Auditory scene analysis as Bayesian inference in sound source models

Maddie Cusimano (mcusi@mit.edu), Luke Hewitt (lbh@mit.edu),
Joshua B. Tenenbaum (jbt@mit.edu), Josh H. McDermott (jhm@mit.edu)
Department of Brain and Cognitive Sciences, 46 Vassar Street
Cambridge, MA 02139 USA

Abstract

Inferring individual sound sources from the mixture of soundwaves that enters our ear is a central problem in auditory perception, termed auditory scene analysis (ASA). The study of ASA has uncovered a diverse set of illusions that suggest general principles underlying perceptual organization. However, most explanations for these illusions remain intuitive (or are narrowly focused), without formal models that predict perceived sound sources from the acoustic waveform. Whether ASA phenomena can be explained by a small set of principles is therefore unclear. We present a Bayesian model based on representations of simple acoustic sources, for which a deep neural network is used to guide Markov chain Monte Carlo inference. Given a sound waveform, our system infers the number of sources present, parameters defining each source, and the sound produced by each source. This model qualitatively accounts for perceptual judgments on a variety of classic ASA illusions, and can in some cases infer perceptually valid sources from simple audio recordings.

Keywords: Bayesian modeling; Probabilistic programs; Auditory perception; Auditory scene analysis; Natural scenes; Perceptual organization; Perceptual grouping; Source separation

Introduction

Listening to music, the ambience of a city street or even your relatively quiet office, one striking aspect of our phenomenal experience is the presence of multiple streams of sound. We tend to experience these streams as arising from distinct sources; walking down the street, you might hear your footsteps crunching in the snow, cars moving in the distance, and your friend speaking. Indeed, the acoustic signal received by the ear is often a mixture of soundwaves generated by various sources, and apprehending these individual sources facilitates interacting with the world. Inferring sources from sound is a central problem in auditory perception and is commonly termed auditory scene analysis (ASA, Bregman (1990)). ASA is inherently ill-posed: infinitely many combinations of sources can generate the same signal. The problem is only solvable due to regularities in natural audio, which the auditory system must internalize as priors to enable source inference.

Historically, synthetic auditory stimuli akin to visual illusions have been used to uncover perceptual priors (Bregman, 1990), demonstrating listeners' tendencies to perceive particular types of source structure. Research over the past five decades has documented a wide variety of such phenomena, revealing the richness of auditory perceptual organization. However, at present, we lack a formal account of these auditory demonstrations, let alone everyday auditory scenes.

Prior attempts to model ASA can be loosely divided into two approaches. The first is mechanistic, describing ASA as a sequence of neurally-inspired transformations. Such models almost exclusively address the perceptual grouping of sequential synthetic tones (e.g., Mill, Bohm, Bendixen, Winkler, and Denham (2013)) or fundamentally depend on the presence of a fixed number of sources (e.g., Krishnan, Elhilali, and Shamma (2014)). It is unclear how these models could scale to natural scenes, given that such scenes consist of 1) a variable number sources, and 2) differently and richly structured sources.

A second approach involves a computational level analysis, framing ASA as the inference of probable causal explanations for sensory data. Bayesian inference provides a rational framework for integrating sensory observations with prior knowledge about the causal processes that generate those observations. The small amount of previous related work demonstrates the difficulty and promise of this approach. A common limitation, as in other domains, is that expressive models typically face insurmountable inference problems (Ellis, 2006). Turner (2010) modeled a variety of ASA phenomena as generative inference, but prespecified potential sources for each illusion, avoiding the structural aspect of inference. Yates, Larigaldie, and Beierholm (2017) explained the perception of tone sequences by inferring the number of sources and the assignment of tones to each source in a nonparametric model (applied to symbolic input). Their work demonstrates the utility of generative models in inferring the structure of auditory scenes, but the scenes were very limited, and their model cannot be applied to raw audio (in which the basic elements are not explicit).

Here, we present a computational model aimed at providing the foundation for a comprehensive account of human ASA. We believe such a foundation necessitates inference from the audio signal and the ability to describe diverse sources. Given an observed sound, our model infers the number of sources, the parameters defining each source, and the sounds produced by each source. We frame this analysis as Bayesian inference in a probabilistic model of auditory scenes. In particular our model is a probabilistic program (Goodman & Stuhlmüller, 2014), expressing uncertainty over both continuous and structural latent variables. Due to the rich structure of our model, inference from raw waveforms poses a significant computational challenge. We overcome this by first training a deep neural network on sounds generated by the model, and then using this network to guide Markov chain Monte Carlo (MCMC) inference. We show that our model qualitatively replicates perceptual organization in a variety of synthetic ASA demonstrations and some simple audio recordings.

Table 1: Scene sampling procedure

scene() :
  $n_{sources} \sim \text{Uniform}(\{1,\ldots,5\})$
  for $i = 1,\ldots,n_{sources}$
    $type_i \sim \text{Uniform}(\{\text{tone}, \text{noise}\})$
    $m_i \sim \text{Geometric}(...)$ $\quad \leftarrow$ *Number of elements in source*
    for $j = 1,...,m_i$:
      $onset_{ij} \sim \text{Uniform}([0, length])$
      $duration_{ij} \sim \text{Uniform}([0, length])$
    condition(no overlapping elements)

    if $type_i = \text{tone}$: **toneSource**()
    else: **noiseSource**()

toneSource() :
  $\mu_i \sim \text{Uniform}([f_{min}, f_{max}])$
  for $j = 1,...,m_i$:
    $frequency_{ij} \sim \text{Normal}(\mu = \mu_i, \sigma_f^2)$
  $volume_i \sim \text{Exponential}(\lambda)$

noiseSource() :
  $\nu_i \sim \text{GaussianProcess}_f(\mu(f) = 0, K_1)$
  for $j = 1,\ldots,m_i$:
    $s_{ij} \sim \text{GaussianProcess}_{tf}(\mu(t,f) = \nu(f), K_2)$

Model

Our generative model is a probabilistic program that consists of two components:

1. A sampling procedure which generates a hierarchical symbolic description of a scene, S, in terms of sources. Given our immediate goal to formally explain a diverse set of classic ASA illusions, we define these source models in terms of the simplest primitive sound elements: pure tones and noises. This defines the prior distribution $p(S)$ over complete auditory scenes.

2. A stochastic renderer which uses this symbolic scene representation to sample an audio signal D, thus defining the model likelihood $p(D|S)$.

Given a sound waveform D, these components induce a posterior distribution over auditory scenes, $p(S|D)$.

Generative Model

Table specifies our full probabilistic program for auditory scenes; an example run is depicted in Figure 1. A scene description consists of one or more parameterized sources, which each emit a sequence of one or more sound elements. Our model includes two source models that vary by the *type* of sound element they produce: tones or noises. Furthermore, a source is constrained to produce only one element at any time, with the occurrence of each element described by its onset and duration. Besides these temporal variables, elements are additionally defined by source-dependent spectral variables: tone elements by their (constant) frequency, and noise elements by their time-varying spectrum.

We base the form of our source models on regularities in natural audio. In particular natural sounds tend to exhibit local correlations in both time and frequency (McDermott, Wrobleski, & Oxenham, 2011). We instantiate these correlations as resulting from generative source models. For tone sources, the log frequencies $f_1,\ldots,f_n$ of n elements emitted are drawn from a Gaussian distribution around the source mean μ. For noise sources, a mean *spectrum* ν is first drawn from a one-dimensional Gaussian process with kernel

$$K_1(f,f') \propto \exp\Big(-\frac{|f-f'|}{\ell^f}\Big) \tag{1}$$

Given this, each element's time-varying spectrum is sampled from a two-dimensional noiseless Gaussian process, with constant mean ν and covariance

$$K_2((t,f),(t',f')) \propto \exp\Big(-\frac{|t-t'|}{\ell^t} - \frac{|f-f|}{\ell^f})\Big) \tag{2}$$

The length scales $\ell^t = 0.8s$ and $\ell^f = 10\text{ERB}$ control the extent of the correlations and was chosen using the statistical analysis of natural sounds in (McDermott et al., 2011). These parameters are held constant across all the illusions that we model.

Renderer

Given this symbolic description of the auditory scene, the renderer generates the acoustic waveform 'produced' by each source. Tone elements are rendered into a sequence of sine waves with the appropriate frequency, onset, and duration. For noise elements, white noise is filtered to match the sampled time-varying spectrum, and windowed to the appropriate onset and duration. These source waveforms are summed to produce the auditory scene waveform.

To compute the likelihood of an observed sound given a sampled scene, the sampled waveform is compared to the observation under a Gaussian noise model. Rather than assuming noise on the waveform itself, the waveforms are converted to a gammatonegram, a time-frequency representation of sound that approximates the filtering properties of the human ear (Ellis, 2009; Glasberg & Moore, 1990). That is, the likelihood is the probability that the observed gammatonegram is a noisy measurement of the sampled gammatonegram:

$$p(D|S) = \mathcal{N}(G_D|\mu = G_S, \sigma^2) \tag{3}$$

To visualize posterior samples throughout the rest of the paper, we plot gammatonegrams as in Figure 2. The observation gammatonegram is plotted in purple with fully labeled axes, while the scene components are shown through gammatonegrams of each of the source waveforms.

Inference

As in other perceptual problems, inference in our model is difficult due to the many local optima that arise when inferring sound elements from raw audio, and to the combinatorially large number of different ways in which elements may

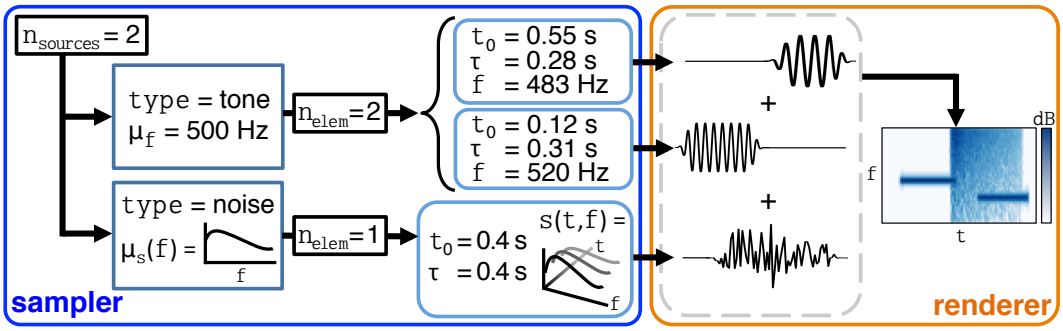

Figure 1: Process of sampling a gammatonegram from the generative model.

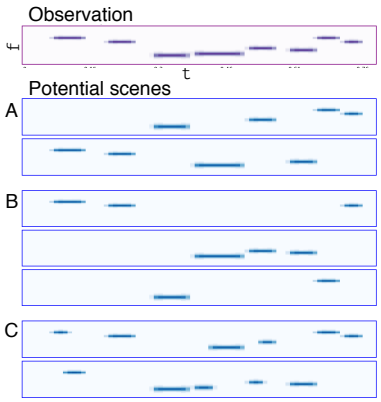

Figure 2: Infinitely many source combinations could produce a given sound. A, B, and C are each combinations in which multiple sources (each row) produce tones that sum to the observation.

group together. Here, inference is compounded by a further challenge: neither the number of sources nor the number of elements are known to our model in advance. Thus, inference involves searching a space of auditory scenes with varying dimensionality.

We address these challenges combining two contemporary tools. We first implement our model as a probabilistic program in the language WebPPL (Goodman & Stuhlmüller, 2014), and then sample scene-gammatonegram pairs from this program to train a deep neural network as a bottom-up feature detector. The architecture of this network was adapted from (Ren, He, Girshick, & Sun, 2015), developed for multiple-object detection in images.

Applied to novel sounds, the network returns bounding boxes to describe the onset and duration of multiple elements from the input gammatonegram, while classifying them as either tone or noise (Figure 3). For a tone element, this bounding box additionally provides information on the tone's frequency, while for a noise element we use onset and offset information to estimate the time-averaged spectrum for noises directly. Thus, the feature detector provides a list of complete candidate elements for a sound.

This list is then used in inference in two ways. First, MCMC is initialized with a random assignment of these elements to sources, constrained to not overlap or mix types within a stream. Second, when MCMC proposes to add new elements, these are drawn from a mixture distribution comprising noisy versions of the elements inferred by the network.

Results

We tested whether the model could qualitatively replicate a range of classic ASA illusions. We mainly focus on illusions involving perceptual 'filling-in', because they are most illustrative of the necessity for inferring the scene from the raw signal. However, we also examine tone-sequence illusion to demonstrate the model's capacity for grouping such stimuli. For each demonstration, the model inferred samples comprising the approximate posterior distribution. We used the resulting samples to simulate psychophysical experiments for comparison with human judgments. All audio recordings, accompanying gammatonegrams, and example posterior samples can be found at http://www.mit.edu/ mcusi/basa/index.html

Grouping in tone sequences

The basic grouping problem in sequential phenomena is demonstrated in Figure 2. An infinite number of sources could have combined to produce the observed audio. For instance, all of the tones could have been produced by a single source or they could be split up in several ways across two or three sources (A, B). Inferring tone elements from sound is also ill-posed, as multiple overlapping elements may combine to produce a long tone (C).

In a classic experiment, Tougas and Bregman (1985) interleaved rising and falling tone sequences, producing the 'X' pattern apparent in Figure 3. They presented listeners with subsets of the tone elements in the 'X' pattern and asked them to rate how clearly the subset resembled something they heard in the tone sequence. Listeners found it difficult to hear rising or falling trajectories in the mixture. Instead, listeners were strongly biased to hear the higher frequency tones as segregated from the lower frequency tones, producing two sequences that 'bounce' and return to their starting points. The generative model qualitatively replicates this preference for frequency proximity, with 90 percent of posterior samples containing this organization Figure 3.

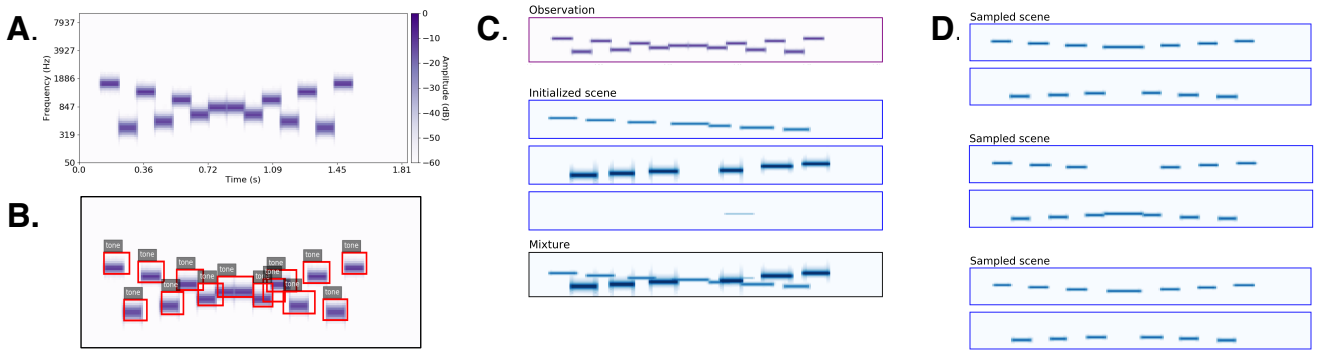

Figure 3: Stages of inference. A: gammatonegram of observed tone sequence, B: neural network bounding boxes computed for A, C: bottom-up initialization of MCMC based on bounding boxes in B, D: posterior samples.

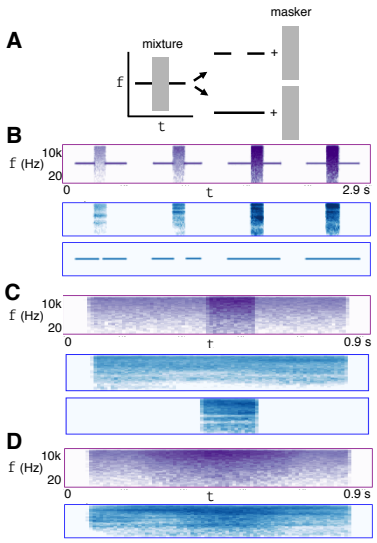

Figure 4: Model results for the continuity illusion, with tones and noise. A) Schematic of basic tone continuity illusion and possible interpretations. B) Inferred sources for four stimulus variants that vary in the relative intensities of tone and noise. Tone is inferred to be continuous when the noise is sufficiently intense to have masked it. C) Analogous effect for noise whose amplitude increases abruptly. D) A gradual amplitude modulation is instead attributed to a single source.

Perceptual 'filling-in'

When sources produce sounds that overlap in time and frequency, sufficiently intense sounds can obscure the presence of less intense sounds. That is, if a less intense sound is added to a sufficiently intense sound, the less intense sound will not be heard – a phenomenon termed 'masking' (Warren, Obusek, & Ackroff, 1972). In such cases, the addition of the less intense sound does not alter the peripheral auditory representation to a detectable extent. However, the perceptual interpretation can be modulated by context. For instance, a noise flanked by tones (Figure 4A) could equally well consist of two short tones adjacent to the noise, or a single longer tone overlapping the noise. Listeners hear this latter interpretation as long as the noise is intense enough to have masked the tone were it to continue through the noise(Warren et al., 1972), an effect replicated in our model (Figure 4B). In the model, the presence of an extended tone decreases the likelihood if the noise is not sufficiently intense, because the energy from the tone is not present in the gammatonegram. On the other hand, if the noise is sufficiently loud such that the tone does not affect the gammatonegram, inference of one over two tones can be understood via Bayes' Occam's razor. This illusion saliently demonstrates that basic 'parts' that we perceive are not explicit in the waveform.

We also tested whether the model can recapitulate trends in perceptual completion in two other domains. First, we tested the model on a variant of the continuity illusion described above involving only amplitude modulated noise. When an initially soft noise undergoes a sudden rise in intensity, listeners perceive the initial source as continuing unchanged behind a distinct, louder noise burst (Warren et al., 1972). In contrast, if the amplitude modulation occurs gradually, a single source is heard to change in intensity. In accordance with human listeners, the model strongly prefers two sources when the noise amplitude changes abruptly (over 1 ms), and almost never when the amplitude changes gradually (over 250 ms) (Figure 4C, D).

Analogous phenomena occur over the frequency spectrum, dubbed 'spectral completion'. McDermott and Oxenham (2008)'s basic paradigm for investigating spectral completion is shown in Figure 5A. Listeners heard a long masker noise, which overlapped with a brief target noise halfway through its duration. The spectrum of the target was ambiguous because the middle band of its spectrum could plausibly be masked by the masker. Listeners were asked to adjust the middle band of a comparison noise until it perceptually matched the target. To compare model inferences with human psychophysics, we measured the expectation of the spectrum level of the model's

inferred target in the middle band.

The human results and model predictions are plotted in Figure 5. In B, i and ii are control conditions that show that participants are able to match the spectrum level of the target if no masker is present. With iii and iv listeners attribute energy to the target in the center band in the presence of the masker, suggestive of spectral completion. v and vi show that the effect depends on the presence of the masker in the appropriate frequency range. The model follows these same trends. To rule out the possibility that participants were attempting to match the raw stimulus levels in Biii and iv, McDermott and Oxenham (2008) ran an additional experiment in which they varied the level of the masker and the unobscured portions of the target in opposite directions. They reasoned that if judgments were based on the actual stimulus level, then listeners would adjust the comparison towards the masker. Listeners instead show a complex dependence on these stimulus levels. When the masker was louder (Ci-iii), judgments followed the target level, as though listeners were inferring a smooth target spectrum. But when the target was louder, judgments ceased to follow the target, and instead only attributed as much energy to the target as could be masked by the masker. The model also closely follows these trends.

Sound recordings

Finally, we tested the generalization of the model to simple recorded audio. To choose the sounds, we compiled a bank of relatively simple single sources, consisting of natural sounds (sound textures and animal vocalizations) and pitched and unpitched percussion instruments. We then randomly mixed small sets of these recordings. To initialize inference, we found it necessary to increase the threshold for accepting the neural network's bounding box proposals (presumably because the sounds no longer are as faithful to the generative model with which the network was trained). Typical posterior samples for several natural scenes are shown in Figure 6. The model produces reasonable interpretations for these real-world sources. More examples may be found on our website.

Discussion

We presented a probabilistic program of auditory scenes based on simple source priors, aiming to formally account for a variety of classic ASA illusions. Inference in the model succeeds at qualitatively matching human perceptual organization in these illusions, as well as some simple audio recordings. Future work will look to broaden the scope of illusions tested and quantitatively match trends in human perception. Most immediately, we will look to expand on our results by comparing our model to quantitative data on temporal continuity (Warren et al., 1972) and performing model lesion studies to better understand the contribution of each component. Furthermore, tone sequences comprise a large portion of the auditory perceptual organization literature, and we will more exhaustively explore the space of those illusions going forward.

Another large class of ASA effects not explored here involve harmonic sounds, which have energy at frequencies that are integer multiples of some fundamental frequency. Sound frequencies with this relationship tend to be heard as produced by a single source. The current source models, while chosen with the aim of parsimoniously explaining a range of ASA phenomena, are clearly insufficient to address these illusions. A source model of harmonic sounds will also be necessary to eventually explain natural scenes. For instance, when building our set of natural scenes it was necessary to exclude sounds with strong harmonic structure, which typifies many animal vocalizations and musical instruments. Thus, future work will include expanding the source models to harmonic sources, potentially building on the Gaussian process framework explored here. For instance, a harmonic source may be thought of as composing a time-varying periodic source with a time-varying filter (akin to the current noise model).

One contribution of this model to perception more broadly is its use of the audio waveform as input to a structured generative model. Previous work on perceptual grouping in vision (Froyen, Feldman, & Singh, 2015) instead use symbolic atoms as the input to their models. However, as mentioned above, scene analysis fundamentally involves inference of the lower level 'parts' themselves from raw sensory input. Using the audio waveform as input means that our model can handle this basic problem. It can also be applied to arbitrary waveforms, including natural sounds, thus allowing broader and more comprehensive comparisons with perception. Waveform-based inference necessitated combining bottom-up neural network initializations with MCMC inference of scene structure. This combined architecture, necessary to implement our computational level analysis of perception, may have implications for mechanistic accounts of ASA, as speed is a significant constraint on perceptual systems.

Acknowledgments

Work funded by NIH grant R01-DC014739-01A1 and a McDonnell Scholar Award to JHM.

References

Bregman, A. S. (1990). *Auditory scene analysis*. MIT Press.

Ellis, D. P. W. (2006). Model-based scene analysis. In D. Wang & G. J. Brown (Eds.), *Computational auditory scene analysis: Principles, algorithms and applications.* Piscataway, NJ: IEEE Press.

Ellis, D. P. W. (2009). *Gammatone-like spectrograms.* Retrieved from `http://www.ee.columbia.edu/~dpwe/resources/matlab/gammatonegram/`

Froyen, V., Feldman, J., & Singh, M. (2015). Bayesian hierarchical grouping: Perceptual grouping as mixture estimation. *Psychological review*, *122*(4), 575.

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1-2), 103–138.

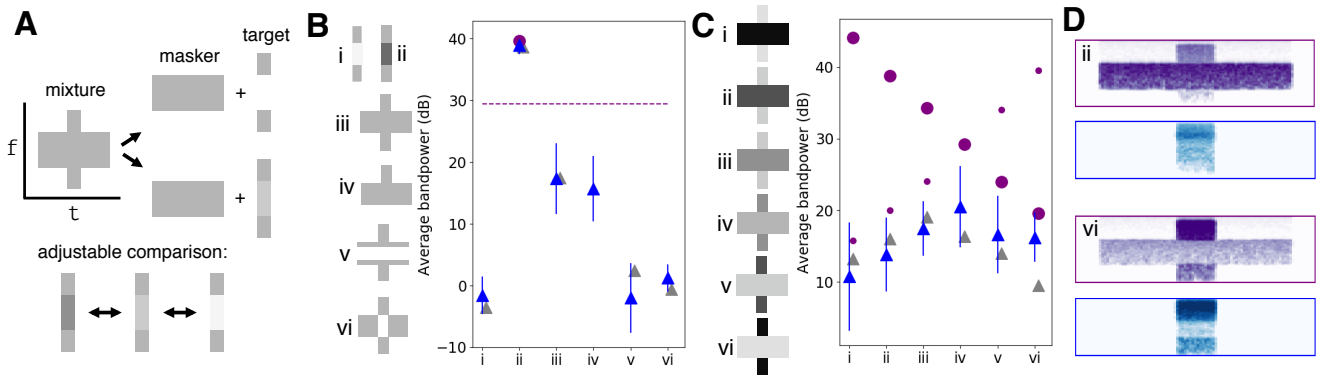

Figure 5: Spectral completion. Purple: stimulus, blue: model, gray: human data from McDermott and Oxenham (2008). A) Stimulus schematic. B) Dotted line: spectrum level corresponding to neutral gray in the schematic. Large dot: spectrum level of the dark band in ii. Here and in C, triangles plot spectrum level matched by humans in middle band of comparison. C) Small dots: unobscured target level. Large dots: masker level. D) Example posterior samples.

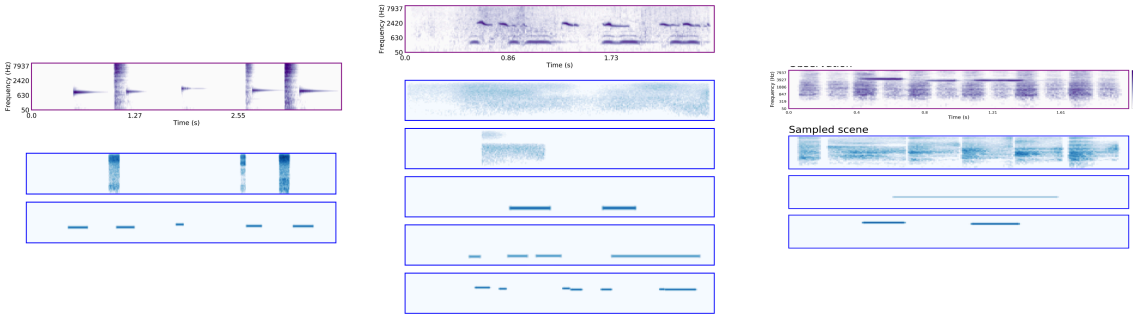

Figure 6: Examples of generative inference for natural auditory scenes composed of recorded real-world sound sources.

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages*. http://dippl.org. (Accessed: 2018-1-31)

Krishnan, L., Elhilali, M., & Shamma, S. (2014). Segregating complex sound sources through temporal coherence. *PLOS Computational Biology*, *10*, e1003985.

McDermott, J. H., & Oxenham, A. J. (2008). Spectral completion of partially masked sounds. *Proceedings of the National Academy of Sciences*, *105*(15), 5939–5944.

McDermott, J. H., Wrobleski, D., & Oxenham, A. J. (2011). Recovering sound sources from embedded repetition. *Proceedings of the National Academy of Sciences*, *108*(3), 1188–1193.

Mill, R. W., Bohm, T. M., Bendixen, A., Winkler, I., & Denham, S. L. (2013). Modelling the emergence and dynamics of perceptual organisation in auditory streaming. *PLOS Computational Biology*, *3*, e1002925.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).

Tougas, Y., & Bregman, A. S. (1985). Crossing of auditory streams. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(6), 788.

Turner, R. E. (2010). *Statistical models for natural sounds*. Doctoral dissertation, University College London.

Warren, R. M., Obusek, C. J., & Ackroff, J. M. (1972). Auditory induction: Perceptual synthesis of absent sounds. *Science*, *176*(4039), 1149–1151.

Yates, T., Larigaldie, N., & Beierholm, U. (2017). A nonparametric bayesian prior for causal inference of auditory streaming. In *Proceedings of the 39th annual conference of the cognitive science society* (pp. 1381–1386). Austin, TX: Cognitive Science Society.