35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Bayesian inference in auditory scenes

Mounya Elhilali, *Member, IEEE*

Abstract— The cocktail party problem is a multi-faceted challenge which encompasses various aspects of auditory perception. Its processes underlie the brain's ability to detect, identify and classify sound objects; to robustly represent and maintain speech intelligibility amidst severe distortions; and to guide actions and behaviors in line with complex goals and shifting acoustic soundscapes. Here, we present a perspective that considers the powerful Bayesian inference as a unifying framework to integrate the role of sensory cues as well as stimulus-driven priors and top-down schemas including attention.

I. INTRODUCTION

Auditory scene analysis is often described in terms of the cues and processes that govern it: simultaneous vs. sequential grouping cues; bottom-up vs. top-down processes. Grouping cues describe the principles used by the auditory system to determine whether different acoustic features emanate from a common sound source or should belong to the same perceptual auditory object or stream. They define the concurrent organization of sound features based on pitch, synchrony and spectral structure (Bird & Darwin, 1997; Darwin, 1997; Roberts & Bailey, 1996); as well as sequential cues to organize elements over time based on –for instance- frequency separation, presentation rate, modulation patterns, and spatial locations (Darwin & Carlyon, 1995; Moore & Gockel, 2002). These organizational principles are complemented with top-down processes set by context; expectations and attentional state in guiding how an auditory scene is perceived (Bregman, 1990).

While studies of grouping cues and processes have greatly benefited our understanding of the phenomenon of auditory scene analysis, there is still a lack of integrative theories that provide a framework of how the auditory system orchestrates all these players together to guide our perception of the surrounding soundscape. Specifically, the role of grouping cues and organizational principles has often been interpreted in the context of a rule-based Gestalt segregation regime. The quest for defining the parameters of these grouping cues and neural underpinning of segregation rules has often driven the study of the biological and perceptual correlates of auditory scene analysis. In this context, the role of top-down processes, particularly goal-directed attention, is often interpreted abstracted from that of sensory-driven segregation cues.

Here, we present a perspective that considers the powerful Bayesian inference as a an alternative unifying framework to integrate the role of sensory cues as well as stimulus-driven priors and top-down schemas including attention. Bayesian inference has been widely applied in other modalities particularly vision and sensorimotor (Kwon & Knill, 2013; Lee & Mumford, 2003; Moreno-Bote, Knill, & Pouget, 2011). It has had limited treatment in the auditory literature (e.g. see (Grossberg, Govindarajan, Wyse, & Cohen, 2004; Winkler, Denham, & Nelken, 2009)). Here, we argue that a Bayesian inference framework is a powerful tool to encompass the role of bottom-up and top-down processes in auditory streaming; to predict their interactions in biasing auditory perception in an optimal or quasi-optimal fashion. This framework provides a proper computational scheme for integrating the uncertainty surrounding sensory information, nondeterministic neural representations of incoming cues as well as malleability of prior knowledge, making a probabilistic interpretation appropriate. We discuss the use of such framework in the context of auditory scene analysis, and provide support for a number of physiological and perceptual studies that provide support for such scheme in processes of auditory scene analysis.

II. SCENE ANALYSIS FRAMEWORK

In presence of a multitude of often ambiguous sensory cues and cognitive factors, perception of auditory scenes can be thought of as an inference process (Knill & Richards, 2008); where the system

*This research was supported by NSF grant IIS-0846112, NIH grant 1R01AG036424 and ONR grants N000141010278 and N00014-12-1-0740.

M. Elhilali with the Faculty of the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, USA, (email: mounya@jhu.edu)

integrates all information (both sensory and cognitive) to come up with likely interpretations of the soundscape. This process is not solely driven by the external stimuli surroundings us. Rather, it integrates both sensory cues with 'internal' top-down information built from learned schemas, prior knowledge and behavioral goals.

Initially, this process maps the acoustic input onto a rich representation that encodes detailed parameters about the signal. Such mapping of sound is parameterized by the acoustic dimensions explicitly represented along different stages of the auditory system starting at the periphery all the way to auditory cortex (see section II.A). The neural mapping of such representation is therefore an estimate of the likelihood state of the soundscape based on its acoustic attributes. This mapping needs to be integrated by a set of priors about the soundscape before inferring knowledge about the state of world. In the current study, we consider priors in the form of bottom-up statistics that are acquired throughout the process of scene analysis (a type of dynamic prior) as well as a selective bias driven by goal-directed attention (Whiteley & Sahani, 2012). The integration of the likelihood estimate with the priors can be optimally through a Bayesian relationship.

In addition, it is also important to acknowledge the dynamic nature of the acoustic space. Therefore, this integration process via Bayesian inference is not a static one; but varies with the changing nature of the acoustic environment. This issue is discussed further in section II.B.

A. Cortical multidimensional representation

Auditory cortex is a natural neural locus of interest in defining the proper representation of the acoustic scene (Nelken, 2004; Sharpee, Atencio, & Schreiner, 2011). Reviewing the literature of cortical sound processing, neurons in the primary auditory cortex appear to be selective not only to the spectral energy at a given frequency, but rather to the specifics of the local spectral shape such as its bandwidth (Schreiner & Sutter, 1992; Schreiner, 1995), symmetry (Versnel, Kowalski, & Shamma, 1995), and dynamics or temporal modulations (Lu, Liang, & Wang, 2001; Miller, Escabi, Read, & Schreiner, 2002; Schreiner, Mendelson, Raggio, Brosch, & Krueger, 1997). The resulting representation of sound in A1 is a *multidimensional* representation, which can be thought of as an array of filters. Cortical filter responses, also called spectro-temporal receptive fields (STRFs) vary along at least three dimensions: (1) *Best frequencies* (BF) that span the entire auditory range; (2) *Bandwidths* that span a wide range from very broad (2-3 octaves) to narrowly tuned (< 0.25 octave); (3) *Dynamics* that are limited to few Hertz (1-30 Hz).

Mathematically, the mapping from a single-dimension acoustic waveform to a higher dimensional cortical space can be captured via a series of transformations, depicting two main operations: (1) an *early* transformation that captures cochlear and midbrain processing. It transforms the one-dimensional acoustic stimulus to an auditory time-frequency spectrographic representation; (2) a *cortical* transformation; so-named because it reflects the more complex spectrotemporal analysis presumed to take place in mammalian primary auditory cortex. While this formulation is not strictly biophysical and bypasses numerous interesting neural transformations occurring in pre-cortical stages; it abstracts an interpretation that is sufficient to understand perceptual phenomena such as stream segregation and auditory object formation

It is important to note that the use of this cortical space as our operating platform is not a statement to reduce the role of auditory cortex to a network of feature detectors (Nelken & Bar-Yosef, 2008). Rather, it is the culmination of a series of transformation undergone by the one-dimensional acoustic input to highlights numerous intricate details about the sound, both spectrally and temporally. Such attributes are believed to define the dimensions where a representation of complex acoustic scenes is highlighted better.

B. Recursive Bayesian estimation

The mapping of sound into a higher dimensional cortical space allows different features to occupy non-overlapping parts of perceptual space, hence enhancing discrimination between different auditory objects in the scene. This operation is reminiscent of classification and regression techniques such as support vector machines and kernel-based classifiers (Duda, Hart, & Stork, 2000; Herbrich, 2001). The rich cortical representation can generate predictions which can be fed back to reconcile with incoming inputs subject to known constraints of auditory objects. This premise can be thought of as a dynamic inference process that tracks the evolution of sound features in the cortical space (Elhilali & Shamma, 2008). The scheme models the underlying dynamical system as Markov-chain model where the future state at time $t + 1$; depends on its current state, along with the stimulus input at time t. By keeping the system relationships linear, the optimal solution would be a Kalman filter estimation (Chui & Chen, 1999). In the implementation of such scheme presented in (Elhilali & Shamma, 2008), one can combine this tracking stage with a clustering process, where the system tracks multiple streams based on how

well each of their predictions match with the incoming input. While the implementation based on Kalman-filtering is over simplified and presents a number of limitations, it still provides a tractable and optimal framework for predicting the dynamic behavior of the cortical space as sounds evolve over time and adjust to the changing nature of the acoustic scene.

III. Adaptation of the Sensory Space

In addition to the inherent dynamics of the acoustic input; and therefore its neural representation, the cortical space itself has been shown to undergo dynamic adaptation driven by attentional goals. Recent neurophysiological data has shown that cortical receptive fields undergo a *rapid* plasticity which changes their tuning almost on the fly in response to changing behavioral goals. The STRFs adapt their spectral and temporal properties in order to enhance behavioral performance, which can monitored through external (reward or aversive) feedback signals.

Specifically, an animal engaged in a spectral task enhances its cortical response at the target tone location (Fritz, Elhilali, Klein, & Shamma, 2003; Fritz, Elhilali, & Shamma, 2007), while an animal engaged in detecting a temporal event such as gap changes its STRF temporal dynamics to enhance its temporal response (Fritz, Elhilali, & Shamma, 2005). Overall, the spectral and temporal nature of target/reference cues dictate the specific form of STRF change, but only if the animal is behaviorally engaged. No such changes were observed in naïve animals, or in trained animals with poor behavioral performance.

Largely, these results suggest the presence of attention-triggered adaptive changes in primary auditory cortex that can swiftly change STRF shape by transforming receptive fields to enhance figure/ground separation. Such changes provide evidence of integration of selective priors with the acoustic mapping of sound cues in order to bias the representation in order to perform a task of interest. Whether this integration is done 'optimally' and provides evidence for a Bayesian framework remains to be proven; though there is no evidence to counter this assertion of optimality. Still, attention does undoubtedly play a role in the scene analysis process in the brain (Fritz, Elhilali, David, & Shamma, 2007; Shinn-Cunningham, 2008); and can be shown to improve scene analysis in engineering applications as well (Patil & Elhilali, 2012).

IV. Conclusion

Here, we postulate that auditory scenes are parsed via a balancing act between three separate components: sensory information emerging from the soundscape, bottom-up priors and top-down attentional demands. As a result of the push-pull between these different factors, the cortical representation undergoes adaptation in order to heighten the system's noise robustness by enhancing figure/ground separation (increasing signal to noise ratio), and boosts its computational efficiency by constricting processing of redundant and irrelevant backgrounds in the acoustic scene. This scheme can be translated into a statistical optimization, whose solution tracks the cortical state which underlies specific percepts and behaviors. Optimality is defined in a Bayesian manner which reconciles the sensory evidence with priors conveying bottom-up and top-down controls.

V. References

1. Bird, J., & Darwin, C. J. (1997). Effects of a difference in fundamental frequency in separating two sentences. In A. Palmer et al. (Eds.), *Psychophysical and physiological advances in hearing* (pp. 263-269). London: Whurr.
2. Bregman, A. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, Mass.: MIT Press.
3. Chui, C., & Chen, G. (1999). *Kalman filtering with real time applications*. New York, NY: Springer-Verlag.
4. Darwin, C. J. (1997). Auditory grouping.*1*(9), 327-333.
5. Darwin, C. J., & Carlyon, R. P. (1995). Auditory grouping. In B. C. J. Moore (Ed.), *Hearing* (pp. 387-424). Orlando, FL: Academic Press.
6. Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification,* Wiley.
7. Elhilali, M., & Shamma, S. A. (2008). A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. *Journal of the Acoustical Society of America, 124*(6), 3751-3771.
8. Fritz, J., Elhilali, M., Klein, D. J., & Shamma, S. A. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature neuroscience.* 6(11):1216-1223.
9. Fritz, J., Elhilali, M., & Shamma, S. (2005). Active listening: Task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex. *Hearing Research, 206*(1-2), 159-176.

10. Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention--focusing the searchlight on sound. *Current Opinion in Neurobiology, 17*, 437-455.
11. Fritz, J. B., Elhilali, M., & Shamma, S. A. (2007). Adaptive changes in cortical receptive fields induced by attention to complex sounds. *Journal of Neurophysiology, 98*(4), 2337-2346.
12. Grossberg, S., Govindarajan, K. K., Wyse, L. L., & Cohen, M. A. (2004). ARTSTREAM: A neural network model of auditory scene analysis and source segregation.*17*(4), 511-536.
13. Herbrich, R. (2001). *Learning kernel classifiers: Theory and algorithms*. Cambridge, MA: MIT Press.
14. Knill, D. C., & Richards, W. (Eds.). (2008). *Perception as bayesian inference* Cambridge University Press.
15. Kwon, O. S., & Knill, D. C. (2013). The brain uses adaptive internal models of scene statistics for sensorimotor estimation and planning. *PNAS, 110*(11), E1064-73.
16. Lee, T. S., & Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex.*20*(7), 1434-1448.
17. Lu, T., Liang, L., & Wang, X. (2001). Temporal and rate representations of time-varying signals in the auditory cortex of awake primates.*4*(11), 1131-1138.
18. Miller, L. M., Escabi, M. A., Read, H. L., & Schreiner, C. E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex.*87*(1), 516-527.
19. Moore, B. C. J., & Gockel, H. (2002). Factors influencing sequential stream segregation.*88*, 320-333.
20. Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *PNAS, 108*(30), 12491-12496.
21. Nelken, I. (2004). Processing of complex stimuli and natural scenes in the auditory cortex.*14*(4), 474-480.
22. Nelken, I., & Bar-Yosef, O. (2008). Neurons and objects: The case of auditory cortex. *Frontiers in Neuroscience, 2*(1), 107-113.
23. Patil, K., & Elhilali, M. (2012). Goal-oriented auditory scene recognition. *Proceedings of the 13th Annual Conference of INTERSPEECH*.
24. Roberts, B., & Bailey, P. J. (1996). Regularity of spectral pattern and its effects on the perceptual fusion of harmonics.*58*(2), 289-299.
25. Schreiner, C. E. (1995). Order and disorder in auditory cortical maps.*5*(4), 489-496.
26. Schreiner, C. E., Mendelson, J., Raggio, M. W., Brosch, M., & Krueger, K. (1997). Temporal processing in cat primary auditory cortex.*532*, 54-60.
27. Schreiner, C. E., & Sutter, M. L. (1992). Topography of excitatory bandwidth in cat primary auditory cortex: Single-neuron versus multiple-neuron recordings.*68*(5), 1487-1502.
28. Sharpee, T. O., Atencio, C. A., & Schreiner, C. E. (2011). Hierarchical representations in the auditory cortex.*21*(5), 761-767.
29. Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences, 12*(5), 182-186.
30. Versnel, H., Kowalski, N., & Shamma, S. A. (1995). Ripple analysis in ferret primary auditory cortex. III. topographic distribution of ripple response parameters. *J.Aud.Neurosc., 1*, 271-286.
31. Whiteley, L., & Sahani, M. (2012). Attention in a bayesian framework. *Frontiers in Human Neuroscience, 6*, 100. doi: 10.3389/fnhum.2012.
32. Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: Predictive regularity representations and perceptual objects, Trends Cogn Sci.;13(12):532-40.