Natural Sound Statistics Predict Auditory Grouping Principles

Wiktor Młynarski (mlynar@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Josh H. McDermott (jhm@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Abstract

Events and objects in the world must be inferred from sensory signals to support behavior. Because sensory signals are transduced with measurements that are temporally and spatially local, the estimation of a particular object or event can be viewed as the result of grouping these local measurements into representations of their common causes. In the auditory system, perceptual grouping is believed to exploit acoustic regularities of natural sounds, such as the tendency of frequencies to be harmonically related or to share a common onset. However, acoustic grouping cues have traditionally been identified using intuitions and informal observation, and investigated using simple, artificial stimuli. As a result, the relevance of known grouping cues to real-world auditory scene analysis remains unclear, and additional or alternative cues remain a possibility. Here we derive auditory grouping cues from co-occurrence statistics of local acoustic features in natural sounds. This process recovers established cues but also reveals previously unappreciated aspects of grouping. The results provide confirmation that auditory grouping is adapted to natural stimulus statistics, and show how these statistics can be harnessed to reveal novel grouping phenomena.

Keywords: natural scene statistics; probabilistic models; perception; audition;

Methods

We first derived a set of primitive auditory patterns by learning a dictionary of spectrotemporal features from a corpus of natural sounds using sparse convolutional coding with non-negativity constraints (Fig. 1). Specifically, we trained a dictionary of 80 spectrotemporal features (kernels) using non-negative, convolutional sparse coding (Fig. 1A,B) on a corpus of speech sounds (TIMIT) and musical instruments.

We then extracted sets of features that either were or were not strongly co-activated in the corpus. For each feature, we computed the average activations of all other features conditioned on the activation of the selected feature exceeding its 95th percentile (Fig. 2). We considered a kernel as co-occurring or not with the selected feature depending on whether this conditional average activation was larger or smaller than its baseline activation (the mean feature activation across the entire speech corpus).

The co-occurrence statistics form a three-dimensional tensor that is not easily inspected (containing the co-activation map for each feature, one of which is shown in Fig. 2B. To relate these statistics to a compact representation of acoustic properties, we designed a discriminative model based on logistic regression. The model projects each acoustic feature onto templates in the time-frequency or modulation planes (the two most common domains in which to examine sound; Fig. 4 A, left and right column respectively), and uses the difference in the projections for two features to predict whether they have high co-occurrence probability or not.

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. On each trial, subjects heard two such stimuli and judged which of them contained two sound sources. One feature pair was selected from the 10% of feature pairs with highest co-occurrence probability, and the other from the 10% of feature pairs with lowest co-occurrence probability. To set a ceiling level on task performance, in another condition, one stimulus was an excerpt of a single speech signal while the other was an excerpt of a mixture of talkers. Because speech contains 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. As a control condition we conducted the same task but with stimuli generated from co-occurrence statistics of modulated noise.

Results

Human listeners reliably identified unlikely sets of features as sounds consisting of two sources (Fig. 3B, center), only slightly below the level for speech mixtures (Fig. 3B, left). In contrast, listeners were unable to identify the unlikely feature pairs generated from co-occurrence statistics from modulated noise (Fig. 3B, right). This result suggests that humans have internalized aspects of the co-occurrence tensor and associate the learned statistics with the perception of grouping.

The discriminative template model provides some insight into what is captured by the co-occurrence statistics. The model learned four templates, two in each of the time-frequency and modulation planes (Fig. 4). Features with similar projections onto the templates were predicted to co-occur, and the four templates were sufficient to differentiate co-occurring from non-co-occurring features with reasonable accuracy (81%).

Inspection of the learned templates reveals interpretable structure. The first spectrotemporal template (Fig. 4A, top left) can be interpreted as computing a spectral centroid, implying that features with similar frequency content are likely to co-occur. Spectral differences are known to influence the grouping of sounds across time (Bregman, 1994; van Noorden, 1977), but this result suggests that they also should affect the grouping of concurrent features. The second spectrotemporal template appears to compute a temporal derivative - features that will have similar projections will tend to have temporally aligned onsets or offsets, recapitulating the established grouping cue of common onset (Rasch, 1978; Darwin & Ciocca, 1992). This template also appears to register misaligned sets of harmonics, another established grouping cue (Moore, Glasberg, & Peters, 1986; De Cheveigné, McAdams, & Marin, 1997; Popham, Boebinger, Ellis, Kawahara, & McDermott, 2018).

The modulation plane features compute differences between different regions of the modulation plane, and thus indicate that features with different spectral shapes (tone vs. clicks, for example) do not co-occur. To our knowledge this type of cue has not been previously noted in the auditory scene analysis literature.

Discussion

The results illustrate a way of measuring natural sound statistics likely to be relevant for grouping, and show that they contain interpretable structure. The statistics justify previously known grouping cues such as common onset and harmonicity, but also reveal previously unacknowledged principles of grouping such as frequency separation and spectrotemporal modulation differences. Moreover, we have provided evidence that humans have internalized these statistics and use them to make grouping judgments. These results provide what to our knowledge is the first quantitative link between auditory perceptual grouping and natural sound statistics, and show how these statistics may be harnessed to study auditory scene analysis.

Acknowledgments

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216, by NSF Grant No. BCS-1454094, and by NIH Grant No. 1R01DC014739- 01A1.

References

Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. MIT press.

Darwin, C., & Ciocca, V. (1992). Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component. *The Journal of the Acoustical Society of America*, *91*(6), 3381–3390.

De Cheveigné, A., McAdams, S., & Marin, C. M. (1997). Concurrent vowel identification. ii. effects of phase, harmonicity, and task. *The Journal of the Acoustical Society of America*, *101*(5), 2848–2856.

Moore, B. C., Glasberg, B. R., & Peters, R. W. (1986). Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *The Journal of the Acoustical Society of America*, *80*(2), 479–483.

Popham, S., Boebinger, D., Ellis, D., Kawahara, H., & McDermott, J. (2018). Inharmonic speech reveals the role of harmonicity in the cocktail party problem. *Nature Communications*.

Rasch, R. A. (1978). The perception of simultaneous notes such as in polyphonic music. *Acta Acustica united with Acustica*, *40*(1), 21–33.

van Noorden, L. P. (1977). Minimum differences of level and frequency for perceptual fission of tone sequences abab. *The Journal of the Acoustical Society of America*, *61*(4), 1041–1045.

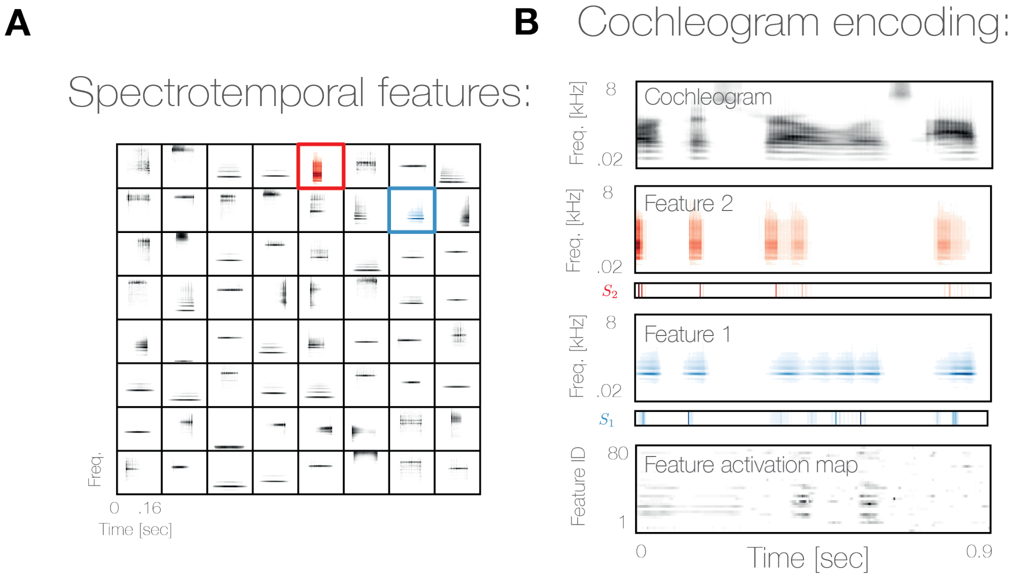

Figure 1: **Spectrotemporal features** A) Dictionary of spectrotemporal features learned from a corpus of speech and instrument sounds (64 out of 80 displayed). B) A cochleogram excerpt (top row) is encoded by a feature activation map (bottom row). Contributions of two individual features to cochleogram encoding are depicted in middle rows. Colors correspond to features highlighted in panel A.

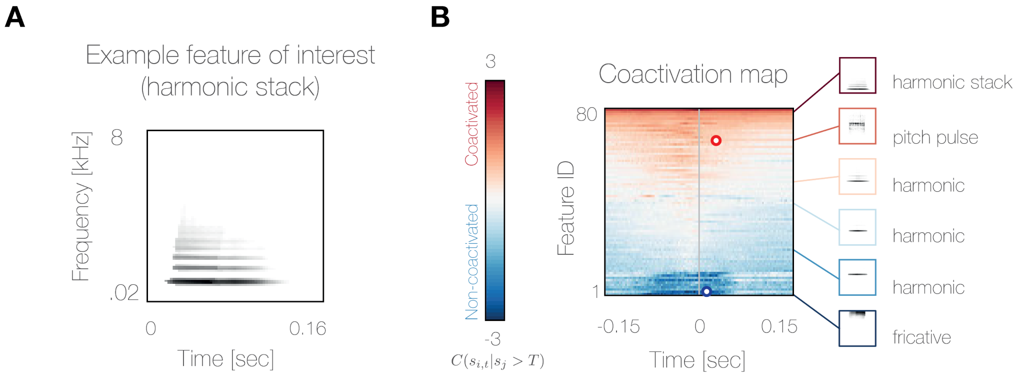

Figure 2: **Feature co-occurrence statistics** A) Example feature of interest B) A co-occurrence matrix for that feature of interest. The color plots the log-ratio of the conditional activation of each feature at each time offset to its baseline activation. Three example co-active and non-co-active features are depicted in the right column.

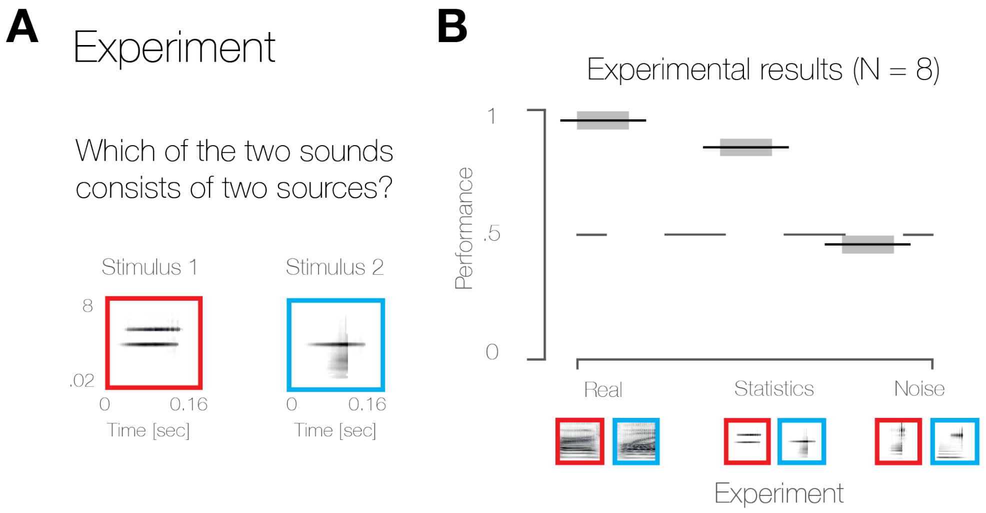

Figure 3: **Experimental results** A) An experimental trial. Subjects judged which of the two sounds was a mixture of two sources. B) Experimental results for mixtures of natural sounds (left), feature pairs derived from natural sound statistics (middle), and feature pairs derived from modulated noise statistics (right).

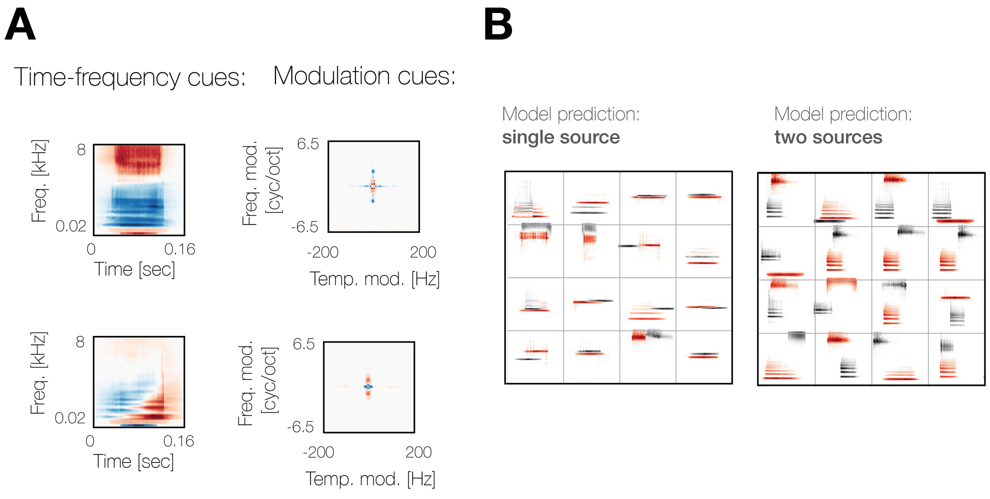

Figure 4: **Grouping cues derived from co-occurrence statistics** A) Time-frequency and modulation cues learned by the discriminative model. The model classifies a feature pair to be generated by two sources if they have different projections onto the learned templates (two learned in the time-frequency plane, and two learned in the modulation plane). B) Example feature pairs judged by the model to be coming from a single source (left) and two sources (right).