Explaining Human Auditory Scene Analysis Through Bayesian Clustering

Nathanael Larigaldie (nathanael.c.larigaldie@durham.ac.uk)
Psychology Department
Durham University, Durham, UK

Ulrik Beierholm (ulrik.beierholm@durham.ac.uk)
Psychology Department
Durham University, Durham, UK

Abstract:

The way auditory stimuli are being processed to form perceptual unitary or segregated groups of sounds is still an ongoing discussion in the Auditory Scene Analysis literature. Mechanistic approaches to model this phenomenon have been somewhat successful but are often overly complicated and constrained to specific paradigms. Our approach is that of simplicity. We have previously proposed a higher-level source inference model in the Bayesian statistical framework that only implements a few simple but sensible rules applied to the stimuli's statistics. Yet, it still captures results from behavioral data (Yates, Larigaldie, & Beierholm, 2017). We have expanded on this model to show its ability to adapt to a wider range of well-known perceptual auditory phenomena. Several original experiments have also been conducted to explore a broader range of stimuli statistics. Our model's responses give insight into possible underlying processes in the brain that could provide a guide towards more behavioral experiments or medical exploration.

Keywords: Auditory Scene Analysis; Causal inference; Bayesian modeling; Perception; Audition;

Introduction

While Gestalt psychology has mainly focused on object grouping in the visual modality, a lot of auditory streams segregation and combination phenomena have been described (A. S. Bregman, 1994).

However, across all modalities it is still unclear how our perceptual systems can cope with both omnipresent uncertainty and a virtually infinite number of perceptual cues to treat, order and categorize in real time. Uncertainty in high-level perception is usually successfully modeled using the Bayesian framework (Körding et al., 2007) (Trommershäuser, Körding, & Landy, 2012). But as the number of perceptual cues increases linearly, the amount of possible clusters explodes factorially. As a result, most Bayesian models cannot be applied to more ecological environments as they are limited to a very low number of perceptual cues before calculations become intractable.

On the other hand, lower-level mechanistic approaches can be less limited in terms of number of percepts to be considered (for a review, see (Bee & Micheyl, 2008), but usually produce complex and situational models. Furthermore, from a cognitive perspective, they are hard to interpret in terms of meaningful brain functions.

We have been developing a non-parametric Bayesian model (Yates et al., 2017). The aim of our model is to tackle limitations from both approaches by being able to consider a potentially unlimited number of perceptual objects in reasonable time while not sacrificing simplicity nor interpretability. Furthermore, it is designed to be abstract enough to easily incorporate new perceptual cues, and to be usable across modalities.

Briefly, the model is a Bayesian clustering algorithm sequentially treating perceptual cues in order to infer the probability that they were produced by a common source, via dimensional proximity and parsimony of hypotheses.

The model assumes that given a source, all perceivable stimuli being created by this source should either have close characteristics on every dimension, or would require some time to change its state. That is, it is unlikely that a source could create two very different stimuli in a very short time. This implies that inference over such structure can be done by clustering of percepts. For example, if two sounds with frequencies F1 and F2 are produced by the same source, the pitch cannot change infinitely fast as an oscillator would require infinite impulse of energy to change its frequency discontinuously. This assumption can in general be summarized by proximity over a K-dimensional plane – with K being the number of perceptual cues considered for each stimulus.

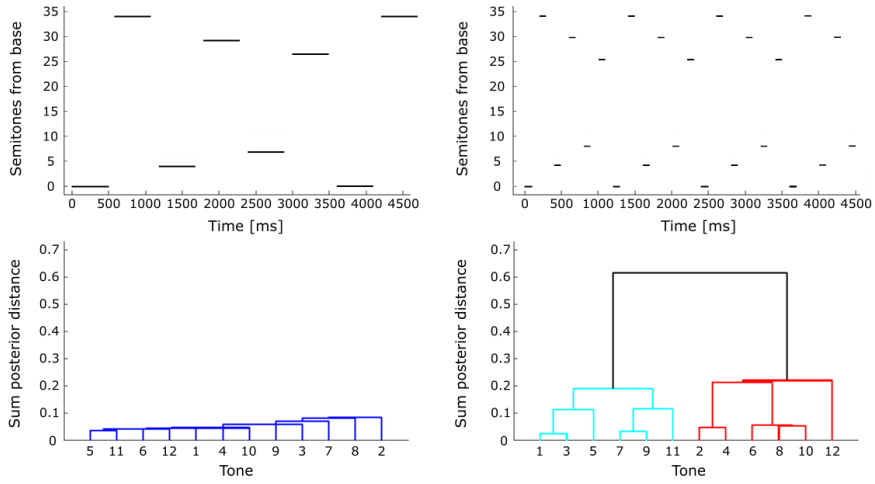

Figure 1: Stimuli used in the second experiment from Bregman & Campbell (1971). A slow sequence is perceived as a single stream, while a faster sequence in split in two. Stimuli are shown at the top, bottom is dendrogram tree-plots based on the posterior distribution over clustering. Across dendrograms, a unique color is assigned to clusters with more than 50 percent distance from other clusters

On top of this generative process, our model also assumes parsimony in the number of plausible clusters. This is done by introducing a non-parametric prior in the form of a Chinese Restaurant Process (Aldous, 1985), gradually decreasing the probability of considering a new source plausible as more sounds have already been assigned to previous sources.

Implementing only these two reasonable assumptions is enough to successfully reproduce several well-known auditory phenomena. Original experimental data were also collected to further explore the model's predictiveness.

Methods and Results

Model specifications

The first aforementioned assumption can be modeled with the following generative process:

$$p(S_n|C_n = C_{n-t}, S_{n-t}) = \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{\left(\frac{\Delta c}{\Delta t}\right)^2}{2\sigma^2}}$$

Where S_n is sound number n, C_n the cluster it belongs to (in other words: the source that caused it), Δc the difference in the considered characteristic (for instance, frequency), Δt the difference in time between two sounds' onsets and σ is a constant that can be fitted to participants' responses. It follows that a source is most likely to cause stimuli whose characteristics change slowly over time. Indeed, as $^{\Delta c}/_{\Delta t}$ increases, the probability of the newer sound to be in the same cluster as the previous sound decreases with a normal decay.

The second assumption is modeled by a non-parametric prior weighing these probabilities according to the number of sounds already in each cluster:

$$p(C_N = i|C_1 \dots C_{N-1}) = \frac{n_i}{(N-1)+\alpha}$$

when cluster i has already been inferred, and:

$$p(C_N = i|C_1 \dots C_{N-1}) = \frac{\alpha}{(N-1)+\alpha}$$

when none of the previous clusters is equal to i.

n_i is the number of sounds in cluster i, N is the total number of sounds considered and α is a constant that can be fitted to participants' responses. It follows that as more and more sounds are being considered, the probability of assignment to a new cluster decreases. On top of this, the model follows a *rich get richer* property, as clusters already comprised of many tones have a higher chance of getting more tones than clusters with fewer tones. Taken together, these properties can be considered as an implementation of Ockham's razor. For details of implementation see Yates et al. (2017).

Phenomena reproduction

A number of phenomena can be replicated by the model, but we will here only present two using auditory frequency as sensory cues.

The first phenomenon is the second experiment taken from Bregman & Campbell (1971), highlighting how the speed of presentation affects perception of streams of tones. Behavioral data shows that faster tones lead to an increased probability of subjects reporting two

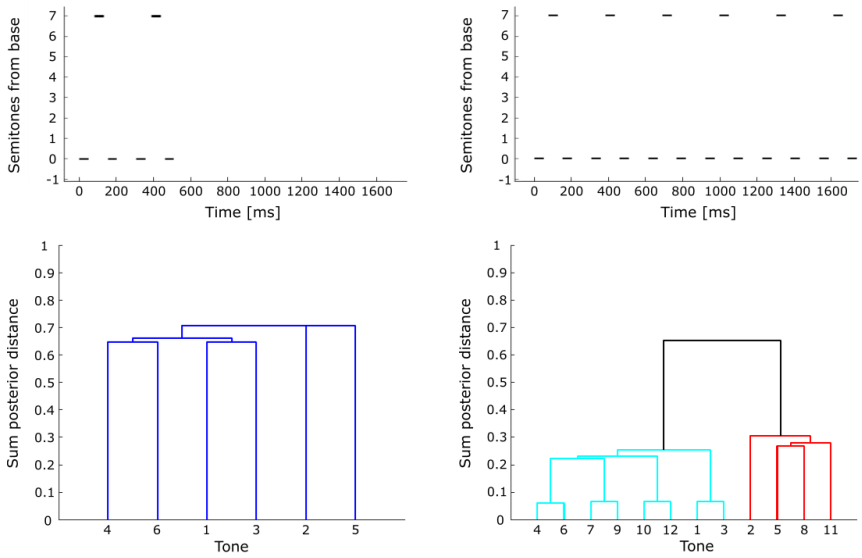

Figure 2: Stimuli used in the second experiment from Bregman (1978). A short sequence is perceived as a single cluster, while a longer sequence is split into two clusters. Stimuli are shown at the top, bottom is dendrogram tree-plots based on the posterior distribution over clustering.

perceived streams of sounds rather than one. Figure 1 shows how our model successfully captures this. The model likelihood term constrains how fast streams can change in frequency, hence too fast changes makes two streams more likely. The "slow sequence" had 100ms ISIs, 500ms tone duration and pitch differences of [0 4 8 26 30 34] semitones from the lowest tone. The "fast sequence" reduced the tone duration to 100ms.

The second phenomenon is taken from Bregman (1978) and shows that auditory streaming is cumulative. Sequences of tones that split into two perceived streams may initially be perceived as one stream. Figure 2 shows how our model successfully captures this. The non-parametric prior makes a single stream more likely when little information has been received. The "short sequence" has 26.6ms ISIs, 7 semi-tones pitch differences with two repetitions. The "long sequence" was instead repeated eight times.

Novel experiments

If auditory scene analysis is indeed a process of perceptual clustering of auditory stimuli into separate streams, then we would expect subjects to be able to cluster more than two sets of stimuli, and consequently perceive more than two streams. A set of novel experiments using a new paradigm have been designed in order to explore the influence of several sensory cues on the formation of auditory streams, and a potentially higher maximum number of perceived streams. Only one of these experiments, using frequency as a sensory cue, will be presented here.

Figure 3 is a schematic representation of the stimuli. The key realization is that subjects lose the order information of tones when assigned to different streams. Therefore, subjects should not detect a difference between sequences 1 and 2 when medium tones do not share a stream with either low or high tones, as long as they are ignorant as to which tone started the medium stream. This is insured by introducing a general fade in effect at the start of every sequence.

Results in Figure 4 show that, as expected, higher frequency differences significantly decreased participants' capacity to tell the two sequences apart, implying that the middle tones are perceived as a separate cluster to either the high or low tones. This strengthens the argument that auditory streaming formation is being influenced by proximity in frequency space, and that humans may hold 3 or more auditory streams simultaneously.

Overall, we show how several aspects of auditory scene analysis can be modeled based on very few normative assumptions. Experiments support the qualitative predictions of the model, an aspect that future work will expand on.

Acknowledgments

The authors are grateful for generous support from the Leverhulme foundation (UB) and Durham University (NL). Previous development of the model was based on work with Timothy Yates.

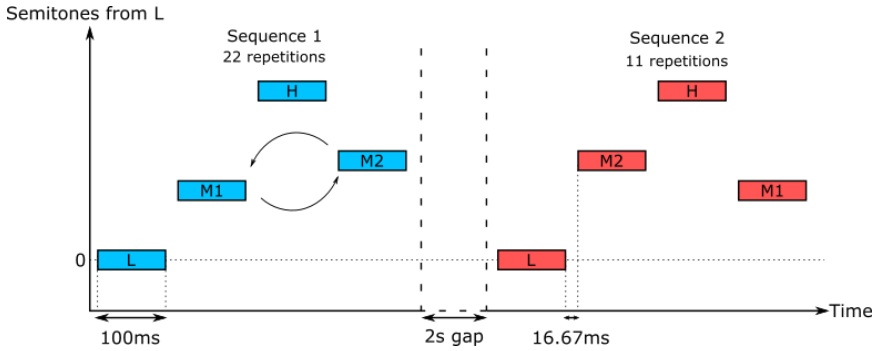

Figure 3: Schematic representation of stimuli used. Medium tones inversion was only present in half of the trials.

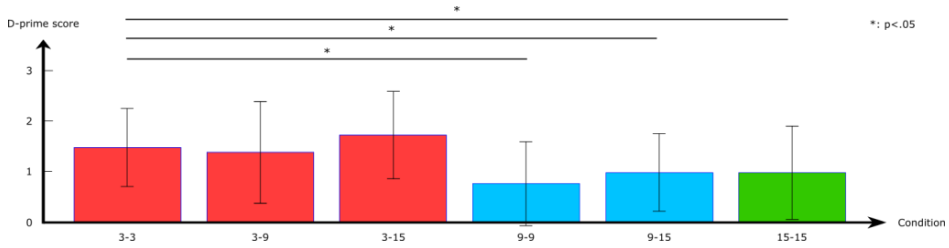

Figure 4: Mean d-prime scores of participants according to conditions. Error bars are standard deviations. First numbers in conditions indicate a difference in semitones from L to M1, second number is the difference between M2 and H. The difference in frequency between M1 and M2 was always 3 semi-tones.

References

Aldous, D. J. (1985). Exchangeability and related topics. In D. J. Aldous, I. A. Ibragimov, J. Jacod, & P. L. Hennequin (Eds.), *École d'Été de Probabilités de Saint-Flour XIII — 1983* (pp. 1–198). Springer Berlin Heidelberg.

Bee, M. A., & Micheyl, C. (2008). The "Cocktail Party Problem": What Is It? How Can It Be Solved? And Why Should Animal Behaviorists Study It? *Journal of Comparative Psychology (Washington, D.C.: 1983)*, *122*(3), 235–251. https://doi.org/10.1037/0735-7036.122.3.235

Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press.

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, *89*(2), 244–249.

Bregman, Albert S. (1978). Auditory streaming is cumulative. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(3), 380–387. https://doi.org/10.1037/0096-1523.4.3.380

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal Inference in Multisensory Perception. *PLOS ONE*, *2*(9), e943. https://doi.org/10.1371/journal.pone.0000943

Trommershäuser, J., Körding, K. P., & Landy, M. S. (2012). *Sensory Cue Integration*. https://doi.org/10.1093/acprof:oso/9780195387247.001.0001

Yates, T., Larigaldie, N., & Beierholm, U. (2017). A Non-Parametric Bayesian Prior For Causal Inference Of Auditory Streaming. *BioRxiv*, 139188. https://doi.org/10.1101/139188