Chapter 5
Modeling the Cocktail Party Problem

Mounya Elhilali

Abstract Modeling the cocktail party problem entails developing a computational framework able to describe what the auditory system does when faced with a complex auditory scene. While completely intuitive and omnipresent in humans and animals alike, translating this remarkable ability into a quantitative model remains a challenge. This chapter touches on difficulties facing the field in terms of defining the theoretical principles that govern auditory scene analysis, as well as reconciling current knowledge about perceptual and physiological data with their formulation into computational models. The chapter reviews some of the computational theories, algorithmic strategies, and neural infrastructure proposed in the literature for developing information systems capable of processing multisource sound inputs. Because of divergent interests from various disciplines in the cocktail party problem, the body of literature modeling this effect is equally diverse and multifaceted. The chapter touches on the various approaches used in modeling auditory scene analysis from biomimetic models to strictly engineering systems.

Keywords Computational auditory scene analysis · Feature extraction · Inference model · Multichannel audio signal · Population separation · Receptive field · Source separation · Stereo mixture · Temporal coherence

5.1 Introduction

In everyday life, humans are constantly challenged to attend to specific sound sources or follow particular conversations in the midst of competing background chatter—a phenomenon referred to as the "cocktail party problem" (Cherry 1953). Whether at a real cocktail party, walking down a busy street, or having a conver-

M. Elhilali (✉)
Laboratory for Computational Audio Perception, Center for Speech and Language Processing, Department of Electrical and Computer Engineering, The Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA
e-mail: mounya@jhu.edu

J.C. Middlebrooks et al. (eds.), *The Auditory System at the Cocktail Party*, Springer Handbook of Auditory Research 60,
DOI 10.1007/978-3-319-51662-2_5

sation in a crowded coffee shop, sounds reaching a listener's ears from a particular sound source almost never exist in isolation. They persistently occur in the presence of other competing sources and distractors that form a person's acoustic environment. This soundscape needs to be organized into meaningful percepts, a process formally called "auditory scene analysis" (ASA) (Cherry 1957; Bregman 1990).

The ASA challenge is not confined to humans. Animals too, including mammals, penguins, songbirds, and fishes, have to overcome similar difficulties to navigate their complex auditory scenes, avoid predators, mate, and locate their newborns (Izumi 2002; Aubin 2004). A similar challenge also faces engineering systems, from military communication and surveillance devices to smart phones. Much like biological systems, these technologies have to navigate their soundscapes to pick out relevant sound sources (e.g., speech) while ignoring interference from the surround (Loizou 2013).

It is important to note that auditory scene analysis is not a monolithic process that is easily defined within an exact framework. Despite its seemingly effortless and intuitive nature, it is a multifaceted challenge that encompasses various processes. It underlies the brain's ability to detect, identify, and classify sound objects; to robustly represent and maintain these representations amidst severe distortions; to guide actions and behaviors in line with complex goals and shifting acoustic soundscapes; to adapt to and learn from the environment; as well as to integrate potentially multimodal sensory cues with information in memory, prior knowledge, and expectations to provide a complete understanding of the scene.

Given its multilayered nature, modeling auditory scene analysis has often been faced with a lack of a unified vision or agreed-on benchmarks that clearly define the objectives to be achieved. These goals have varied from tracking only relevant targets in a scene to a complete scan of all elements in the scene. Despite this complexity, interest in addressing the problem computationally is driven by a number of aims: (1) The ability of the brain to parse informative sensory inputs and track targets of interests amidst severe, unknown, and dynamic interferers is ultimately what gives the biological system its lead over state-of-the-art engineering systems. Modern technologies strive to replicate this intelligent processing in computational systems. This goal remains one of the holy grails of audio and speech systems (Wang and Brown 2006). (2) Computational models of ASA can provide a strong perspective in guiding neural and perceptual investigations of the problem in both humans and animals (Cisek et al. 2007). (3) Defining theoretical principles that govern aspects of the cocktail party problem will guide the field to develop better benchmarks to compare performance across systems as well as match up different implementations against the biological system for well-defined subtasks. (4) Mathematical ASA models can also act as a platform to examine commonalities across different sensory modalities and shed light on questions of optimality and efficiency of performance of the biological or engineering system under different operating conditions and for various tasks and environments.

5.2 Defining the Problem in the Cocktail Party Problem

Exploring the computational principles of the cocktail party challenge requires articulating the exact nature of the problem itself as well as considering the architecture of models that could tackle this challenge. As is the case with the study of any complex system, it is important to define the system's input to the task at hand and the nature of the output. At the input level, the most biological reasonable expectation of the input is the acoustic signal that impinges on the listener's ears either monaurally or binaurally. This corresponds to a single-microphone or two-microphone recording of the soundscape. Naturally, some engineering applications extend this notion to the possibility of multiple microphones, which expands the spatial resolution of the system, though taking it away from the realm of biological plausibility. This design takes full advantage of the role of spatial processing in analyzing complex soundscapes without limiting the engineering application to the same constraints of the biology. This view has indeed opened the door to many successful "solutions" to certain aspects of the cocktail party problem by using independent component analysis (ICA) (Hyvarinen et al. 2001) and other blind source separation (BSS) (Naik and Wang 2014) and beamforming techniques (van der Kouwe et al. 2001).

While choosing the number of input channels for a computational model is a relatively straightforward decision based on the desired fidelity of the model to biological processes, defining the actual goal for modeling the cocktail party problem is an ill-posed query (Haykin and Chen 2005; Lewicki et al. 2014). Brain mechanisms engaged in processing complex scenes can be interpreted at many levels. One level is as an *analysis* or *segmentation* goal that defines auditory scene analysis as a stream segregation problem, as envisioned by Bregman and Campbell (Bregman and Campbell 1971; Bregman 1981). In this view, the cocktail party problem describes the task whereby a listener is confronted with intertwined sound sequences from multiple sources and the brain must form separate perceptual streams (or "sound objects"). A computational implementation of this level focuses on segregating different sound sources based on their acoustic attributes, including their spatial location, and binding the appropriate elements together to represent the perceived streams in a multisource auditory scene. Although this definition identifies a goal for the computational algorithm, it maintains a significant degree of ambiguity when it comes to defining the exact relationship between the physical nature of the sound source and the perceived stream, which is not a one-to-one mapping.

Think, for instance, of a scenario in which the audience at a symphony hall is enjoying an orchestral concert. Although the sound sources can be discretely distinguished in acoustic space, the perceptual experience of this rich auditory scene is not trivial to segregate. Should the model distinguish woodwinds from the rest of the instruments or should it focus on flutes versus clarinets versus bassoons? Uniquely defining the granularity of this segregation task is simply impossible and ultimately depends either on the goals of the model/system, or—in the case of

modeling human behavior—on the specific task given to the listener along with any behavioral metrics. This subsequently raises questions as to the limits of incorporating information about the sources in the segregation process. Should the model have knowledge about what a flute or a clarinet sounds like?

More importantly, the segmentation of an auditory scene poses additional, larger, questions: should the segregation be confined to a two-stream problem consisting of segregating a foreground (or target) stream from the background that incorporates the entire remainder of the scene; or should the segregation truly represent "all" possible individual sound streams within the scene itself? When framed as a figure–ground segregation problem, the degree of complexity is greatly reduced. It is still incomplete, however, until additional processes (e.g., selective attention) are incorporated to help dictate what the target or foreground characteristics are. It also requires specifying the underlying priors as to "what" the target (or target class) is, what its attributes are, and whether there are descriptive or statistical models that define them.

Alternatively, one can take a different approach and cast the overall goal of the cocktail party model as arising from a *recognition* point of view. In this case, the objective is to provide a recognizable label of the soundscape. This view aligns with frameworks commonly employed in computer vision and traditions of visual scene perception (Riesenhuber and Poggio 2002; Xu and Chun 2009) and has found applications in many sound technologies and speech systems (Chen and Jokinen 2010). Such systems are developed to provide various informative descriptors about a given a scene; e.g. is human speech present in a recording? Which melody is playing right now? Can footsteps be tracked in a surveillance microphone? Is there an abnormal heart murmur in a stethoscope signal? Clearly, the range of information that can be potentially conveyed from an auditory scene can be limitless.

Existing technologies have successfully focused on particular aspects of this recognition task, especially recognizing a single target amidst interfering backgrounds such as human speech (Virtanen et al. 2012) or tune/melody recognition systems (Collins 2009). Alternatively, some systems focus on recognizing the environment that gave rise to the scene itself (Patil and Elhilali 2013; Barchiesi et al. 2015), while other systems target abnormal or unexpected events in a scene for surveillance and medical systems (Anemuller et al. 2008; Kaya and Elhilali 2013) or even attempt to learn from the surrounding soundscape (Buxton 2003).

Finally, another body of work interprets the cocktail party problem from a *synthesis* point of view, where the intent of the computational model is to synthesize individual streams following the segregation process (e.g., musical track separation [Collins 2009]), or extract cleaner or denoised versions of a target stream by suppressing undesired backgrounds, echoes, and reverberations, as is goal of speech enhancement (Loizou 2013). In these systems, the ultimate goal is to generate a simplified or cleaned version of the auditory scene that captures only one or a few signals of interest.

Overall, the lack of uniformity across the body of work addressing the computational bases of auditory scene analysis raises additional challenges when it comes to assessing the success of such systems: it becomes task dependent and

contingent on the perspective of the modeler. The lack of well-defined goals is one of the main hurdles that restricts progress in the field, constrains comparative studies of existing models, and limits incremental innovation that builds on the existing body of work. Ultimately, the cocktail party problem is an inherently cross-disciplinary challenge spanning domains of neuroscience, cognitive science, behavioral sciences, ethology, psychology, psychophysics, and medicine, as well as engineering and computer sciences. Naturally, the perspective of each of these disciplines puts the emphasis on different aspects of the problem and biases the computational theory to tackle the cocktail party problem at different levels of abstraction and granularity.

5.3 Principles of Modeling the Cocktail Party Problem

The cocktail party problem falls in the category of general information processing systems, which can be nicely framed in the context of Marrian models that emphasize different levels of granularity for understanding the underlying processes (Marr 1982). Although Marr's specific tri-level explanation may ultimately be incomplete (Poggio 2012), it nonetheless provides an integrated framework for understanding different levels of information processing. At the highest level, the *computational theory* describes the overall goal of the system and what a model of auditory scene analysis needs to achieve. In the case of the cocktail party problem, this remains one of the most challenging levels to describe. As highlighted in Sect. 5.2, the cocktail party effect is not a well-defined problem with an agreed-on objective. Most models strive to provide an informative mapping of a complex audio signal whether in the form of segregated streams, recognition of sound events, or synthesized variations of the same scene. At the next level of granularity, the *algorithm* describes the approach undertaken to achieve this goal. This level encompasses approaches based on analysis, recognition, or synthesis. At the lowest level, the *implementation* level details the practical realization of the algorithmic computation in terms of computational primitives or neural mechanisms at different levels of the hierarchy.

5.3.1 Algorithmic Strategies

The overall strategy undertaken by most models of the cocktail party problem focuses on invoking processes that extract "discriminative" cues from the incoming sensory input in such a way as to facilitate the differentiation of distinct sound streams or target selection. This is a particularly daunting task because these cues operate not only locally, but also globally, as sound streams evolve over time. These strategies have generally clustered into a few standard approaches, as outlined next.

5.3.1.1 The Population-Separation Theory

The premise of the "population-separation" theory and its related "peripheral channeling" account is that the perceptual organization of sounds into segregated streams is determined by the physical overlap between neural populations driven by sensory properties of the input. Van Noorden originally championed this principle in his doctoral work (van Noorden 1975), where he particularly emphasized the role of peripheral population separation. Specifically, sounds that activate separate peripheral channels (defined as tonotopic frequency channels or left–right lateral channels) would give rise to segregated stream percepts. A number of studies have in fact provided support for this observation confirming that formation of segregated auditory streams is strongest when sounds occupy separate peripheral channels (van Noorden 1977; Hartmann and Johnson 1991).

Subsequent experiments have contested the specific premise of peripheral channeling, showing that separate streams can in fact be formed even when sources share a common frequency range, as long as they differ along another acoustic dimension. Numerous psychoacoustic studies have shown that stream segregation can occur for sounds that differ in timbre (Cusack and Roberts 2000), bandwidth (Cusack and Roberts 1999), amplitude modulation rate (Grimault et al. 2002), binaural pitch (Akeroyd et al. 2005), unresolved pitch (Vliegen and Oxenham 1999), phase (Roberts et al. 2002), or perceived spatial location (Darwin and Hukin 1999; Gockel et al. 1999). Although most of these stimulus manipulations do not evoke peripheral channeling per se, they generate sound sources that activate separate neural channels at the brainstem or higher levels of auditory processing. In this way, these findings still support the more general population separation premise that activation of distinct neural populations (whether at peripheral or central nuclei of the auditory pathway) is a prerequisite for their perceptual segregation into distinct streams.

The population separation theory is supported by a number of neurophysiological studies that corroborate the role of feature selectivity in the auditory system in mediating the organization of sensory cues into segregated perceptual streams. Evidence of a correlation between responses at individual neuronal sites and perceptual judgments of streaming has been reported in animal models at various levels of processing from the cochlear nucleus (Pressnitzer et al. 2008) all the way to auditory cortex (Micheyl et al. 2007; Itatani and Klump 2011). Neural correlates of stream formation have also been explored in humans, using electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI) (Simon, Chap. 7). Overall, human studies corroborate the role of feature selectivity and tonotopic organization along the auditory pathway in facilitating stream segregation.

Computationally, the role of population separation in the organization of auditory streams can be interpreted as providing a discriminable representation of acoustic cues that allows mapping the stimulus into a separable space. By projecting sensory information into a new feature space that provides non- or minimally overlapping manifolds of the data, the neural representation enhances

discriminability between different auditory streams in the scene, allowing them to be separated. This operation is reminiscent of classification and regression techniques such as support vector machines and kernel-based classifiers (Duda et al. 2000; Herbrich 2001).

5.3.1.2 The Temporal Coherence Theory

The general population-separation theory accounts for a number of perceptual findings about stream segregation induced by sufficiently salient differences across sound dimensions (Moore and Gockel 2002). However, it does not account for crucial aspects of stream segregation that relate to the relative timing between sound events. Specifically, as sounds evolve over time, the relative timing between individual components in a complex scene plays a crucial role in dictating whether these components will segregate as separate streams or group together as a single stream. For instance, frequency components that start together (i.e., share a common onset) are likely to be perceived as grouped together (Darwin and Carlyon 1995), while delays of a few tens of milliseconds can suffice to induce a segregated percept (Sheft 2008). Similarly, frequency channels that evolve together in time over hundreds of milliseconds are likely to be perceived as one group, whereas elements that are out of phase relative to each other are likely to segregate (Micheyl et al. 2013). These longer time constants over which sound features evolve directly influence the nature of the stimulus-induced neural response. Indeed, sound components—if sufficiently far apart, for example, in frequency—will activate clearly distinct frequency-selective neural populations regardless of whether there are perceived as segregated or grouped (Elhilali et al. 2009), hence violating the population separation premise.

The *temporal coherence* theory has been proposed to complement the population-separation theory by addressing its main shortcoming, notably by incorporating information about the relative timing across neural responses to sounds over longer time constants (Shamma et al. 2011). This concept emphasizes the notion of temporal coherence whereby neural populations whose responses are in phase relative to each other over long time windows (hundreds of milliseconds) should be treated as if they represent a perceptually coherent stream; conversely, neural populations whose responses are asynchronous should be treated as representing sounds that probably belong to different streams.

By combining together the ideas of feature selectivity (which is at the core of the population-separation theory) and grouping by temporal coherence, one obtains a general model of auditory stream formation as illustrated in Fig. 5.1. This model includes two main bottom-up stages: a feature analysis stage followed by a coherence analysis stage. The analysis of sound features begins with a frequency mapping, which simulates the spectral analysis performed at the level of the cochlea. The output of this initial frequency analysis is used to extract a variety of spectral and temporal sound features, including spectral shape and bandwidth, harmonicity, temporal periodicity, and interaural time and level differences. For

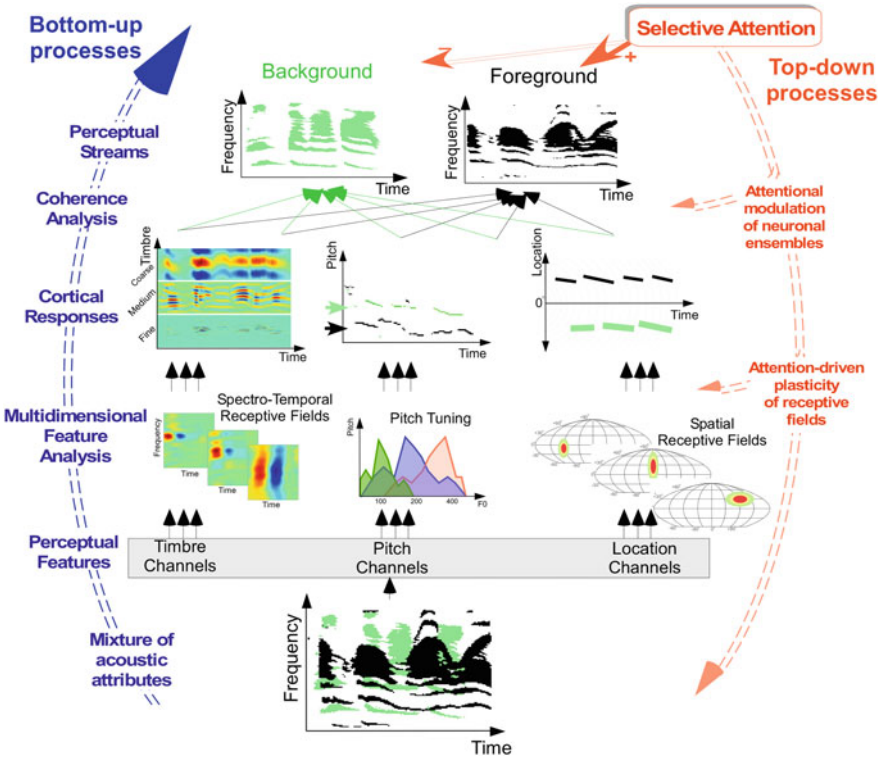

Fig. 5.1 Schematic of the temporal coherence strategy for modeling the cocktail party problem. An incoming signal (bottom of figure) consisting of a mixture of acoustic waveforms emanating from multiple sources is first analyzed through an array of auditory feature channels. These features extract cues (e.g., spatial location, pitch, etc.) that enable the segregation of sound attributes onto different perceptual streams. This process projects the low-dimensional acoustic signal onto a multidimensional space where different sound components occupy separate subspaces of the feature space, effectively segregating common sets of elements of the input and facilitating the process of formation of auditory objects or streams. This process takes advantage of the intricate time–frequency–space selectivity of neurons along the auditory pathway up to the level of auditory cortex. A coherence process tracks the trajectory of this feature space over "cortical" time constants of the order of few hundred milliseconds and binds together the elements that covary together, hence forming a representation of the foreground stream away from the background (top of figure). Top-down processes, particularly selective-attention (arrows on right-hand side) can modulate this entire process by exerting feedback projections that can reshape selectivity of cortical neurons or modulate ensemble of neurons. This process facilitates figure/ground segregation. [Figure adapted from Shamma et al. (2011).]

computational convenience, and illustration, these various feature detectors are assumed to be organized in "maps." However, it is important to note that an orderly topographic representation of sound features is not required for the general model to operate. The key point is that the model includes neurons selective to different sound features, or different values of a particular feature. Temporal coherence then

operates on these neural outputs to bind together elements that covary over time, while segregating those that are out of synchrony relative to each other (Krishnan et al. 2014).

It is worth noting that the principle of temporal coherence falls in the general category of correlational models that have been proposed many decades ago to address the cocktail party problem (von der Malsburg 1994; Wang and Brown 1999). The correlation output is generated by an autonomous process via neural coupling that allows neurons to synchronize together if driven by temporally bound features, forming a topographic map. This concept has been formalized in computational scene analysis models using oscillatory networks, where each population of synchronized oscillators represents an auditory stream (Wang and Brown 1999; Brown et al. 2001). In the majority of these models, correlation is defined as pairwise instantaneous temporal coincidence between temporal trajectories along different acoustic features.

The concept of "temporal coherence" takes a different view than instantaneous associations across sound elements (Krishnan et al. 2014). It emphasizes correlations among slow-varying temporal outputs of feature-selective neurons over longer time scales—of the order of hundreds of milliseconds (Elhilali et al. 2009, 2010). These time scales are commensurate with dynamics observed in the mammalian primary auditory cortex. The contrast between the variable time scales of correlations between an oscillatory model and a temporal coherence model is highlighted in Eq. (5.1):

$$Cr_{ij} = \frac{1}{\Gamma}\int r_i(t) r_j(t) dt \text{ vs. } Ch_{ij} = \frac{1}{\Gamma}\int \left[r_i(t) *_t h_{\tau_k}(t)\right]\left[r_j(t) *_t h_{\tau_k}(t)\right]^* dt \quad (5.1)$$

where $r_i(t)$ is the stimulus-driven response in the ith feature channel, Γ is an appropriately chosen normalization constant, $*_t$ represents convolution over time t, and $h_{\tau_k}(t)$ is the impulse response of a modulation-selective filter with time constant τ_k. * is the conjugate symmetry operator that accounts for the fact that the filter $h_{\tau_k}(t)$ is modeled as a complex-valued system that reflects both the magnitude and phase alignment of the stimulus with the time integration channel τ_k. So, although both correlation and coherence are computing a coincidence across different feature channels; they are operating at different time scales. The former is an instantaneous correlation across pairs of feature channels, whereas the latter is an operator that tracks longer-term correlations, parameterized by filters $h_k(.)$ over time constants τ_k. The coherence operator therefore is effectively tracking the trajectory of activity across feature channels, which results in a different tracing of coincidence across feature channels.

It is essential to note that the term temporal coherence used in the literature in the context of feature binding (Shamma et al. 2011) refers to stimulus-induced temporal coherence of neural activity and should not be confused with intrinsically generated temporal coherence, for example, oscillations in the gamma band. The current chapter refers specifically to the former. However, stimulus-induced neural

responses may interact with (and enhance or suppress) intrinsically generated temporal patterns of neural activity (Lakatos et al. 2005).

The role of temporal coherence in providing a framework for feature binding is not unique to the auditory modality, but has been advanced in other contexts and in other sensory modalities. It has been suggested that a similar principle operates in the visual modality (Alais et al. 1998; Blake and Lee, 2005). In addition, it has also been speculated that temporal coherence between cortical areas corresponding to different sensory modalities can, in principle, support cross-modal binding, for example, lip-reading, though not much is known about the exact role of cross-modal interactions in auditory stream formation (Almajai and Milner 2011; Mirbagheri et al. 2012).

5.3.1.3 The Inference Theory

The concept of temporal coherence reviewed in Sect. 5.3.1.2 is based on a notion of tracking the temporal evolution of sound elements. A closely related strategy, posited as the underlying neural process for organizing a complex acoustic scene, is that of prediction-based or inference models (Winkler et al. 2009). Inference-based computation provides a framework for integrating all available cues (e.g., sensory, contextual, cognitive) to derive likely interpretations of the soundscape. Initially, this process maps the acoustic input onto a high dimensional representation or onto feature maps (akin to processes underlying population separation). This mapping parameterizes the acoustic environment along dimensions that represent an estimate of the likelihood of a particular decomposition of the soundscape, based on acoustic attributes. This representation can further be integrated with priors that represent sensory statistics or dynamics of the acoustic features, as well as potential contextual information and any additional prior knowledge. This evidence is then integrated using an optimal Bayesian framework or alternative strategies to infer knowledge about the state of the auditory scene and its constituent streams (Friston 2010; Elhilali 2013).

This inference process can take many forms. Arguably, one of the most biologically plausible implementations invokes predictive coding, which processes sensory information in terms of predictive interpretations of the underlying events in the scene (Mumford 1992; Rao and Ballard, 1999). The circuitry underlying such processing has been studied at various hierarchical levels and has been speculated to include microcircuitry spanning sensory, parietal, and frontal cortex (Bastos et al. 2012). In the context of the cocktail party problem, such mechanisms have been linked to the concept of regularity tracking as an underlying mechanism for perception in auditory scenes (Winkler et al. 2009). In this scheme, the brain's strategy is to capture the behavior of sound sources in the scene and their time-dependent statistics by inferring the evolution of sound streams: constantly generating new expectations that reflect the fidelity of the sensory evidence, and matching these predictions with the ongoing dynamics of the scene. This strategy has led to successful computational models of auditory scene analysis, framed either as discovery

of predictable patterns in the scene (Mill et al. 2013; Schroger et al. 2014) or as a tracking operator that follows the evolution of states in the auditory scene and integrates past behavior of sound sources with their expected trajectories (Elhilali and Shamma 2008). In many regards, the predictive tracking algorithm can be related to temporal coherence analysis, provided the temporal dynamics of both processes operate at similar "slow" time scales (4–20 Hz) commensurate with the neuronal dynamics at the level of primary auditory cortex (Krishnan et al. 2014).

5.3.1.4 Spatial Models

The spatial location of sound sources is one of the strong cues that facilitate the process of auditory scene analysis (Culling and Stone, Chap. 3). Acoustic events that emanate from the same location in space tend to be perceived as belonging to the same stream whereas events that originate from different locations tend to be assigned to different streams (Gilkey and Anderson, 2014). The effect of interferers on the perception of a target is greatly reduced when the signal and masker are perceived to be at different spatial locations, in a phenomenon referred to as spatial release from masking (Arbogast et al. 2002). The extent to which spatial separation of sound sources supports bottom-up stream segregation is an active topic of research (Middlebrooks, Chap. 6). Nevertheless, there is no doubt that spatial cues are crucial components in sound lateralization as well as object selection in complex soundscapes. As such, they have featured in a prominent role in a number of computational models of auditory scene analysis that operate with two or multiple microphones.

Models of the cocktail party for stereo and multimicrophone applications have indeed taken advantage of the spatial layout of the scene, either in conjunction with other acoustic cues or based solely on spatial processing. Bio-inspired models rely on binaural cues represented by interaural level, phase, or timing differences to facilitate the separation of sound components that originate from different locations (Stern et al. 2005). Central to these bio-inspired spatial models is the mechanism of cross-correlation or coincidence detection which allows a direct comparison of signals from the two ears. Building on a theory put forth by Jeffress (1948), an interaural cross-correlation is computed across different channels that often represent frequency-selective neural populations. A central processing stage generally follows to integrate cross-correlation responses across frequency and time (Colburn and Kulkarni 2005; Trahiotis et al. 2005).

In more engineering-centric models, binaural cues are used in conjunction with more probabilistic methods as complementary priors or to inform constraints on the location of sound sources (Marin-Hurtado et al. 2012; Alinaghi et al. 2014). In this body of work, the statistical structure of the sources or space itself plays a more prominent role in facilitating the segregation of the different signals. The most popular approach is this literature is blind source separation (BSS) which refers to a family of techniques that exploit the statistical structure of sources to separate their signals in a blind (i.e. unsupervised) manner (Bell and Sejnowski 1995; Naik and

Wang 2014). Generally, these algorithms are very effective at separating the sound sources under certain conditions that are gradually being relaxed by ongoing research efforts (Jutten and Karhunen 2004; Jadhav and Bhalchandra 2008).

Many engineering applications leverage the spatial analysis of a scene using multiple microphones. The rich sampling of the soundscape at multiple pick-up points opens the door to alternative techniques such as spatial sampling and beamforming (Van Veen and Buckley 1988; Krim and Viberg 1996). Such techniques aim at extracting a target source situated at a specific spatial direction using the sensor array. They focus on determining direction-of-arrival of sounds of interest, and are effectively filtering methods that operate in three-dimensional space to boost signals from a direction of interest. Although these techniques fall short of capitalizing on merits of spatial hearing, some have in fact benefited from human sound-source localization by employing adaptive beamformers that can judge the direction of target sounds, or take advantage of head-related transfer functions to reproduce out-of-head localization, or even incorporate simulations of room acoustics (Doclo and Moonen 2003; Farmani et al. 2015).

5.3.2 Neural Infrastructure

Most of the strategies discussed in Sect. 5.3.1 rely on intricate machinery or physical computations to achieve the required analysis of the complex scene. It is generally accepted that the pathway traveled by incoming acoustic information along the auditory system carries out the task of decomposing the sensory signal into its constituting elements and mapping them into perceptual streams (Nelken 2004). This neural transformation aims at extracting various acoustic features such as frequency, spectral profile, amplitude and frequency modulations, and interaural cues (Middlebrooks et al. 1980; Schreiner 1998). This feature representation is a canonical scheme for a discriminative representation of the scene that mediates the organization of features into segregated streams (Bizley and Cohen 2013).

Computationally, the incoming signal can be modeled as undergoing a series of mappings from acoustic space to a new feature space whose dimensionality facilitates the segregation or grouping of sound components into corresponding perceptual streams. At the core of this transformation is the concept of a receptive field, which has been instrumental in providing a functional descriptor of sensitivity of auditory neurons, as well as offering a computational medium for parameterizing the auditory feature space (Eggermont 2013). A receptive field can be thought of as a two-dimensional descriptor of the time-frequency sound features that best drive an auditory neuron, hence the name spectrotemporal receptive field (STRF) (Elhilali et al. 2013). It can be viewed as a time-dependent spectral transfer function, or a frequency-dependent dynamical filter (deCharms et al. 1998; Klein et al. 2000). In other words, if one views a neuron as a dynamical system, the STRF provides a descriptor of the linearized system function along both time and frequency, which maps the values of an input s at different time instants to a value of the output

(or response) r at the current time t (Korenberg and Hunter 1996), as described in Eq. (5.2):

$$r(t) = \sum_{f} \int \mathrm{STRF}(\tau, f) s(t - \tau, f) d\tau \tag{5.2}$$

Receptive field descriptors have been successfully approximated at subcortical (Escabi and Schreiner 2002; Bandyopadhyay and Young 2013), as well as cortical stages (Depireux et al. 2001; Miller et al. 2002). By and large, convergent evidence suggests that the accumulation of transformations through these diverse receptive fields from the periphery up to auditory cortex is instrumental in providing the rich high-dimensional space necessary for segregating components of an acoustic scene (Sharpee et al. 2011; Christison-Lagay et al. 2015).

Indeed, a number of studies suggest that the organization of sound elements into mental representations of auditory objects may reside as early as primary auditory cortex (A1) (Nelken and Bar-Yosef 2008; Bizley and Cohen 2013). The neural representation of sounds as viewed through cortical receptive fields covers a rich feature space that spans at least three key dimensions (Fig. 5.2b): (1) *best frequencies* (BF) that cover the entire auditory range (Schreiner 1998; Klein et al. 2003); (2) *bandwidths* that span a wide range from very broad (2–3 octaves) to narrowly tuned (<0.25 octave) (Schreiner and Sutter 1992; Versnel et al. 1995); (3) *dynamics* that range from very slow to relatively fast (1–30 Hz) (Lu et al. 2001; Miller et al. 2002). This variability along different acoustic attributes is at the core of a multidimensional neural representation of sound mixtures, which in turn

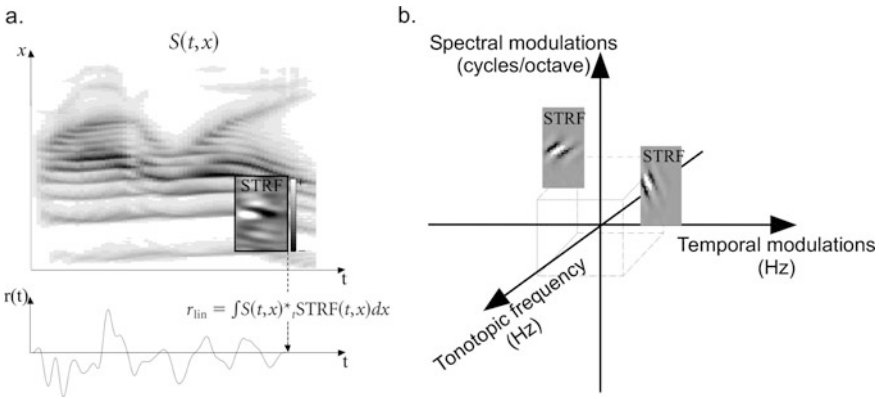

Fig. 5.2 Schematic of the concept of receptive field. **(a)** A spectrotemporal receptive field (STRF) operates as a two-dimensional filter that integrates stimulus information across time and frequency that best matches its selectivity. The corresponding neural response reflects the signal components that best drive the filter itself. **(b)** The selectivity of STRFs spans a high-dimensional space that spans tonotopic frequency, temporal modulations, and spectral modulations. Each STRF can be thought of as a mapping to a small portion of this space. The collection of responses through a network of neurons would correspond to a mapping onto a high-dimensional space

facilitates the execution of a number of strategies for modeling the cocktail party problem (Cooke and Ellis 2001; Elhilali and Shamma 2008). State-of-the-art models of auditory scene analysis also build on the same foundation, of a rich feature space extended to nonlinear manifolds. Current techniques using deep belief architectures, convolutional neural networks, and multivariate analysis have also been shown to exploit a rich time–frequency mapping similar to that observed in neural receptive fields to facilitate tasks of source separation (Le Roux et al. 2015; Simpson 2015).

5.4 Bottom-up Models of the Cocktail Party Problem

Together, the strategies driving modeling efforts of the cocktail party problem draw on viewpoints prompted by multidisciplinary efforts spanning the engineering, psychology, and neuroscience communities. On one end of the spectrum, numerous studies have attempted strict engineering approaches such as the successful application of blind source separation techniques (Roweis 2001; Jang and Lee 2003), statistical speech models (Varga and Moore 1990; Yoon et al. 2009), and other machine learning algorithms (Ming et al. 2013; Souden et al. 2013). Most of these approaches construct systems that exploit statistical knowledge about the target of interest (e.g., existing database of the target speaker's voice), mine data about the physical or source characteristics of a target (e.g., knowledge about sources of noise), or utilize spatial characteristics of the receivers (usually in a multimicrophone setting) to hone in on desired signals to be segregated (Kristjansson et al. 2006; Madhu and Martin 2011). The statistical characteristics and possibly independence or uniqueness of the different sources (or at least the sound class of interest) are at the core of these approaches.

Despite their undeniable success, these algorithms often violate fundamental aspects of the manner in which humans and animals perform this task. They are generally constrained by their own mathematical formulations, are mostly applicable and effective in multisensor configurations, and/or require prior knowledge and training on the task at hand. By design, these systems target particular configurations of the sensory environment or require existing training databases or general knowledge about the task or target of interest. This reliance on training data or task-specific prior knowledge generally limits the applicability of these algorithms to general-purpose tasks. In this regard, the gap between these computational approaches and biological audition is still wide. A major effort in such engineering-centric systems deals with which patterns to extract from the scene and how to best capitalize on existing knowledge to perform the segregation, recognition, or synthesis task.

The best success stories in the category of engineering-centric systems are automatic speech recognition systems (Waibel and Lee 1990; Rabiner and Juang 1993) that focus on recognition of speech sounds even in the presence of unknown interferers and background noise. Although these systems are not immune to noise,

they have made great strides in improving the recognition accuracy by combining acoustic and language models that represent statistical representations of the sounds that make up each word and sequence of words as dictated by the grammatical rules of the language. This training knowledge is often combined by powerful machine learning tools such as convolutional systems and deep learning techniques (Hinton et al. 2012; Deng et al. 2013). These powerful tools, combined with abundance of training data, distance the challenge from the details of the feature analysis and compensate any weaknesses in the chosen signal representations by the strength of the statistical structure of the models. Unfortunately, these formulations limit any progress in truly understanding the strengths of the multiscale and parallel processing underlying sound processing in the auditory system and limit translating successes from these engineering approaches into cocktail party models that can truly mimic brain functions.

On the other end of the spectrum are perceptually driven studies that focus on factors influencing auditory scene analysis, in particular the segregation/binding cues that govern the simultaneous and sequential integration of sound patterns into objects emanating from a same environmental event (Bregman 1990; Carlyon 2004). These efforts have triggered a lot of interest in constructing *biologically inspired systems* that can perform intelligent processing of complex sound mixtures. Early instantiations of these models were strongly focused on the peripheral representations of sound. These models focused on peripheral selectivity, possibly allowing competition between different channels to result in a dominant foreground stream (Beauvois and Meddis 1996; McCabe and Denham 1997).

Other studies took more pragmatic approaches to modeling the cocktail party problem; particularly capitalizing on the salient acoustic attributes that can be tracked for individual sources to segregate them from competing backgrounds. Early work by Parsons (1976) and Weintraub (1985) focused on tracking the fundamental frequency of concurrent speakers. The role of a particular auditory feature (e.g., pitch) was later extended to additional acoustic cues and grouping dimensions following the basic premise of Gestalt principles and population separation theory, but with different computational implementations of the binding and integration stage (Brown and Cooke 1994).

The extraction of acoustic features has also been a cornerstone of correlation-based models mentioned in Sect. 5.3.1, by exploiting synchrony between different oscillators as a reflection of a grouped perceptual stream (Brown and Cooke 1998; Wang and Brown 1999). Synchrony of individual oscillators is initiated by regularity in the sound's spectrotemporal elements, and hence lateral connections between oscillators are implemented to encode harmonicity and proximity in time and frequency. A similar concept of feature extraction is also at the core of coherence-based models that emphasize the role of temporal integration over relatively long time scales; hence viewing feature analysis through the lens of temporal properties at the level of the mammalian primary auditory cortex (Shamma et al. 2011; Krishnan et al. 2014).

By and large, biomimetic models of auditory analysis of complex scenes have universally invoked the extraction of acoustic features as a foundation of any

subsequent processing. However, these implementations largely favor a bottom-up processing view (Fig. 5.1), relying on the salience of stimulus events. The models—with few exceptions—often abstract away intricate and indispensable contributions of goal-directed top-down processing and shy away from incorporating truly adaptive and task-dependent neural processing under top-down control.

5.5 Top-Down Processes and the Cocktail Party Problem

Along with the physical properties of sounds in the environment, listeners exploit learned knowledge from their recent and lifelong experiences to further complement processing of complex auditory scenes (Bregman 1990; Ciocca 2008). These learned "schemas" encompass a listener's familiarity with the statistical structure of sound sources (e.g., natural sounds), recent and long-term memories about specific sources, expectation about the state of the world (e.g., speech sounds produced through a human vocal tract), as well as their attentional state which helps steer brain processes towards targets of interest while ignoring background interferers. These processes are believed to play a crucial role in tackling the cocktail party problem because they impose constraints on the space of possible solutions. They can be viewed as top-down or feedback projections that control the system's performance to meet desired behaviors.

Of all schema-based processes, attention is one of the most widely studied top-down mechanisms affecting the cocktail party problem (Shinn-Cunningham, Chap. 2). It is a crucial component in the scene analysis process because it dictates what the targets of interest are, and orients the listener to the desired sound source or sources. It ultimately acts as a processing bottleneck that appropriately allocates neural resources to informative events in the acoustic scene and selectively filters the most relevant sensory inputs (Whiteley and Sahani 2012). While clearly behaviorally crucial, the specific roles of attention in auditory scene analysis remain an unsettled question in the field. It is certainly true that attention can strongly affect stream segregation. For instance, the ability to switch at will between hearing certain tone sequences as one or two streams can be thought of as an effect of attention, but the question of whether attention is *necessary* for streaming remains a matter of debate (Carlyon et al. 2001; Macken et al. 2003).

The bulk of the current literature suggests that at least some forms of stream segregation occur in the absence of attention, in what is termed "primitive" stream segregation (Bregman 1990; Sussman et al. 2007). As outlined in Sect. 5.3, the vast majority of cocktail party models have indeed implemented successful renditions of the problem solution in absence of any role of selective attention. Stream segregation may also be thought of as a process that facilitates attention (rather than only vice versa) in that it becomes possible to pay exclusive attention to tones of a single frequency only if they are successfully segregated from other tones in the sequence (Shinn-Cunningham 2008).

In the case of alternating tone sequences, early work by Van Noorden provided a useful distinction by defining two boundaries, the fission boundary and the coherence boundary (van Noorden 1975). The fission boundary defines the frequency difference (or other dimension) below which segregation is not possible, while the coherence boundary defines the point above which integration is not possible. The area in between these two boundaries can be thought of as the region in which attention can play a particularly important role in determining whether one or two streams are heard.

Though some computational models of the cocktail party problem have attempted to reproduce these effects (Beauvois and Meddis 1996; Wang and Chang 2008), they have not truly incorporated any mechanisms manipulating the attentional state of listener/model in a way that mimics the presumed feedback control exerted by attentional projections on feedforward sensory processing.

At the physiological level, a growing body of literature has established that auditory experience throughout adulthood can have profound global effects by reshaping cortical maps and significant local effects by transforming receptive field properties of neurons in primary auditory cortex (Suga et al. 1997; Weinberger 2001). The exact form of this remarkable plasticity is determined by the salience or task relevance of the spectral and temporal characteristics of the acoustic stimuli (Kilgard et al. 2001). Recent findings have also shown that cortical responses are heavily modulated by the attentional state of the brain and undergo rapid, short-term, and task-dependent changes that reflect not only the incoming sensory cues but also behavioral goals (Fritz et al. 2007; Mesgarani and Chang 2012). In this kind of adaptive plasticity, selective functional reconfiguration or resetting of the underlying cortical circuitry leads to changes in receptive field properties that may enhance perception in a cocktail party (Shamma and Fritz 2014).

Unfortunately, there is a notable lack of the incorporation of cognitive or adaptive mechanisms into mathematical models of auditory cortical processing and, ultimately, implementations of cocktail party models. This deficiency is itself motivated by lack of information and ignorance of the neural mechanisms underlying the ability of cortical circuits to adapt online to changing behavioral demands. In contrast, active and adaptive processing has more commonly been explored in models of the visual system. These implementations typically model parallels of predictive coding in the visual thalamus (LGN), contextual modulation in primary visual cortex (V1), attentional modulation in higher cortical areas (V2 and V4, and area MT), as well as decision making in parietal and frontal cortex. A commonly used formulation for such systems is that of generative models, whereby sensory input can be explained as being caused by hidden "causes" or "states" in the world (Duda et al. 2000). The model then estimates the probability of these causes based on inputs incoming up to a certain point in time. Modeling based on hidden causes or states is amenable to predictive coding, similar to concepts discussed in Sect. 5.3.1.3. In other words, the models employ a probabilistic formulation where optimization functions can then be defined as maximizing posterior probabilities, which is equivalent to minimizing the prediction error generated by this model. Some studies have presented successful implementations of these models as

hierarchical systems of early and higher visual cortical processing (Rao and Ballard 1999; Lee and Mumford 2003). This body of work has often relied on a linear formulation of the generative model, hence benefiting from existing linear hidden state estimation techniques such as Kalman filtering. The tracking of these latent states was also formulated to adapt the model parameters continuously to the statistics in the visual scene, hence giving the system a desired plastic behavior. Other techniques have also been explored to go beyond the generative model approach. Systems based on belief propagation, graphical models, as well as inference in recurrent networks have shown variable success in interpreting top-down feedback as prior probabilities (Rao 2005).

Recent models and frameworks for modeling the cocktail party effect and its biological bases have begun focusing on the role of schema-based processes, particularly attention in both its bottom-up and top-down forms in biasing selection and organization of sensory events (Shamma et al. 2011; Kaya and Elhilali 2014). Ultimately, progress in the integration of top-down processes in cocktail party models is closely tied to progress in unraveling neural mechanisms underlying cognitive effects on sensory processing, as well as models of feedback loops in shaping auditory processing of complex scenes.

5.6 Summary

The challenge of auditory scene analysis is a problem facing biological and engineering systems alike. Computational auditory scene analysis is a young field that aims at providing theoretical insights and solutions to the cocktail party problem that can inform neuroscience research as well as benefit audio applications. Though a lofty goal, translating perceptual phenomena related to the cocktail party problem to exact mathematical formulations requires more concise definitions of the problem, well-defined constraints on the desired system, as well as clear measureable outcomes and behaviors. Indeed, the cocktail party problem is a phenomenological description of multiple tasks related to processing complex soundscapes. These range from detection and recognition to tracking, description, and audio resynthesis. Translating these problems into computational models leaves the field somewhat fragmented.

Nonetheless, a rich body of computational models has offered insights into how the brain might tackle the cocktail party challenge. These invoke the rich feature selectivity that underlies neural processing through the auditory pathway from the periphery all the way to auditory cortex. The neural transformations up to sensory cortex offer part of the solution to the segregation of sound mixtures along informative dimensions for further processing. Additional processes such as temporal coherence play a role in the binding process that combines relevant acoustic cues onto perceptual streams corresponding to perceived objects. Computational models also capitalize on the structure of sound sources to track the regularities or dynamics of sound events over time.

All in all, models inspired from brain processes have laid the conceptual groundwork for interpreting the transformation from an acoustic space of a mixture of sound sources to a perceptual space with segregated streams. Translating this foundation into practical engineering applications and evaluating its effectiveness remains one of the big challenges in the field. In conjunction, additional factors, particularly with regard to schema-based processes (e.g., attention, learning), add extra hurdles in developing full solutions to the cocktail party problem that could come close to emulating the biological system. As the growing yet limited knowledge of the neural underpinnings of schema-based processes sheds light on their role in cocktail parties, truly intelligent systems will undoubtedly emerge that can mimic the complex processing exhibited by the brain when dealing with the cocktail party problem.

Acknowledgements Dr. Elhilali's work is supported by grants from The National Institutes of Health (NIH: R01HL133043) and the Office of Naval Research (ONR: N000141010278, N000141612045, and N000141210740).

Compliance with Ethics Requirements
Mounya Elhilali declares that she has no conflict of interest.

References

Akeroyd, M. A., Carlyon, R. P., & Deeks, J. M. (2005). Can dichotic pitches form two streams? *The Journal of the Acoustical Society of America, 118*(2), 977–981.

Alais, D., Blake, R., & Lee, S. H. (1998). Visual features that vary together over time group together over space. *Nature Neuroscience, 1*(2), 160–164.

Alinaghi, A., Jackson, P. J., Liu, Q., & Wang, W. (2014). Joint mixing vector and binaural model based stereo source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22*(9), 1434–1448.

Almajai, I., & Milner, B. (2011). Visually derived wiener filters for speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing, 19*(6), 1642–1651.

Anemuller, J., Bach, J., Caputo, B., Havlena, M., et al. (2008). The DIRAC AWEAR audio-visual platform for detection of unexpected and incongruent events. In *International Conference on Multimodal Interaction*, (pp. 289–293).

Arbogast, T. L., Mason, C. R., & Kidd, G. (2002). The effect of spatial separation on informational and energetic masking of speech. *The Journal of the Acoustical Society of America, 112*(5 Pt 1), 2086–2098.

Aubin, T. (2004). Penguins and their noisy world. *Annals of the Brazilian Academy of Sciences, 76* (2), 279–283.

Bandyopadhyay, S., & Young, E. D. (2013). Nonlinear temporal receptive fields of neurons in the dorsal cochlear nucleus. *Journal of Neurophysiology, 110*(10), 2414–2425.

Barchiesi, D., Giannoulis, D., Stowell, D., & Plumbley, M. D. (2015). Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Magazine, 32*(3), 16–34.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., et al. (2012). Canonical microcircuits for predictive coding. *Neuron, 76*(4), 695–711.

Beauvois, M. W., & Meddis, R. (1996). Computer simulation of auditory stream segregation in alternating-tone sequences. *The Journal of the Acoustical Society of America, 99*(4), 2270–2280.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*(6), 1129–1159.

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience, 14*(10), 693–707.

Blake, R., & Lee, S. H. (2005). The role of temporal structure in human vision. *Behavioral and Cognitive Neuroscience Review, 4*(1), 21–42.

Bregman, A. S. (1981). Asking the 'what for' question in auditory perception. In M. Kubovy & J. Pomerantz (Eds.), *Perceptual organization* (pp. 99–118). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: 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.

Brown, G. J., & Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech & Language, 8*(4), 297–336.

Brown, G. J., & Cooke, M. (1998). Temporal synchronization in a neural oscillator model of primitive auditory stream segregation. In D. L. Wang & G. Brown (Eds.), *Computational auditory scene analysis* (pp. 87–103). London: Lawrence Erlbaum Associates.

Brown, G. J., Barker, J., & Wang, D. (2001). A neural oscillator sound separator for missing data speech recognition. In *Proceedings of International Joint Conference on Neural Networks, 2001* (*IJCNN* '01) (Vol. 4, pp. 2907–2912).

Buxton, H. (2003). Learning and understanding dynamic scene activity: A review. *Image and Vision Computing, 21*(1), 125–136.

Carlyon, R. P. (2004). How the brain separates sounds. *Trends in Cognitive Sciences, 8*(10), 465–471.

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance, 27*(1), 115–127.

Chen, F., & Jokinen, K. (Eds.). (2010). *Speech technology: Theory and applications*. New York: Springer Science+Business Media.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*(5), 975–979.

Cherry, E. C. (1957). *On human communication*. Cambridge, MA: MIT Press.

Christison-Lagay, K. L., Gifford, A. M., & Cohen, Y. E. (2015). Neural correlates of auditory scene analysis and perception. *International Journal of Psychophysiology, 95*(2), 238–245.

Ciocca, V. (2008). The auditory organization of complex sounds. *Frontiers in Bioscience, 13,* 148–169.

Cisek, P., Drew, T., & Kalaska, J. (Eds.). (2007). *Computational neuroscience: Theoretical insights into brain function*. Philadelphia: Elsevier.

Colburn, H. S., & Kulkarni, A. (2005). Models of sound localization. In A. N. Popper & R. R. Fay (Eds.), *Sound source localization* (pp. 272–316). New York: Springer Science+Business Media.

Collins, N. (2009). *Introduction to computer music*. Hoboken, NJ: Wiley.

Cooke, M., & Ellis, D. (2001). The auditory organization of speech and other sources in listeners and computational models. *Speech Communication, 35,* 141–177.

Cusack, R., & Roberts, B. (1999). Effects of similarity in bandwidth on the auditory sequential streaming of two-tone complexes. *Perception, 28*(10), 1281–1289.

Cusack, R., & Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Perception and Psychophysics, 62*(5), 1112–1120.

Darwin, C. J., & Carlyon, R. P. (1995). Auditory grouping. In B. C. J. Moore (Ed.), *Hearing* (pp. 387–424). Orlando, FL: Academic Press.

Darwin, C. J., & Hukin, R. W. (1999). Auditory objects of attention: The role of interaural time differences. *Journal of Experimental Psychology: Human Perception and Performance, 25*(3), 617–629.

deCharms, R. C., Blake, D. T., & Merzenich, M. M. (1998). Optimizing sound features for cortical neurons. *Science, 280*(5368), 1439–1443.

Deng, L., Li, J., Huang, J., Yao, K., et al. (2013). Recent advances in deep learning for speech research at Microsoft. In *2013 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 26–31, 2013 (pp. 8604–8608).

Depireux, D. A., Simon, J. Z., Klein, D. J., & Shamma, S. A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of Neurophysiology, 85*(3), 1220–1234.

Doclo, S., & Moonen, M. (2003). adaptive. *EURASIP Journal of Applied Signal Processing, 11,* 1110–1124.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification*. Hoboken, NJ: Wiley.

Eggermont, J. J. (2013). The STRF: Its origin, evolution and current application. In D. Depireux & M. Elhilali (Eds.), *Handbook of modern techniques in auditory cortex* (pp. 1–32). Hauppauge, NY: Nova Science Publishers.

Elhilali, M. (2013). Bayesian inference in auditory scenes. In *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, Japan, (pp. 2792–2795).

Elhilali, M., & Shamma, S. A. (2008). A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation. *The Journal of the Acoustical Society of America, 124*(6), 3751–3771.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron, 61*(2), 317–329.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A., & Shamma, S. (2010). Rate vs. temporal code? A spatio-temporal coherence model of the cortical basis of streaming. In E. Lopez-Poveda, A. Palmer & R. Meddis (Eds.), *Auditory physiology*, *perception and models* (pp. 497–506). New York: Springer Science+Business Media.

Elhilali, M., Shamma, S. A., Simon, J. Z., & Fritz, J. B. (2013). A linear systems view to the concept of STRF. In D. Depireux & M. Elhilali (Eds.), *Handbook of modern techniques in auditory cortex* (pp. 33–60). Hauppauge, NY: Nova Science Publishers.

Escabi, M. A., & Schreiner, C. E. (2002). Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. *The Journal of Neuroscience, 22*(10), 4114–4131.

Farmani, M., Pedersen, M. S., Tan, Z. H., & Jensen, J. (2015). On the influence of microphone array geometry on HRTF-based sound source localization. In *2015 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)*, (pp. 439–443).

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138.

Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention–focusing the searchlight on sound. *Current Opinion in Neurobiology, 17*(4), 437–455.

Gilkey, R., & Anderson, T. R. (Eds.). (2014). *Binaural and spatial hearing in real and virtual environments*. New York: Psychology Press.

Gockel, H., Carlyon, R. P., & Micheyl, C. (1999). Context dependence of fundamental-frequency discrimination: Lateralized temporal fringes. *The Journal of the Acoustical Society of America, 106*(6), 3553–3563.

Grimault, N., Bacon, S. P., & Micheyl, C. (2002). Auditory stream segregation on the basis of amplitude-modulation rate. *The Journal of the Acoustical Society of America, 111*(3), 1340–1348.

Hartmann, W., & Johnson, D. (1991). Stream segregation and peripheral channeling. *Music Perception, 9*(2), 155–184.

Haykin, S., & Chen, Z. (2005). The cocktail party problem. *Neural Computation, 17*(9), 1875–1902.

Herbrich, R. (2001). *Learning kernel classifiers: Theory and algorithms*. Cambridge, MA: MIT Press.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE, 29*(6), 82–97.

Hyvarinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. Hoboken, NJ: Wiley.

Itatani, N., & Klump, G. M. (2011). Neural correlates of auditory streaming of harmonic complex sounds with different phase relations in the songbird forebrain. *Journal of Neurophysiology, 105*(1), 188–199.

Izumi, A. (2002). Auditory stream segregation in Japanese monkeys. *Cognition, 82*(3), B113–B122.

Jadhav, S. D., & Bhalchandra, A. S. (2008). Blind source separation: Trends of new age—a review. In *IET International Conference on Wireless*, *Mobile and Multimedia Networks*, *2008*, Mumbai, India, January 11–12, 2008 (pp. 251–254).

Jang, G. J., & Lee, T. W. (2003). A maximum likelihood approach to single-channel source separation. *Journal of Machine Learning Research, 4*(7–8), 1365–1392.

Jeffress, L. A. (1948). A place theory of sound localization. *Journal of Comparative and Physiological Psychology, 41*(1), 35–39.

Jutten, C., & Karhunen, J. (2004). Advances in blind source separation (BSS) and independent component analysis (ICA) for nonlinear mixtures. *International Journal of Neural Systems, 14*(5), 267–292.

Kaya, E. M., & Elhilali, M. (2013). Abnormality detection in noisy biosignals. In *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, Japan (pp. 3949–3952).

Kaya, E. M., & Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience*, *8*(327), doi:10.3389/fnhum.2014.00327

Kilgard, M. P., Pandya, P. K., Vazquez, J., Gehi, A., et al. (2001). Sensory input directs spatial and temporal plasticity in primary auditory cortex. *Journal of Neurophysiology, 86*(1), 326–338.

Klein, D. J., Depireux, D. A., Simon, J. Z., & Shamma, S. A. (2000). Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design. *Journal of Computational Neuroscience, 9*(1), 85–111.

Klein, D. J., Konig, P., & Kording, K. P. (2003). Sparse spectrotemporal coding of sounds. *EURASIP Journal on Applied Signal Processing, 2003*(7), 659–667.

Korenberg, M., & Hunter, I. (1996). The identification of nonlinear biological systems: Volterra kernel approaches. *Annals of Biomedical Engineering, 24*(4), 250–268.

Krim, H., & Viberg, M. (1996). Two decades of array signal processing research: The parametric approach. *IEEE Signal Processing Magazine, 13*(4), 67–94.

Krishnan, L., Elhilali, M., & Shamma, S. (2014). Segregating complex sound sources through temporal coherence. *PLoS Computational Biology, 10*(12), e1003985.

Kristjansson, T., Hershey, J., Olsen, P., Rennie, S., & Gopinath, R. (2006). Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system. In *International Conference on Spoken Language Processing*, Pittsburgh, PA, September 17–21, 2006.

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., et al. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology, 94*(3), 1904–1911.

Lee, T. S., & Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America, 20*(7), 1434–1448.

Le Roux, J., Hershey, J. R., & Weninger. F. (2015). Deep NMF for speech separation. In *2015 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)*, Brisbane, Australia, April 19–24, 2015 (pp. 66–70).

Lewicki, M. S., Olshausen, B. A., Surlykke, A., & Moss, C. F. (2014). Scene analysis in the natural environment. *Frontiers in Psychology, 5,* 199.

Loizou, P. C. (2013). *Speech enhancement: Theory and practice* (2nd ed.). Boca Raton, FL: CRC Press.

Lu, T., Liang, L., & Wang, X. (2001). Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nature Neuroscience, 4*(11), 1131–1138.

Macken, W. J., Tremblay, S., Houghton, R. J., Nicholls, A. P., & Jones, D. M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of Experimental Psychology: Human Perception and Performance, 29*(1), 43–51.

Madhu, N., & Martin, R. (2011). A versatile framework for speaker separation using a model-based speaker localization approach. *IEEE Transactions on Audio, Speech and Language Processing, 19*(7), 1900–1912.

Marin-Hurtado, J. I., Parikh, D. N., & Anderson, D. V. (2012). Perceptually inspired noise-reduction method for binaural hearing aids. *IEEE Transactions on Audio, Speech and Language Processing, 20*(4), 1372–1382.

Marr, D. (1982). *Vision*. San Francisco: Freeman and Co.

McCabe, S. L., & Denham, M. J. (1997). A model of auditory streaming. *The Journal of the Acoustical Society of America, 101*(3), 1611–1621.

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature, 485*(7397), 233–236.

Micheyl, C., Carlyon, R. P., Gutschalk, A., Melcher, J. R., et al. (2007). The role of auditory cortex in the formation of auditory streams. *Hearing Research, 229*(1–2), 116–131.

Micheyl, C., Hanson, C., Demany, L., Shamma, S., & Oxenham, A. J. (2013). Auditory stream segregation for alternating and synchronous tones. *Journal of Experimental Psychology: Human Perception and Performance, 39*(6), 1568–1580.

Middlebrooks, J. C., Dykes, R. W., & Merzenich, M. M. (1980). Binaural response-specific bands in primary auditory cortex (AI) of the cat: Topographical organization orthogonal to isofrequency contours. *Brain Research, 181*(1), 31–48.

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, 9*(3), e1002925.

Miller, L. M., Escabi, M. A., Read, H. L., & Schreiner, C. E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal of Neurophysiology, 87*(1), 516–527.

Ming, J., Srinivasan, R., Crookes, D., & Jafari, A. (2013). CLOSE—A data-driven approach to speech separation. *IEEE Transactions on Audio, Speech and Language Processing, 21*(7), 1355–1368.

Mirbagheri, M., Akram, S., & Shamma, S. (2012). An auditory inspired multimodal framework for speech enhancement. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Portland, OR.

Moore, B. C. J., & Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica, 88,* 320–333.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics, 66*(3), 241–251.

Naik, G., & Wang, W. (Eds.). (2014). *Blind source separation: Advances in theory, algorithms and applications*. Berlin/Heidelberg: Springer-Verlag.

Nelken, I. (2004). Processing of complex stimuli and natural scenes in the auditory cortex. *Current Opinion in Neurobiology, 14*(4), 474–480.

Nelken, I., & Bar-Yosef, O. (2008). Neurons and objects: The case of auditory cortex. *Frontiers in Neuroscience, 2*(1), 107–113.

Parsons, T. W. (1976). Separation of speech from interfering speech by means of harmonic selection. *The Journal of the Acoustical Society of America, 60*(4), 911–918.

Patil, K., & Elhilali, M. (2013). Multiresolution auditory representations for scene recognition. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, October 20–23, 2013.

Poggio, T. (2012). *The levels of understanding framework, revised.* Computer Science and Artificial Intelligence Laboratory Technical Report MIT-CSAIL-TR-2012–014. Cambridge, MA: Massachusetts Institute of Technology.

Pressnitzer, D., Sayles, M., Micheyl, C., & Winter, I. M. (2008). Perceptual organization of sound begins in the auditory periphery. *Current Biology, 18*(15), 1124–1128.

Rabiner, L., & Juang, B. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice Hall.

Rao, R. P. (2005). Bayesian inference and attentional modulation in the visual cortex. *NeuroReport, 16*(16), 1843–1848.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87.

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology, 12*(2), 162–168.

Roberts, B., Glasberg, B. R., & Moore, B. C. (2002). Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. *The Journal of the Acoustical Society of America, 112*(5), 2074–2085.

Roweis, S. T. (2001). One microphone source separation. *Advances in Neural Information Processing Systems, 13,* 793–799.

Schreiner, C. E. (1998). Spatial distribution of responses to simple and complex sounds in the primary auditory cortex. *Audiology and Neuro-Otology, 3*(2–3), 104–122.

Schreiner, C. E., & Sutter, M. L. (1992). Topography of excitatory bandwidth in cat primary auditory cortex: Single-neuron versus multiple-neuron recordings. *Journal of Neurophysiology, 68*(5), 1487–1502.

Schroger, E., Bendixen, A., Denham, S. L., Mill, R. W., et al. (2014). Predictive regularity representations in violation detection and auditory stream segregation: From conceptual to computational models. *Brain Topography, 27*(4), 565–577.

Shamma, S., & Fritz, J. (2014). Adaptive auditory computations. *Current Opinion in Neurobiology, 25,* 164–168.

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences, 34*(3), 114–123.

Sharpee, T. O., Atencio, C. A., & Schreiner, C. E. (2011). Hierarchical representations in the auditory cortex. *Current Opinion in Neurobiology, 21*(5), 761–767.

Sheft, S. (2008). Envelope processing and sound-source perception. In W. A. Yost, A. Popper, & R. R. Fay (Eds.), *Auditory perception of sound sources* (pp. 233–280). New York: Springer Science+Business Media.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences, 12*(5), 182–186.

Simpson, A. J. (2015). Probabilistic binary-mask cocktail-party source separation in a convolutional deep neural network. *arXiv Preprint* arXiv:1503.06962.

Souden, M., Araki, S., Kinoshita, K., Nakatani, T., & Sawada, H. (2013). A multichannel MMSE-based framework for speech source separation and noise reduction. *IEEE Transactions on Audio, Speech and Language Processing, 21*(9), 1913–1928.

Stern, R., Brown, G., & Wang, D. L. (2005). Binaural sound localization. In D. L. Wang & G. Brown (Eds.), *Computational auditory scene analysis: Principles, algorithms and applications* (pp. 147–186). Hoboken, NJ: Wiley-IEEE Press.

Suga, N., Yan, J., & Zhang, Y. (1997). Cortical maps for hearing and egocentric selection for self-organization. *Trends in Cognitive Sciences, 1*(1), 13–20.

Sussman, E. S., Horvath, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception and Psychophysics, 69*(1), 136–152.

Trahiotis, C., Bernstein, L. R., Stern, R. M., & Buel, T. N. (2005). Interaural correlation as the basis of a working model of binaural processing: An introduction. In A. N. Popper & R. R. Fay (Eds.), *Sound source localization* (pp. 238–271). New York: Springer Science+Business Media.

van der Kouwe, A. W., Wang, D. L., & Brown, G. J. (2001). A comparison of auditory and blind separation techniques for speech segregation. *IEEE Transactions on Speech and Audio Processing, 9*(3), 189–195.

van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences*. *Ph.D. dissertation*. Eindhoven, The Netherlands: Eindhoven University of Technology.

van Noorden, L. P. A. S. (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.

Van Veen, B. D., & Buckley, K. M. (1988). Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine, 5*(2), 4–24.

Varga, A. P., & Moore, R. K. (1990). Hidden Markov model decomposition of speech and noise. In *Proceedings of the IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, Albuquerque, NM, April 3–6, 1990 (pp. 845–848).

Versnel, H., Kowalski, N., & Shamma, S. A. (1995). Ripple analysis in ferret primary auditory cortex. III. Topographic distribution of ripple response parameters. *Journal of Auditory Neuroscience, 1,* 271–286.

Virtanen, T., Singh, R., & Bhiksha, R. (Eds.). (2012). *Techniques for noise robustness in automatic speech recognition*. Hoboken, NJ: Wiley.

Vliegen, J., & Oxenham, A. J. (1999). Sequential stream segregation in the absence of spectral cues. *The Journal of the Acoustical Society of America, 105*(1), 339–346.

von der Malsburg, C. (1994). The correlation theory of brain function. In E. Domany, L. Van Hemmenm, & K. Schulten (Eds.), *Models of neural networks* (pp. 95–119). Berlin: Springer.

Waibel, A., & Lee, K. (1990). *Readings in speech recognition*. Burlington, MA: Morgan Kaufmann.

Wang, D., & Chang, P. (2008). An oscillatory correlation model of auditory streaming. *Cognitive Neurodynamics, 2*(1), 7–19.

Wang, D. L., & Brown, G. J. (1999). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks, 10*(3), 684–697.

Wang, D. L., & Brown, G. J. (Eds.). (2006). *Computational auditory scene analysis: Principles, algorithms and applications*. Hoboken, NJ: Wiley-IEEE Press.

Weinberger, N. M. (2001). Receptive field plasticity and memory in the auditory cortex: Coding the learned importance of events. In J. Steinmetz, M. Gluck, & P. Solomon (Eds.), *Model systems and the neurobiology of associative learning* (pp. 187–216). Mahwah, NJ: Lawrence Erlbaum Associates.

Weintraub, M. (1985). *A theory and computational model of auditory monaural sound separation*. *Ph.D. dissertation*. Stanford University.

Whiteley, L., & Sahani, M. (2012). Attention in a bayesian framework. *Frontiers in Human Neuroscience*, *6*(100), doi:10.3389/fnhum.2012.00100

Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: Predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences, 13*(12), 532–540.

Xu, Y., & Chun, M. M. (2009). Selecting and perceiving multiple visual objects. *Trends in Cognitive Sciences, 13*(4), 167–174.

Yoon, J. S., Park, J. H., & Kim, H. K. (2009). Acoustic model combination to compensate for residual noise in multi-channel source separation. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, April 19–24, 2009 (pp. 3925–3928).