. Geisler, The University of Texas at Austin, Austin, TX, and approved October 16, 2019 (received for review March 5, 2019)

Events and objects in the world must be inferred from sensory signals to support behavior. Because sensory measurements are temporally and spatially local, the estimation of an object or event can be viewed as the grouping of these measurements into representations of their common causes. Perceptual grouping is believed to reflect internalized regularities of the natural environment, yet grouping cues have traditionally been identified using informal observation and investigated using artificial stimuli. The relationship of grouping to natural signal statistics has thus remained unclear, and additional or alternative cues remain possible. Here, we develop a general methodology for relating grouping to natural sensory signals and apply it to derive auditory grouping cues from natural sounds. We first learned local spectrotemporal features from natural sounds and measured their co-occurrence statistics. We then learned a small set of stimulus properties that could predict the measured feature co-occurrences. The resulting cues included established grouping cues, such as harmonic frequency relationships and temporal coincidence, but also revealed previously unappreciated grouping principles. Human perceptual grouping was predicted by natural feature co-occurrence, with humans relying on the derived grouping cues in proportion to their informativity about co-occurrence in natural sounds. The results suggest that auditory grouping is adapted to natural stimulus statistics, show how these statistics can reveal previously unappreciated grouping phenomena, and provide a framework for studying grouping in natural signals.

cocktail party problem | natural sound statistics | source separation

Sensory receptors sample the world with local measurements, integrating energy over small regions of time and space. Because the objects and events on which we must base behavior are temporally and spatially extended, their inference can be viewed as the process of grouping these measurements to form representations of their underlying causes in the world. Grouping has been viewed as a fundamental function of the nervous system since the dawn of perceptual science (1, 2).

Grouping mechanisms are presumed to embody the probability that sets of sensory measurements are produced by a common cause in the world (3–5). Yet, dating back to the Gestalt psychologists, grouping has most often been studied using artificial stimuli composed of discrete elements (6, 7)—arrays of dots or line segments in vision or frequencies in sound. One challenge in relating such research to the real world is that it is often difficult to describe natural images and sounds in terms of discrete elements. As a result, grouping phenomena have been related to natural stimulus statistics in only a handful of cases where human observers have been used to label local image features (8–12). Grouping research has otherwise been limited to testing intuitively plausible grouping principles that can be instantiated in hand-designed artificial stimuli.

Grouping is critical in audition, where it is believed to help solve the “cocktail party problem”—the problem of segregating a sound source of interest from concurrent sounds (7, 13–15) (Fig. 1). As in other sensory systems, auditory grouping is

believed to exploit acoustic regularities of natural stimuli, such as the tendency of frequencies to be harmonically related (16–19) or to share a common onset (20–24). However, because acoustic grouping cues have traditionally been identified using informal observation and investigated using simple synthetic stimuli, much remains unknown. First, the extent to which known principles of perceptual grouping are related to natural stimulus statistics is unclear. Second, because the science of grouping has thus far been largely driven by human intuition, additional or alternative grouping principles remain a possibility.

We sought to link auditory grouping principles to the structure of natural sounds by measuring feature co-occurrences in natural signals and assessing their relation to perception. Our approach is distinguished from that of prior work in being independent of prior hypotheses about the underlying features or regularities that might relate to grouping. We first derived a set of primitive auditory patterns by learning a dictionary of spectrotemporal features from a corpus of natural sounds (recordings of speech and musical instruments) using sparse convolutional coding (25, 26). We then measured co-occurrence statistics for these features in the natural sound corpus. We found that superpositions of naturally co-occurring features were more likely to be heard as a single source than pairs of features that do not commonly co-occur, indicating that the auditory system has internalized the

Significance

Events and objects must be inferred from sensory signals. Because sensory measurements are temporally and spatially local, the estimation of an object or event can be viewed as the grouping of these measurements into representations of their common causes. Perceptual grouping is believed to reflect internalized regularities of the natural world, yet grouping cues have traditionally been identified using informal observation. Here, we derive auditory grouping cues by measuring and summarizing statistics of natural sound features. Feature co-occurrence statistics reproduced established cues but also, revealed previously unappreciated grouping principles. The results suggest that auditory grouping is adapted to natural stimulus statistics, show how these statistics can reveal previously unappreciated grouping phenomena, and provide a framework for studying grouping in natural signals.

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

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Published under the PNAS license.

Data deposition: Examples of spectrotemporal features and stimuli used in all experiments have been deposited at <http://mcdermottlab.mit.edu/grouping-statistics/index.html>.

¹To whom correspondence may be addressed. Email: mlynar@mit.edu or jhm@mit.edu.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1903887116/-DCSupplemental>.

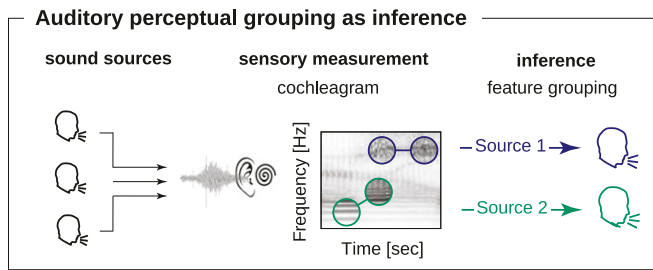


Fig. 1. Auditory perceptual grouping. Multiple sources in the world generate signals that sum at the ear. Local sensory measurements must then be grouped to form source inferences.

co-occurrence statistics over evolution or development. We next developed a method to summarize the observed co-occurrence statistics with a set of cues. The cues defined stimulus properties predictive of whether features were likely to co-occur. To facilitate their interpretation, the cues were instantiated as linear templates. The learned templates captured traditional grouping cues, such as harmonicity and common onset, but also revealed grouping principles not typically noted in the auditory grouping literature. Our results suggest that auditory grouping cues are adapted to natural stimulus statistics and that considering these statistics can reveal previously unappreciated aspects of grouping.

Results

In order to study grouping in natural sound signals without relying on a prior hypothesis of the features or principles that would be involved, we used convolutional sparse coding (25, 26) to first learn a set of features from which natural sounds can be composed. These features were learned from recordings of single sources represented as “cochleagrams”—time–frequency decompositions intended to approximate the representation of sound in the human cochlea. We conceive of distinct sound sources as generated by distinct physical processes in the world and formed a corpus of single sources from recordings of speech or individual musical instruments. Speech and instruments clearly do not exhaust the space of natural sounds, but they are the primary sound classes for which single-source recordings are available in large quantities, as is critical for the stable measurement of the statistical properties that we study here. Speech and instruments also utilize a fairly wide range of physical sound-producing processes (rigid objects excited in different ways, aerodynamic events, periodic and aperiodic energy, etc.), and therefore, it seemed plausible that they might exhibit most of the statistical properties relevant to grouping.

The spectrotemporal features were optimized to reconstruct the training corpus given the constraints of nonnegativity (on both feature kernels and their coefficients) and sparsity (on the coefficients). These constraints produce features that can be thought of as “parts” of the cochleagram, similar to nonnegative representations of natural images (27). The learned features capture simple and local time–frequency patterns, including single frequencies, sets of harmonic frequencies, clicks, and noise bursts (Fig. 2A), loosely analogous to the spectrotemporal features that might be detected in early stages of the auditory system (28). The features reconstructed the training corpus relatively well (Fig. 2B), and they did so significantly better than 3 alternative, nonlearned feature sets (Fig. 2C) (significantly different by *t* tests, $t > 100$, $P < 0.001$ in all cases). We note that features could also be obtained via alternative methods (for instance, via optimization for tasks), which could yield distinct features (29–31). Examples of spectrotemporal features and stimuli used in all experiments can be found on the accompanying webpage (32).

Each feature can itself be viewed as an initial elementary stage of grouping sound energy likely to be due to a single source. However, because natural sound signals are represented with many such features (as a set of time-varying, sparsely activated coefficients) (Fig. 2B), these features must in turn be grouped in order to estimate sound sources from the feature representation.

Feature Co-occurrence Statistics in Natural Sounds. After a signal is decomposed into a feature representation, the problem of grouping thus consists of determining which features are activated by the same source—an inherently ill-posed inference problem (Fig. 1). We measured co-occurrence statistics that should constrain this inference. In principle, the inference of sources from feature activations could be constrained by the full joint distribution of all features. In practice, this distribution is challenging to learn and to analyze (26). Instead, we measured dependencies between pairs of features, which are tractable to measure and analyze and which we found to contain rich structure. The key idea was to compare the co-occurrence of features within the same source with the co-occurrence of features in different sources on the grounds that feature activations should be grouped together if they co-occur in a particular configuration substantially more often in the same source than otherwise.

To measure co-occurrences for features in the same source, we took encodings of large corpora of single sources—speech and instrument sounds—and for each feature *f* (Fig. 3A), computed the average activations of all other features at each of a

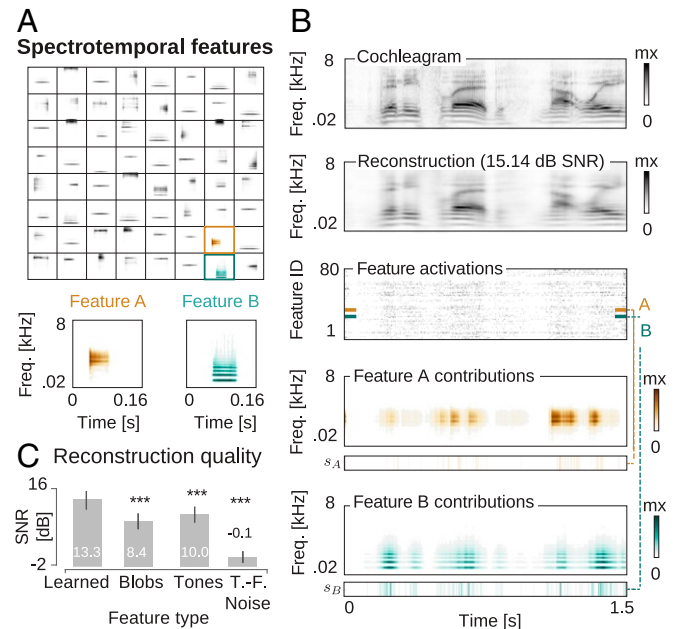


Fig. 2. Spectrotemporal feature decompositions of natural sounds. (A) Spectrotemporal features optimized to reconstruct the corpus of speech and instrument sounds. Two example features are shown at higher resolution. (B) Example speech excerpt (row 1) and its reconstruction (row 2) from time-varying feature coefficients (row 3). The contribution of the 2 example features from A to the reconstruction are shown in rows 4 and 5. The cochleagrams show the convolution of the feature kernel with its time-varying coefficient (shown below each cochleagram). SNR, signal-to-noise ratio. (C) Reconstruction quality of the natural sound corpus. Features learned from natural sound statistics (bar 1) represent cochleagrams with more accuracy than nonlearned features (bars 2 to 4). Error bars plot standard deviation. Asterisks denote statistical significance of *t* tests (vs. learned). *** $P < 0.001$. T-F, time–frequency.

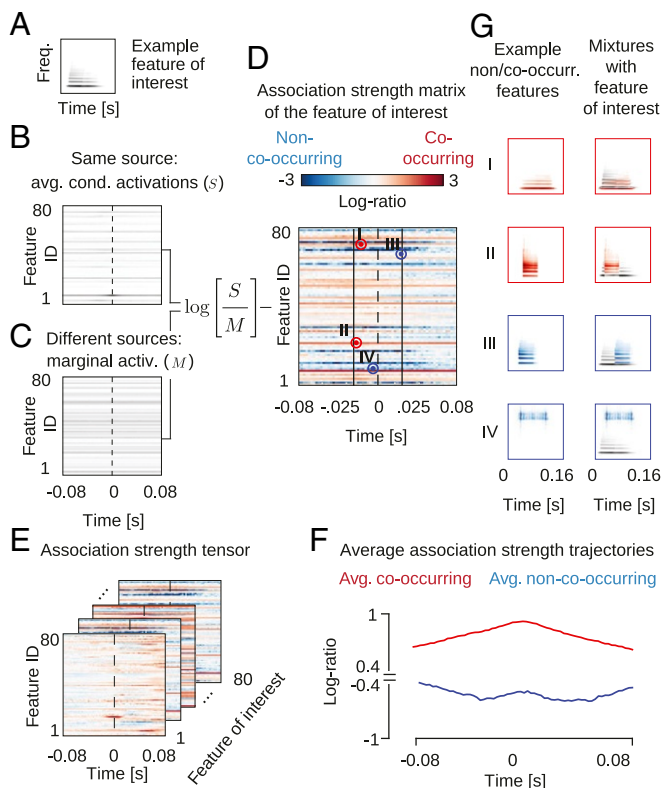


Fig. 3. Co-occurrence statistics of spectrotemporal features. (A) Example feature of interest. (B) Average conditional activations of other features conditioned on the example feature of interest exceeding an activation threshold. (C) Average marginal activations of other features (averaged over time and across the corpus). These are by definition constant over time. (D) Coactivation matrix for the feature of interest formed by the logarithm of the ratio of the mean conditional and marginal activations of the other features. (E) Coactivation tensor formed from the coactivation matrices of all features. (F) Positive and negative tensor entries averaged across features. The strength of association between features decreases with their time offset as expected. (G) Examples of features with high and low association strength with the feature from A. (Left) Features colored red and blue have high and low association strength, respectively, with the example feature of interest from A. (Right) Mixtures of the selected features with the example feature of interest from A.

set of time offsets, conditioned on the activation of the feature f being high (exceeding the 95th percentile of its distribution of activations) (Fig. 3B). This coactivation measure is high for features that tend to be activated at a particular time offset when the selected feature f is activated. To measure co-occurrence for features in distinct sources, we assumed distinct sound sources in the world to be independent (i.e., that the joint distribution of the 2 source signals is equal to the product of their marginal distributions). This assumption is not always correct, as when 2 speakers are conversing and what one person says is in response to another or when 2 musical instruments are played in coordination. However, the independence assumption nonetheless seems likely to approximate what holds much of the time. Given that assumption, the distribution of activations of one feature conditioned on the activation of another can be approximated by its marginal distribution (Fig. 3C). Thus, as a summary measure of the co-occurrence of one feature and another, we computed the ratio of the mean conditional activation of the feature to its mean marginal activation (the mean feature activation averaged over time and across the entire training corpus). Dividing by the mean marginal activation can also be viewed as a normalization step that prevents the resulting

measure from being dominated by how often a feature occurs in the training corpus. In all subsequent analyses, we display the logarithm of this ratio, which we term the “association strength.” We consider a feature as co-occurring or not with the selected feature depending on whether the association strength is positive or negative.

This analysis yielded a matrix for each feature (containing its association strength with each other feature at each of a range of time offsets) (Fig. 3D) and thus, a 3-dimensional tensor for the entire dictionary (Fig. 3E). These matrices are not obviously structured when inspected visually, apart from containing dependencies that on average grow weaker as the time offset between features increases (Fig. 3F). However, the tensor can be used to draw pairs of features that are strongly coactivated in the training corpus or not, and these exhibit intuitively sensible relationships. The examples in Fig. 3G for the harmonic feature shown in Fig. 3A reveal that other features that strongly co-occur with it share a common fundamental frequency (f_0) and fall in the same general frequency range. Conversely, features that are unlikely to co-occur with the example harmonic feature are those that are misaligned in f_0 or that are far apart in frequency. These examples suggest that the co-occurrence statistics can capture reasonable relations between features, but it was not obvious to what extent the full co-occurrence tensor would have been internalized by human listeners and to what extent it would contain comprehensible structure.

Perceptual Grouping Reflects Co-occurrence Statistics. To test whether human listeners have internalized the measured co-occurrence statistics, we conducted a psychophysical experiment with stimuli generated by superimposing sets of features (experiment 1). On each trial, participants heard 2 such stimuli and judged which of them contained 2 sound sources (Fig. 4A). One feature pair was selected from the feature pairs with an association strength in the top 1% of all co-occurring pairs, and the other was from the feature pairs with an association strength in the lowest 5% of the non-co-occurring pairs (Fig. 4B) (i.e., the most negative; the inclusion thresholds were asymmetric because the distribution of associations strengths was asymmetric about 0). To set a ceiling level on task performance, in a second condition, one stimulus was an excerpt of a single speech or instrument sound, while the other was a mixture of 2 such excerpts. Because natural sounds contain a superset of the dependencies measured in the co-occurrence tensor, performance on this condition should provide an upper limit on performance for the task with feature superpositions.

Human listeners reliably identified unlikely combinations of features as sounds consisting of 2 sources (Fig. 4B, left bar) [$t(14) = 10.95$, $P < 0.001$], only slightly below the level for speech mixtures (Fig. 4B, right bar) [$t(14) = 93.75$, $P < 0.001$]. This result suggests that humans have internalized aspects of the co-occurrence tensor and use the learned statistics for perceptual grouping.

To assess the extent to which the perceptual sensitivity was specific to natural co-occurrence statistics, we ran a control experiment using stimuli derived from a co-occurrence tensor measured from synthetic sound textures (33) (experiment 2). The textures were synthesized to match the power spectrum and modulation spectrum of speech but were otherwise unstructured (*Materials and Methods*). Co-occurrence statistics were measured using the same features learned from the natural sound corpus. The experiment thus controlled for the possibility that the features and their encoding process might themselves create dependencies that could support task performance, independent of natural signal statistics. In contrast to stimuli from the natural co-occurrence tensor, the control stimuli produced near-chance performance (Fig. 4C, right bar) [not

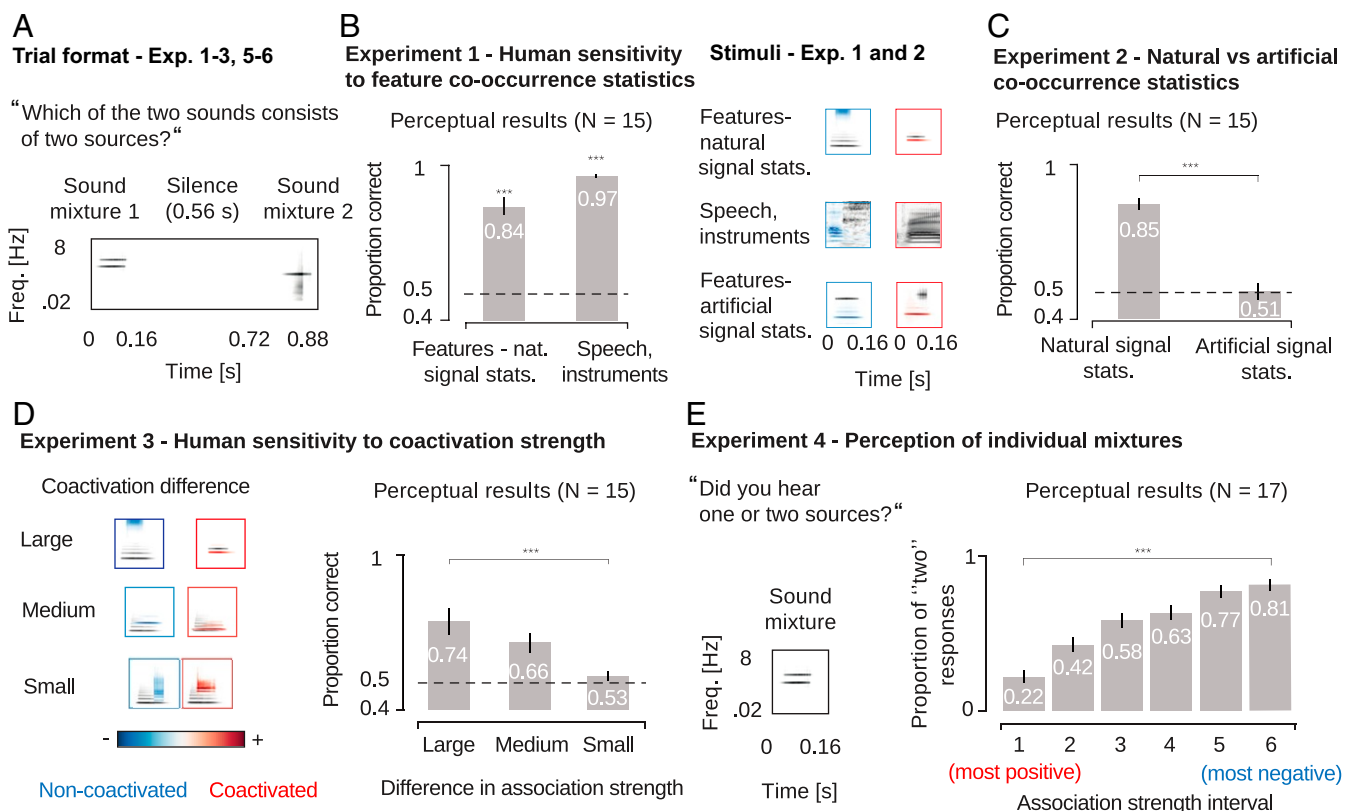


Fig. 4. Perceptual sensitivity to natural feature co-occurrence statistics. (A) Stimulus from an example trial. Listeners heard 2 feature pairs and judged which consisted of 2 sources. (B) Conditions and results of experiment 1. Listeners discriminated 1) feature pairs assembled using natural co-occurrence statistics or 2) mixtures from single excerpts of speech and/or instruments. Asterisks denote statistical significance of *t* tests (vs. chance). $***P < 0.001$. Here and in C–E, error bars plot SEM. (C) Results of experiment 2. Listeners discriminated feature pairs assembled using 1) natural co-occurrence statistics or 2) co-occurrence statistics measured from artificial sound textures. The textures were synthesized to match some of the statistics of speech (related to power and modulation spectra). Asterisks denote statistical significance of *t* tests (vs. chance or between conditions). $***P < 0.001$. (D) Conditions and results of experiment 3. Listeners discriminated feature pairs drawn from different ranges of the coactivation continuum, producing large, medium, or small coactivation differences between the 2 pairs presented on a trial. Asterisks denote statistical significance of repeated measures ANOVA comparing performance in the 3 conditions ($***P < 0.001$). (E) Example trial and results of experiment 4. On each trial, listeners heard a feature pair and judged whether it consisted of 1 or 2 sources (Left). Conditions corresponded to association strength intervals defined for experiment 3. Asterisks denote statistical significance of repeated measures ANOVA comparing performance in the 6 conditions ($***P < 0.001$).

significantly different from chance, $t(14) = 0.1748$, $P = 0.86$ and significantly worse than the natural stimuli, $t(14) = 10.57$, $P < 0.001$. The results suggest that grouping judgments depend on internalized statistics that are to some extent specific to natural sounds.

To further probe the extent to which perceptual grouping judgments would reflect natural co-occurrence statistics, we generated pairs of feature pairs with association strength differences that fell into 1 of 3 ranges (experiment 3; each range differed from that used in experiment 1) (Fig. 4D). If listeners have internalized natural feature co-occurrences, performance should scale with the association strength difference. As shown in Fig. 4D, performance was best when the association strength difference was large and declined as it decreased, yielding a main effect of the association strength difference [$F(1.39, 19.45) = 17.46$, $P < 0.001$]. This result is further consistent with the role of natural co-occurrence statistics in perceptual grouping judgments.

To test whether the association strength would correctly predict whether individual stimuli were heard as 1 or 2 sources, we conducted an additional experiment in which individual stimuli were judged to be 1 or 2 sources (experiment 4). The stimuli were superpositions of pairs of features with association strength that was drawn from bins ranging from negative to positive val-

ues. As shown in Fig. 4E, the tendency to hear a stimulus as a single sound was high for feature combinations with positive association strengths and low for features with negative association strength [$F(5, 80) = 82.35$, $P < 0.001$]. The relation between the empirical pairwise association of features and their perception as a single source provides further evidence for the role of natural co-occurrence statistics in perceptual grouping.

Predicting Grouping Cue Strength from Natural Statistics. Grouping is typically conceptualized in terms of cues—stimulus properties that are predictive of grouping and that could thus help to solve the inference problem at the heart of grouping. We sought to relate grouping cues to co-occurrence statistics both to evaluate the statistical validity of traditionally proposed cues and to learn cues *de novo* from statistics. We formalized a grouping cue to be a function of 2 stimulus features whose value depended on whether the 2 features are likely to belong to the same source or not (Fig. 5A). We quantified the statistical strength of a cue using the co-occurrence tensor, measuring the cue for all pairs of strongly positively associated features and all pairs of strongly negatively associated features and then quantifying the difference in the distributions of cue values for the 2 sets (Fig. 5A).

We first considered the 2 most commonly cited cues from traditional accounts of auditory grouping: common onset and offset

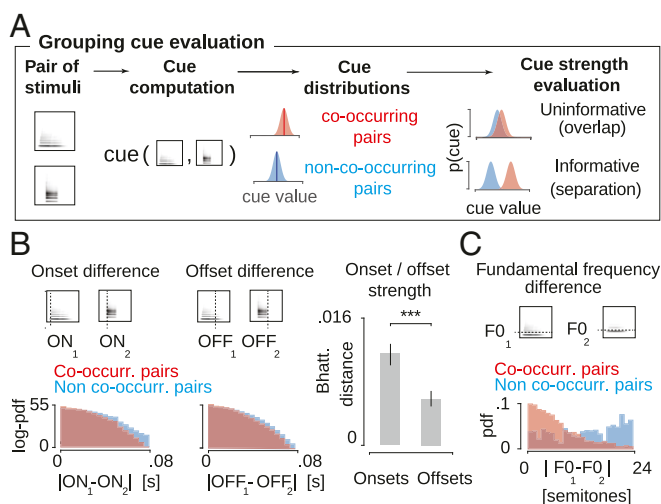


Fig. 5. Grouping cue evaluation. (A) Cues are defined as functions of feature pairs that should differ depending on whether the features are likely to be due to the same source. Cue strength is quantified as the separation of cue distributions for co-occurring and non-co-occurring feature pairs. (B) Evaluation of the cue strength of common onset and common offset. *Left* and *Center* illustrate cue measurement for an example feature pair and the resulting cue distributions for co-occurring and non-co-occurring feature pairs (the top 75% of positive and bottom 75% of negative of feature pairs when ranked according to their association strength). Logarithmic axis serves to reveal the difference between the tails of the distributions. *Right* plots the Bhattacharya distance, a summary measure of the separation of the cue distributions for co-occurring and non-co-occurring feature pairs, predicting that common onset should be a stronger grouping cue than common offset. Error bars plot standard deviation of the bootstrap distribution (obtained by resampling from the sets of feature pairs). Asterisks denote statistical significance of bootstrap test between conditions ($***P < 0.001$). (C) Evaluation of the cue provided by differences in fundamental frequency (f_0), which is small for co-occurring feature pairs and large for non-co-occurring pairs. This analysis was restricted to features that were above a criterion level of periodicity and that thus had a well-defined f_0 .

(20–24) and common fundamental frequency (16–19). We measured onsets and offsets of each feature as the time points where their broadband envelope exceeded or dropped below a threshold value (Fig. 5B, *Upper Left*) and measured the difference in onset or offset time for all pairs of strongly positively or negatively coactivated features (corresponding to the top 75% of positive entries and bottom 75% of negative entries in the association strength tensor) (Fig. 5B, *Lower Left*). Both onset and offset differences were smaller for coactivated features, but the difference was larger for onsets than offsets (quantified with the Bhattacharya distance) (Fig. 5B, *Right*). This difference provides an explanation for the documented difference in the perceptual effect of common onset and offset (whereby grouping from offsets is weaker than grouping from onsets) (22). Similarly, the f_0 difference between features was smaller for coactivated features (Fig. 5C) (measured in features that exceeded a criterion level of periodicity such that the f_0 was well defined). These analyses provide evidence that conventionally cited grouping cues have a sound basis in natural signal statistics.

Grouping Cues Derived by Summarizing Co-occurrence Statistics. We next sought to derive grouping cues from the co-occurrence tensor in order to explore the cues that would emerge independent of human intuition. We searched for acoustic properties that would predict the association strength of feature pairs, restricting the properties to those defined by linear templates in order to facilitate their interpretability. The features were optimized to classify features as belonging to 1 or 2 sounds, as this is

arguably the task faced by the auditory system. The resulting discriminative model learned templates in the time–frequency and modulation domains whose dot product with a spectrotemporal feature kernel was similar for frequently co-occurring features but different for non-co-occurring features.

Specifically, given 2 features, the model computed their projections onto a template. The “cue value” was defined as the magnitude of the difference in the 2 projections. The model used this value to predict whether the features have high association strength or not (via logistic regression) (Fig. 6A). The 2 domains considered (time–frequency and modulation planes) are the most common representations in which to examine sound; the modulation plane is simply the 2-dimensional power spectrum of the time–frequency representation of a sound (34). Templates were learned via gradient descent to maximize discrimination of feature pairs with high and low association strength (roughly the 10% most positive and 10% most negative entries in the tensor) (*Materials and Methods*). Learning occurred sequentially for each template, adding a new template at each iteration until performance reached an asymptote.

The learning procedure resulted in 4 templates, 2 in each of the time–frequency and modulation planes (*Materials and Methods* and Fig. 6B–E) (additional templates only marginally improved performance). We emphasize that the goal of the model was not to fully capture human source separation (which seems likely to require a substantially more complicated model) but rather, to test whether a set of simple acoustic properties would capture important aspects of human auditory grouping. Despite the limitations inherent to linear templates, the 4 templates were sufficient to differentiate co-occurring from non-co-occurring features with reasonable accuracy (81%), indicating that they captured a substantial amount of the variance in feature co-occurrence.

Even though the templates were derived purely from co-occurrence statistics without regard for prior hypotheses or human intuition, inspection of the learned templates reveals interpretable structure. The first cue template (Fig. 6B, *Left*) can be interpreted as computing a spectral centroid, implying that features with similar frequency content are likely to co-occur. We quantified this effect by measuring the spectral centroid of each feature and comparing the centroid difference for feature pairs with high and low cue values (Fig. 6B, *Center* and *Right*). Spectral differences are known to influence the grouping of sounds across time (7, 35, 36), but this result suggests that they also should affect the grouping of concurrent sound energy (the temporal extent of the tensor was ± 80 ms from the center of the reference feature, and the width of feature kernels was 162 ms such that all feature pairs considered in this analysis overlapped in time to a fair extent).

The second template (Fig. 6C) appears to compute a temporal derivative—features that have similar projections tend to be aligned in time (Fig. 6C, column 2), recapitulating the established grouping cue of common onset/offset (20–24). This template also detects misalignments in fundamental frequency (Fig. 6C, column 3), another established grouping cue (16–19, 37, 38).

The modulation plane templates (Fig. 6D and E) compute differences between the power in different regions of the modulation plane and thus capture the tendency of features with different spectral shapes (tone vs. clicks, for example) to belong to distinct sources, regardless of their temporal configuration. To our knowledge, this type of cue has not been previously noted in the auditory scene analysis literature, although modulation rate has been shown to affect the grouping of sequences of sound elements (39).

Perceptual Test of Learned Grouping Cues. The derived cues embodied in the templates varied in their statistical cue strength,

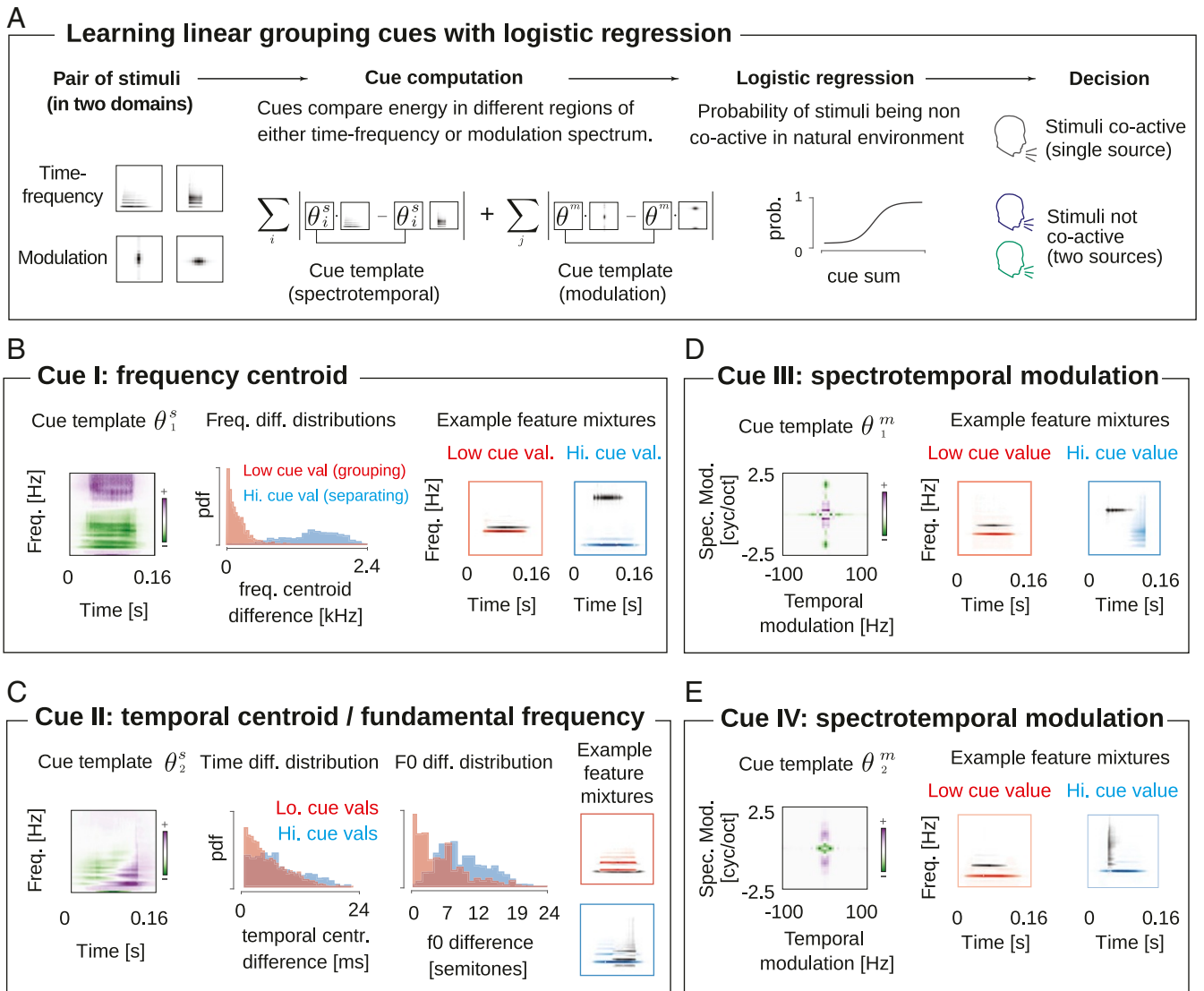


Fig. 6. Learning grouping cues from natural signal statistics. (A) Schematic of discriminative model from which cues were learned. Cues are computed for pairs of features by projecting each feature onto a cue template and taking the absolute value of the difference. The discriminative model takes the sum of these absolute differences for a set of cue templates and predicts whether the feature pair co-occurs or not using logistic regression. The templates could be defined in either the time-frequency or modulation planes. (B–E) Characterization of the 4 learned cues. (Left) Cue templates. (B, Center and C, columns 2 and 3) Distribution of stimulus properties hypothesized to be captured by template. (Right) Example feature pairs with high and low cue values. The cue in C appears to capture 2 conceptually distinct sound properties (temporal offset and fundamental frequency difference) with a single template.

but all were individually predictive of whether feature pairs were associated or not (Fig. 7A, using the analysis of Fig. 5B). To test whether the derived cues affect perceptual grouping, we used each individual template to construct experimental stimuli and measured whether listeners' ability to use the cue in a grouping judgment varied in accordance with its statistical strength in the training corpus of natural sounds (experiment 5). For each cue, we searched for pairs of features with high values of that cue but low values of the other 3 cues such that the cue of interest would provide the only indication that the 2 features were not part of the same source (Fig. 7B, Left). We then presented the pair successively with another pair in which all 4 cues had low values and asked listeners to judge which of the 2 pairs consisted of 2 sources. Listeners were significantly above chance for each cue (Fig. 7B, Right) [$t(14) \geq 4.17$, $P < 0.001$ in all cases], suggesting that all cues contribute to perceptual grouping judgments. Moreover, performance varied with the statistical cue strength, providing addi-

tional evidence that perceptual grouping is based on internalized co-occurrence statistics.

As a further test of the predictive value of the learned cues, we used them to predict the perceptual grouping of 3 types of stimuli: pairs of the learned spectrotemporal kernels, mixtures of artificial sounds synthesized from "blobs" in the time-frequency plane, and mixtures of speech segments windowed by time-frequency apertures. Apertures were used for the speech conditions because mixtures of extended speech excerpts almost never perceptually group to resemble a single source. We searched for stimuli that the cue model confidently judged to be single sources as well as stimuli that the model confidently judged to be mixtures, and on each trial, we presented listeners with one stimulus from each group, asking them to identify the single source (experiment 6). In all 3 cases, listeners' judgments agreed with those of the model [being well above chance for each condition; $t(14) \geq 5.82$, $P < 0.001$ in all cases]. These results provide further evidence for the perceptual

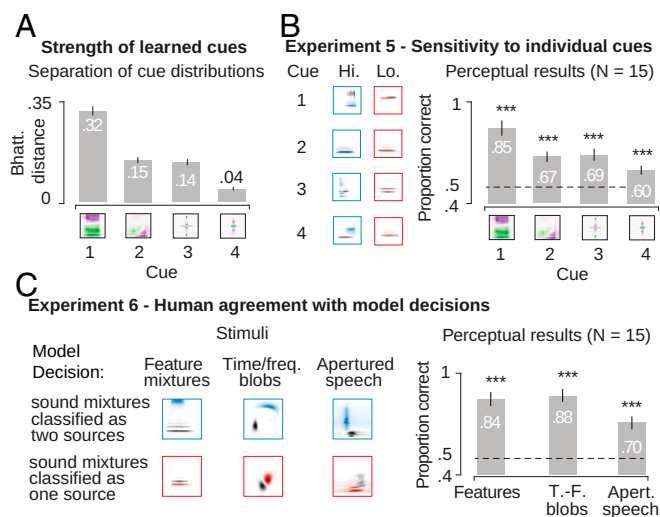


Fig. 7. Perceptual sensitivity to learned grouping cues. (A) Cue strength of the learned cues measured as the Bhattacharyya distance between the cue distributions for co-occurring and non-co-occurring feature pairs. Error bars plot standard deviation of the bootstrap distribution (obtained by resampling from the sets of feature pairs). (B) Description and results of experiment 5, which measured perceptual sensitivity to each of the 4 learned cues. The task was the same as in experiments 1 to 3: listeners heard 2 feature pairs and judged which one consisted of 2 sources. One feature pair on a trial had a low cue value (implying high association strength), and one had a high cue value (implying low association strength). Here and in C, error bars denote SEM, and asterisks denote statistical significance of t tests vs. chance, $***P < 0.001$. (C) Description and results of experiment 6, which measured human agreement with model decisions about the segregation of mixtures of 3 types of stimuli (the spectrotemporal kernels learned from speech and instruments, artificial time-frequency (T.-F.) blobs, and speech excerpts windowed in the time-frequency plane). On each trial, listeners heard 2 mixtures and judged which consisted of 2 sources.

reality of the derived cues and show that they have fairly general predictive power.

Grouping of Feature Sequences. Experiments 1 to 6 demonstrate the perceptual relevance of empirical co-occurrence statistics and of the cues that we derived from them but utilized pairs of features or sound excerpts in close temporal proximity. To test whether the measured co-occurrence statistics would be predictive of the perceptual grouping of more extended sound sequences, we used the co-occurrence tensor to generate sequences of features spaced more widely in time. Each sequence was seeded with an initial feature. Subsequent features were chosen from a probability distribution derived from their association strength with the previous feature, with features with higher association strength having higher probability (Fig. 8A). We then measured whether the co-occurrence statistics could predict the perceptual “streaming” of these sequences. For each of a set of reference sequences, we generated 2 types of mixtures: one with a second sequence with features that had high association strength with the features of the reference sequence and one with features that did not (*Materials and Methods* and Fig. 8B). Listeners were presented with a mixture and judged whether it was generated by 1 or 2 sources (experiment 7).

As shown in Fig. 8C, listeners reliably judged the mixture with the non-co-occurring sequence as 2 sources but showed the opposite tendency for the mixtures with the co-occurring sequence [$t(10) = 9.56$, $P < 0.001$, t test]. Subjectively, the sequences in a non-co-occurring mixture typically differed in their acoustic qualities, and attention could often be directed

to one or the other. There was thus some similarity to classical examples of streaming with alternating tones and other simple sound elements (35, 36), even though the sound sequences here were more stochastic and varied. The results indicate that pairwise co-occurrence statistics capture some of the principles that cause extended sound sequences to perceptually stream.

Discussion

We introduced a framework for measuring natural signal statistics that could underlie perceptual grouping and explored their relationship to perception in the domain of audition. We first learned local acoustic features from natural audio signals (speech and instrument recordings) (Fig. 2) and computed their strength of co-occurrence (Fig. 3). Our results revealed that acoustic features exhibit rich pairwise dependencies, but that these co-occurrences could be summarized to a fair extent with a modest number of “cues.” We formalized the notion of a cue as a stimulus property that predicts the co-occurrence of pairs of features (Fig. 5) and derived cues from the large set of measured pairwise co-occurrence statistics (Fig. 6). The cues that emerged include some previously known to influence grouping (such as common onset and fundamental frequency) as well as others that have not previously been widely acknowledged (such as separation in acoustic and modulation frequency for concurrent features). We found evidence that the auditory system has internalized these statistics and uses them to group features into coherent objects. This was true both for isolated pairs of features (experiments 1 to 4 and 6) and for more extended feature “streams” (experiment 7) as well as for each of the individual cues revealed by the co-occurrence statistics (experiment 5). These results provide a quantitative link between auditory perceptual grouping and natural sound statistics, show how these statistics may be harnessed to study auditory scene analysis, and

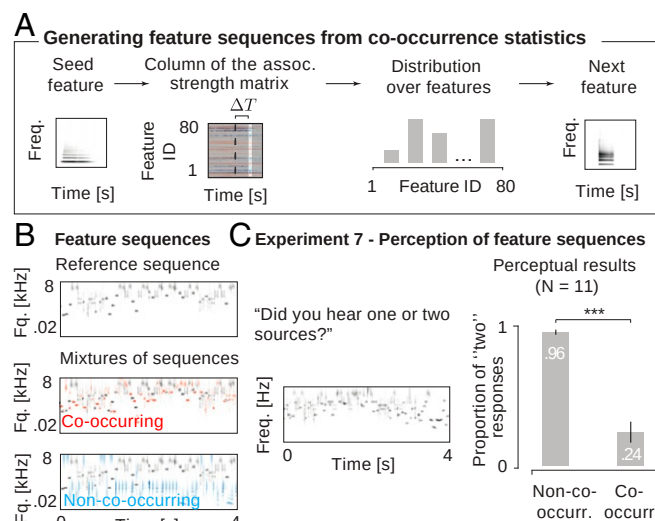


Fig. 8. Streaming of spectrotemporal feature sequences. (A) Sequence generation from co-occurrence statistics. First, a seed feature is chosen. Second, the column of its association strength matrix is extracted for the desired time offset for the next feature (here fixed at 75 ms). Third, the column is transformed to a probability distribution via the softmax function. Fourth, the next feature is drawn from this distribution. These steps are iterated until a sequence of the desired length is obtained. (B) Example reference sequence (Top) mixed with a co-occurring sequence (Middle) and a non-co-occurring sequence (Bottom). (C) Description and results of experiment 7. On each trial, listeners heard a mixture of 2 feature sequences and judged whether it was produced by 1 or 2 sources. Error bars denote SEM. Asterisks denote statistical significance of a paired t test between conditions ($***P < 0.001$).

offer a general framework for relating natural signal properties to perceptual grouping.

Related Work. The derivation of ideal observer models has a long and productive history in perception research (40), and such models have been used to learn cues for a range of natural tasks (41, 42). Previous such attempts to relate perceptual grouping to natural scene statistics have largely been limited to contour grouping in images (8–12). These influential earlier efforts inspired our work here but were reliant on hand-picked features labeled by human observers (object edges), and their analysis was limited to dimensions thought to be important *a priori* (position and orientation). Our results demonstrate how one can derive grouping cues from features learned entirely from natural signals without prior hypotheses about the features or underlying grouping principles. Learning signal features and grouping cues from the structure of natural sounds paid dividends by revealing statistical effects that were not obvious beforehand and that were found to have corresponding perceptual grouping effects. Our methodology also gives additional support to commonly discussed cues by showing that they emerge from the large set of possible cues that might in principle have been derived from natural sound statistics.

Our results complement a long research tradition that has documented behavioral and neural effects of a handful of acoustic grouping cues, relying on intuitively plausible cues and synthetic stimuli (7, 16–22, 37, 43). We provide statistical justification for the 2 most commonly studied cues from this literature (onset and harmonicity) but also identify other statistical effects and show their perceptual relevance. Frequency separation is known to strongly affect the grouping of stimuli over time (35, 36) but is less acknowledged to influence the grouping of concurrent features. Our results show that it is the strongest effect evident in local co-occurrence statistics of natural audio, at least for the corpora that we analyzed, and that it has a correspondingly strong perceptual effect. Modulation differences have also not been widely appreciated as an influence on the grouping of concurrent features (39) but emerged from the analysis of co-occurrence statistics and also proved to have a large perceptual effect. The analysis of natural signal statistics is thus “postdictive,” suggesting normative explanations for known effects, but can also be predictive, pointing us to phenomena that we should test experimentally.

Our quantitative approach to grouping has the added benefit of taking us beyond verbal descriptions of phenomena to enable grouping predictions for arbitrary stimuli. We leveraged this ability to make such predictions for 3 different types of stimuli (experiment 5). The verbal characterizations of cues from classical approaches cannot be tested in this way.

Our approach also complements engineering efforts to solve auditory grouping. Early attempts in this domain were inspired by psychoacoustic observations and implemented hand-engineered grouping constraints based on common onset and periodicity (44–46). More recent attempts to build computational models of sound segregation similarly focus on the intuitively plausible cue of temporal coincidence (23, 24). Current state-of-the-art engineering methods instead rely on learning how to group acoustic energy from labeled sound mixtures (47, 48) but are at present difficult to probe for insight into the underlying acoustic dependencies. Our methodology falls between these 2 traditions, using the rich set of constraints imposed by natural signals but providing interpretable insight into factors that might underlie grouping. Indeed, our choices to restrict the analysis to pairwise dependencies and to learn linear cues that summarize the measured dependencies were made to facilitate inspection of the results.

The choices that we made to enable interpretability come at the expense of predictive power: the cues do not perfectly

predict the empirical statistical association between features, and they do not perfectly predict human judgments. This no doubt reflects in part the complexity of the source separation problem, the optimal solution of which seems likely to require more than pairwise feature associations and linear cue templates. In this respect, the problem that we are modeling may be distinct from other more limited perceptual tasks that have been successfully modeled using simple cue features (41, 42, 49, 50). We suspect that models that make accurate quantitative predictions about human source separation will need to be substantially more complicated than the discriminative model that we used here. However, this complexity may come at the cost of interpretability (51) as in contemporary source separation methods (47, 48). Grouping cues as traditionally conceived (and as derived here) may be limited to coarsely approximating the mechanisms underlying real-world perceptual organization, providing insight and the ability to make qualitative predictions but falling short of a full explanation of human abilities.

Open Issues and Future Directions. Our approach leveraged available recordings of single sound sources. Single-source recordings provide a weak form of supervision in that the resulting feature activations can be assumed to belong together without requiring the use of human labels that were critical to previous work in this vein (8–12). However, because large numbers of single-source recordings are presently available only for speech and musical instruments, our analysis was limited to these sound genres. Humans encounter many other types of sounds, and our results may thus not reflect the full set of dependencies that influence perception. However, speech and instruments instantiate many of the types of physical processes that can generate sound in the world (52): impact sounds, sound produced by blowing air in various ways, periodic and aperiodic source energy filtered by resonant bodies, etc. It thus seems plausible that the dependencies learned from the combination of speech and instruments could approximate many of the statistical properties that matter for auditory grouping. However, the results would no doubt be quantitatively different if it were possible to include other types of sound (53), and an expanded corpus might yield association strengths that are more strongly predictive of perception.

The use of large corpora of recorded audio had the additional consequence that our analysis was restricted to monaural audio. Natural auditory input likely contains important binaural dependencies that contribute to grouping (54–58) that our approach could in principle capture if applied to audio recorded from 2 ears (59). Another limitation of our approach lies in the use of sparse feature decodings, which efficiently describe speech and music sounds but are a poor description of more noise-like sounds, such as textures (33). Textures are an important part of auditory scene analysis (60), and studying the statistical basis of their grouping will likely require an alternative encoding scheme, potentially based on summary statistics (61) rather than localized time–frequency features.

Our results suggest that human listeners have internalized the co-occurrence statistics that we measured: Listeners reliably discriminate between feature pairs with high and low association strength (Fig. 4). The results leave open whether knowledge of the dependencies is built into the auditory system over evolution, whether it is learned during development, and/or whether it continues to be updated during adulthood. If evolved, grouping principles could potentially even predate the origins of speech and music, in which case the match between perception and our corpus statistics might reflect the adaptation of speech and music to the auditory system (which, by hypothesis, would then be shaped primarily by other classes of natural sounds) rather than the other way around (62). However, some types of sound source structure can be learned relatively quickly (63, 64) and can

aid source separation (65), raising the possibility that the local feature dependencies studied here could be learned over development, plausibly from speech and music sounds among others. This could in principle be addressed by exposing listeners to sounds with altered statistical dependencies and then measuring whether perceptual grouping is altered.

A full account of auditory scene analysis will undoubtedly require more complete statistical models of natural sound sources, incorporating more than the pairwise dependencies between local features studied here (26). In addition to multi-element dependencies, a full model will likely require additional hierarchical structure, in which groups of local features are in turn grouped into larger-scale configurations. Such hierarchical organization could be one way to model the grouping effects of repetition (66, 67), which is one powerful grouping phenomenon not accounted for by our analysis.

The instantiation of perceptual grouping in the brain remains a key open issue in systems neuroscience, particularly in audition (23, 68–70). The features that we measured could plausibly be detected by neurons in the auditory system (28), and the co-occurrence statistics that we analyzed could in principle be encoded by connections between neurons, analogous to the association fields for contour grouping that are thought to be instantiated in lateral connections between visual neurons (71). Alternatively, co-occurrence statistics could be encoded by higher-level sensory neurons implementing logical and/or-like

computations (72–74). The latter possibility could be tested by comparing the components of such multidimensional receptive fields with the cue templates that we derived.

Although our methodology starts from an encoding scheme based on local features, in part because these are most readily mapped onto early stages of sensory systems (62, 75, 76), problems of scene analysis can also be approached with generative models more rooted in how sounds are produced (77). For instance, speech and instrument sounds are fruitfully characterized as the product of a source and a filter that each vary over time in particular ways (78, 79), as are sounds in reverberant environments (80), and humans seem to have implicit knowledge of this generative structure (81). Reconciling these generative models for sound with those rooted in neurally plausible local feature decompositions is a critical topic for future research.

Materials and Methods

Methods are described in full detail in *SI Appendix, SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Sophia Dolan, River Grace, Malinda McPherson, and Richard McWalter for running some of the psychophysical experiments; and the laboratory of J.H.M. for comments on the manuscript. This material is based on work supported by the Center for Brains, Minds and Machines funded by NSF Science and Technology Center Award CCF-1231216. Work was also supported by a McDonnell Scholar Award (to J.H.M.) and NIH Grant R01DC014739.

1. M. Wertheimer, Untersuchungen zur lehre von der gestalt. ii. *Psychol. Forsch.* **4**, 301–350 (1923).
2. W. Kohler, *Gestalt Psychology* (Liveright, New York, 1929).
3. F. Attneave, Some informational aspects of visual perception. *Psychol. Bull.* **61**, 183–193 (1954).
4. J. Feldman, Bayesian contour integration. *Percept. Psychophys.* **63**, 1171–1182 (2001).
5. D. Kersten, P. Mamassian, A. Yuille, Object perception as bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004).
6. S. Palmer, *Vision Science: Photons to Phenomenology* (MIT Press, Cambridge, MA, 1999).
7. A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA, 1990).
8. E. Brunswik, J. Kamiya, Ecological cue-validity of 'proximity' and of other gestalt factors. *Am. J. Psychol.* **66**, 20–32 (1953).
9. W. Geisler, J. Perry, B. Super, D. Gallogly, Edge co-occurrence in natural images predicts contour grouping performance. *Vis. Res.* **41**, 711–724 (2001).
10. J. Elder, R. Goldberg, Ecological statistics of gestalt laws for the perceptual organization of contours. *J. Vis.* **2**, 5 (2002).
11. W. Geisler, J. Perry, Contour statistics in natural images: Grouping across occlusions. *Vis. Neurosci.* **26**, 109–121 (2009).
12. M. Sigman, G. Cecchi, C. Gilbert, M. Magnasco, On a common circle: Natural scenes and gestalt rules. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 1935–1940 (2001).
13. C. Darwin, Auditory grouping. *Trends Cogn. Sci.* **1**, 327–333 (1997).
14. R. Carlyon, How the brain separates sounds. *Trends Cogn. Sci.* **8**, 465–471 (2004).
15. J. H. McDermott, The cocktail party problem. *Curr. Biol.* **19**, R1024–R1027 (2009).
16. B. Moore, B. Glasberg, R. Peters, Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *J. Acoust. Soc. Am.* **80**, 479–483 (1986).
17. W. Hartmann, S. McAdams, B. Smith, Hearing a mistuned harmonic in an otherwise periodic complex tone. *J. Acoust. Soc. Am.* **88**, 1712–1724 (1990).
18. A. de Cheveigne, S. McAdams, C. Marin, Concurrent vowel identification. ii. effects of phase, harmonicity, and task. *J. Acoust. Soc. Am.* **101**, 2848–2856 (1997).
19. S. Popham, D. Boebinger, D. Ellis, H. Kawahara, J. McDermott, Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nat. Commun.* **9**, 2122 (2018).
20. R. Rasch, The perception of simultaneous notes such as in polyphonic music. *Acustica* **40**, 21–33 (1978).
21. C. Darwin, Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Q. J. Exp. Psychol.* **33**, 185–207 (1981).
22. C. Darwin, Perceiving vowels in the presence of another sound: Constraints on formant perception. *J. Acoust. Soc. Am.* **76**, 1636–1647 (1984).
23. S. A. Shamma, M. Elhilali, C. Micheyl, Temporal coherence and attention in auditory scene analysis. *Trends Neurosci.* **34**, 114–123 (2011).
24. L. Krishnan, M. Elhilali, S. A. Shamma, Segregating complex sound sources through temporal coherence. *PLoS Comput. Biol.* **10**, e1003985 (2014).
25. M. S. Lewicki, T. J. Sejnowski, "Coding time-varying signals using sparse, shift-invariant representations" in *Advances in Neural Information Processing Systems (NIPS)*, M. I. Jordan, Y. LeCun, S. A. Solla, Eds. (MIT Press, Cambridge, MA, 1999), pp. 730–736.
26. W. Młynarski, J. McDermott, Learning mid-level auditory codes from natural sound statistics. *Neural Comput.* **30**, 631–669 (2018).
27. D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
28. D. Depireux, J. Simon, D. Klein, S. Shamma, Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *J. Neurophysiol.* **85**, 1220–1234 (2001).
29. A. J. Kell et al., A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* **98**, 630–644 (2018).
30. Z. Tüske, R. Schlüter, H. Ney, "Acoustic modeling of speech waveform based on multi-resolution, neural network signal processing" in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. <https://ieeexplore.ieee.org/document/8461871>. Accessed 14 November 2019.
31. L. Ondel, R. Li, G. Sell, H. Hermansky, "Deriving spectro-temporal properties of hearing from speech data" in *Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. <https://ieeexplore.ieee.org/document/8682787>. Accessed 14 November 2019.
32. W. Młynarski, J. H. McDermott, Ecological origins of perceptual grouping principles in the auditory system - Stimulus examples. <http://mcdermottlab.mit.edu/grouping-statistics/index.html>. Deposited 1 December 2018.
33. J. H. McDermott, E. Simoncelli, Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron* **71**, 926–940 (2011).
34. N. C. Singh, F. E. Theunissen, Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am.* **114**, 3394–411 (2003).
35. L. van Noorden, Minimum differences of level and frequency for perceptual fission of tone sequences. *J. Acoust. Soc. Am.* **61**, 1041–1045 (1977).
36. B. C. Moore, H. E. Gockel, Properties of auditory stream formation. *Philos. Trans. R. Soc. Biol. Sci.* **367**, 919–931 (2012).
37. J. Culling, C. Darwin, Perceptual separation of simultaneous vowels: Within and across-formant grouping by f0. *J. Acoust. Soc. Am.* **93**, 3454–3467 (1993).
38. K. Woods, J. McDermott, Attentive tracking of sound sources. *Curr. Biol.* **25**, 2238–2246 (2015).
39. N. Grimault, S. Bacon, C. Micheyl, Auditory stream segregation on the basis of amplitude-modulation rate. *J. Acoust. Soc. Am.* **111**, 1340–1348 (2002).
40. W. Geisler, Contributions of ideal observer theory to vision research. *Vis. Res.* **51**, 771–781 (2011).
41. J. Burge, W. Geisler, Optimal speed estimation in natural image movies predicts human performance. *Nat. Commun.* **6**, 7900 (2015).
42. J. Burge, W. S. Geisler, Optimal defocus estimation in individual natural images. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 16849–16854 (2011).
43. R. Carlyon, Discriminating between coherent and incoherent frequency modulation of complex tones. *J. Acoust. Soc. Am.* **89**, 329–340 (1991).
44. G. J. Brown, M. Cooke, Computational auditory scene analysis. *Comput. Speech Lang.* **8**, 297–336 (1994).
45. D. Wang, G. J. Brown, Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Trans. Neural Netw.* **10**, 684–697 (1999).
46. D. Wang, G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (John Wiley & Sons, Hoboken, NJ, 2006).

47. J. R. Hershey, Z. Chen, J. Le Roux, S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation" in *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. <https://ieeexplore.ieee.org/document/7471631>. Accessed 14 November 2019.
48. Z. Chen, Y. Luo, N. Mesgarani, "Deep attractor network for single-microphone speaker separation" in *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. <https://ieeexplore.ieee.org/document/7952155>. Accessed 14 November 2019.
49. W. Kienzle, M. O. Franz, B. Schölkopf, F. A. Wichmann, Center-surround patterns emerge as optimal predictors for human saccade targets. *J. Vis.* **9**, 7–7 (2009).
50. A. R. Girshick, M. S. Landy, E. P. Simoncelli, Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nat. Neurosci.* **14**, 926–932 (2011).
51. W. S. Geisler, J. Najemnik, A. D. Ing, Optimal stimulus encoders for natural tasks. *J. Vis.* **9**, 17–17 (2009).
52. W. W. Gaver, What in the world do we hear?: An ecological approach to auditory event perception. *Ecol. Psychol.* **5**, 1–29 (1993).
53. M. S. Lewicki, Efficient coding of natural sounds. *Nat. Neurosci.* **5**, 356–363 (2002).
54. J. Culling, Q. Summerfield, Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *J. Acoust. Soc. Am.* **98**, 785–797 (1995).
55. C. Darwin, R. Hukin, Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *J. Acoust. Soc. Am.* **102**, 2316–2324 (1997).
56. R. Litovsky, H. Colburn, W. Yost, S. Guzman, The precedence effect. *J. Acoust. Soc. Am.* **106**, 1633–1654 (1999).
57. R. Weiss, M. Mandel, D. Ellis, Combining localization cues and source model constraints for binaural source separation. *Speech Commun.* **53**, 606–621 (2011).
58. A. Schwartz, J. McDermott, B. Shinn-Cunningham, Spatial cues alone produce inaccurate sound segregation: The effect of interaural time differences. *J. Acoust. Soc. Am.* **132**, 357–368 (2012).
59. W. Młynarski, The opponent channel population code of sound location is an efficient representation of natural binaural sounds. *PLoS Comput. Biol.* **11**, e1004294 (2015).
60. R. I. McWalter, J. McDermott, Adaptive and selective time-averaging of auditory scenes. *Curr. Biol.* **28**, 1405–1418 (2018).
61. J. McDermott, M. Schemitsch, E. Simoncelli, Summary statistics in auditory perception. *Nat. Neurosci.* **16**, 493–498 (2013).
62. E. Smith, M. Lewicki, Efficient auditory coding. *Nature* **439**, 978–982 (2006).
63. T. Agus, S. Thorpe, D. Pressnitzer, Rapid formation of auditory memories: Insights from noise. *Neuron* **66**, 610–618 (2010).
64. C. E. Stilp, T. T. Rogers, K. R. Kluender, Rapid efficient coding of correlated complex acoustic properties. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 21914–21919 (2010).
65. K. Woods, J. McDermott, Schema learning for the cocktail party problem. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E3313–E3322 (2018).
66. G. Kidd, C. Mason, P. Deliwala, W. Woods, Reducing informational masking by sound segregation. *J. Acoust. Soc. Am.* **95**, 3475–3480 (1994).
67. J. H. McDermott, D. Wroblewski, A. J. Oxenham, Recovering sound sources from embedded repetition. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 1188–1193 (2011).
68. Y. I. Fishman, J. C. Arezzo, M. Steinschneider, Auditory stream segregation in monkey auditory cortex: Effects of frequency separation, presentation rate, and tone duration. *J. Acoust. Soc. Am.* **116**, 1656–1670 (2004).
69. D. Pressnitzer, M. Sayles, C. Micheyl, I. Winter, Perceptual organization of sound begins in the auditory periphery. *Curr. Biol.* **18**, 1124–1128 (2008).
70. I. Winkler, S. L. Denham, I. Nelken, Modeling the auditory scene: Predictive regularity representations and perceptual objects. *Trends Cogn. Sci.* **13**, 532–540 (2009).
71. D. Field, A. Hayes, R. Hess, Contour integration by the human visual system: Evidence for a local "association field." *Vis. Res.* **33**, 173–193 (1993).
72. C. Atencio, T. Sharpee, C. E. Schreiner, Hierarchical computation in the canonical auditory cortical circuit. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 21894–21899 (2009).
73. N. S. Harper *et al.*, Network receptive field modeling reveals extensive integration and multi-feature selectivity in auditory cortical neurons. *PLoS Comput. Biol.* **12**, e1005113 (2016).
74. A. Kozlov, T. Gentner, Central auditory neurons have composite receptive fields. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 1441–1446 (2016).
75. B. Olshausen, D. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
76. N. Carlson, V. Ming, M. DeWeese, Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS Comput. Biol.* **8**, e1002594 (2012).
77. M. Cusimano, L. B. Hewitt, J. Tenenbaum, J. H. McDermott, "Auditory scene analysis as bayesian inference in sound source models" in *2019 Conference on Computational Cognitive Neuroscience*, 10.32470/CCN.2018.1039-0 (2018).
78. K. N. Stevens, *Acoustic Phonetics* (MIT Press, 2000).
79. N. Fletcher, T. Rossing, *The Physics of Musical Instruments* (Springer, 2010).
80. W. Gardner, *Reverberation Algorithms* (Kluwer Academic Publishers, Norwell, MA, 1998).
81. J. Traer, J. McDermott, Statistics of natural reverberation enable perceptual separation of sound and space. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E7856–E7865 (2016).