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 unclear. We present a Bayesian model based on representations of simple acoustic sources, for which a 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 models; Auditory scene analysis; Auditory perception; Probabilistic programs; 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. 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, commonly termed auditory scene analysis (ASA, Bregman (1990)). ASA is 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. However, at present, we lack a formal account of these auditory illusions, let alone everyday auditory scenes. 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.

Model

Our model is a Bayesian probabilistic program (Goodman & Stuhlmüller, 2014), expressing uncertainty over both continuous and structural latent variables. The model comprises:

1. **Prior**, $p(S)$: A sampling procedure generates a hierarchical symbolic description of a scene, S. S 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. These source types were chosen because they can describe the majority of stimuli in the ASA literature. Element parameters are drawn from Gaussian processes, with parameters set to instantiate local correlations in time and frequency that are present in natural sounds (McDermott, Wrobleski, & Oxenham, 2011).

2. **Likelihood**, $p(D|S)$: A stochastic renderer uses this symbolic scene representation to sample an audio signal, represented as a gammatonegram D. This time-frequency representation of sound approximates the filtering properties of the human ear (Ellis, 2009; Glasberg & Moore, 1990). To compute the likelihood, a sampled gammatonegram is compared to the observation under a Gaussian noise model.

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

Inference

Inference is difficult due to the many local optima that arise when parsing elements from raw audio, and to the combinatorially large number of assignments of these elements into sources. Also, as neither the number of sources nor elements are known in advance, inference involves searching a space of auditory scenes with varying dimensionality.

We address these challenges combining two contemporary tools: we implement our model as a probabilistic program in the language WebPPL (Goodman & Stuhlmüller, 2014), and then sample (S,D) 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, and Sun (2015), developed for multiple-object detection in images. Applied to novel sounds, the network returns bounding boxes for multiple elements along with their type. These candidate elements are used to initialize and guide MCMC inference.

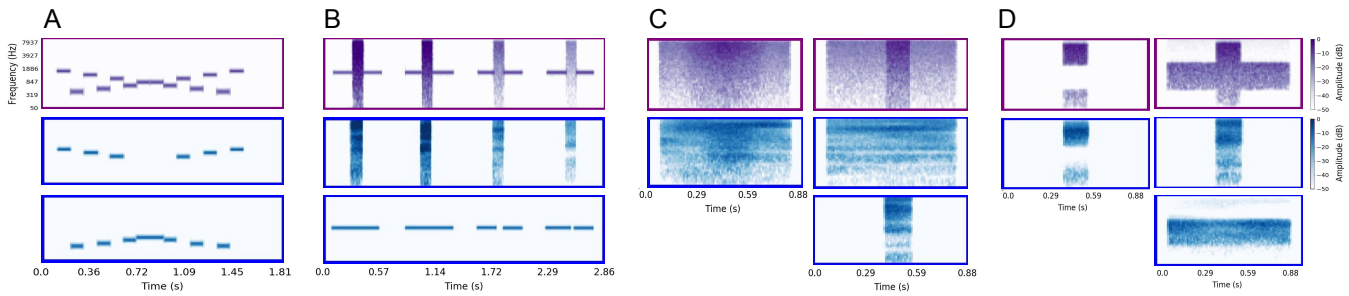

Figure 1: Top row shows gammatonegram of observed sound. Lower rows show sources rendered from one posterior sample.

Results

We tested whether the model could qualitatively replicate a range of classic ASA illusions. For each illusion, the model inferred samples comprising the approximate posterior distribution. We also tested the generalization of the model to recorded audio that was derived from a bank of simple natural sounds. All audio recordings, accompanying gammatonegrams, and example posterior samples can be found at `http://mit.edu/mcusi/www/basa-ccn`.

Grouping in tone sequences

In a classic experiment, Tougas and Bregman (1985) interleaved rising and falling tone sequences, producing an 'X' pattern. 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% of posterior samples segregating tones into higher- and lower-frequency ranges (Figure 1A).

Perceptual 'filling-in'

When sources produce sounds that overlap in time and frequency, sufficiently intense sounds can obscure the presence of less intense sounds – 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 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), and posterior samples from our model reliably reproduce this pattern (decreasing noise amplitudes in Figure 1B).

We also tested whether the model can recapitulate trends in perceptual completion in two other illusions. First, we tested the model on a variant of the continuity illusion described above involving only amplitude modulated noise. Like human listeners, the model infers a quiet source continuing behind a louder source only when the amplitude is modulated suddenly (Figure 1C). Second, analogous phenomena occur over the frequency spectrum, dubbed 'spectral completion' (McDermott & Oxenham, 2008). When a masker is present, the model infers energy at masked frequencies belonging to a short target sound (Figure 1D), in accordance with listeners.

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

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

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

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.

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.

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