REVIEW **Communicated by Michael Lewicki**

The Cocktail Party Problem

Simon Haykin
haykin@mcmaster.ca
Zhe Chen
zhechen@soma.crl.mcmaster.ca
Adaptive Systems Lab, McMaster University, Hamilton, Ontario, Canada L8S 4K1

This review presents an overview of a challenging problem in auditory perception, the cocktail party phenomenon, the delineation of which goes back to a classic paper by Cherry in 1953. In this review, we address the following issues: (1) human auditory scene analysis, which is a general process carried out by the auditory system of a human listener; (2) insight into auditory perception, which is derived from Marr's vision theory; (3) computational auditory scene analysis, which focuses on specific approaches aimed at solving the machine cocktail party problem; (4) active audition, the proposal for which is motivated by analogy with active vision, and (5) discussion of brain theory and independent component analysis, on the one hand, and correlative neural firing, on the other.

> One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. This is such a common experience that we may take it for granted; we may call it "the cocktail party problem." No machine has been constructed to do just this, to filter out one conversation from a number jumbled together.
>
> —Colin Cherry, 1957.

1 Introduction

The *cocktail party problem* (CPP), first proposed by Colin Cherry, is a psychoacoustic phenomenon that refers to the remarkable human ability to selectively attend to and recognize one source of auditory input in a noisy environment, where the hearing interference is produced by competing speech sounds or a variety of noises that are often assumed to be independent of each other (Cherry, 1953). Following the early pioneering work (Cherry, 1953, 1957, 1961; Cherry & Taylor, 1954; Cherry & Sayers, 1956, 1959; Sayers & Cherry, 1957), numerous efforts have been dedicated to the CPP in diverse fields: physiology, neurobiology, psychophysiology, cognitive psychology, biophysics, computer science, and engineering. Due to its multidisciplinary nature, it is almost impossible to completely cover this problem in a single

article.[1] Some early partial treatment and reviews of this problem are found in different disciplinary publications (Bregman, 1990; Arons, 1992; Wood & Cowan, 1995; Yost, 1997; Feng & Ratnam, 2000; Bronkhorst, 2000). Half a century after Cherry's seminal work, however, it seems fair to say that a complete understanding of the cocktail party phenomenon is still missing, and the story is far from being complete; the enigma about the marvelous auditory perception capability of human beings remains a mystery. To unveil the mystery and imitate the human performance with a machine, computational neuroscientists, computer scientists, and engineers have attempted to view and simplify this complex perceptual task as a learning problem, for which a tractable computational solution is sought. Despite their obvious simplicity and distinction from reality, the efforts seeking the computational solutions to imitate a human's unbeatable audition capability have revealed that we require a deep understanding of the human auditory system and the underlying neural mechanisms. Bearing such a goal in mind, it does not mean that we must duplicate every aspect of the human auditory system in solving the machine cocktail party problem. Rather, it is our belief that seeking the ultimate answer to the CPP requires deep understanding of many fundamental issues that are deemed to be of theoretical and technical importance. In addition to its obvious theoretical values in different disciplines, the tackling of the CPP will certainly be beneficial to ongoing research on human-machine interfaces.

There are three fundamental questions pertaining to CPP:

1. What is the cocktail party problem?
2. How does the brain solve it?
3. Is it possible to build a machine capable of solving it in a satisfactory manner?

The first two questions are human oriented and mainly involve the disciplines of neuroscience, cognitive psychology, and psychoacoustics; the last question is rooted in machine learning, which involves computer science and engineering disciplines.

In addressing the cocktail party problem, we are interested in three underlying neural processes:[2]

- **Analysis:** The analysis process mainly involves segmentation or segregation, which refers to the segmentation of an incoming auditory signal to individual channels or streams.[3] Among the heuristics used by a

[1] A recently edited volume by Divenyi (2004) discusses several aspects of the cocktail party problem that are complementary to the material covered in this review.

[2] Categorization of these three neural processes, done essentially for research-related studies, is somewhat artificial; the boundary between them is fuzzy in that the brain does not necessarily distinguish between them as defined here.

[3] For an early discussion of the segmentation process, see Moray (1959).

listener to do the segmentation, spatial location is perhaps the most important. Specifically, sounds coming from the same location are grouped together, while sounds originating from other different directions are segregated.

- **Recognition:** The recognition process involves analyzing the statistical structure of the patterns contained in a sound stream that are helpful in recognizing the patterns. The goal of recognition is to uncover the neurobiological mechanisms through which humans are able to identify a segregated sound from multiple streams with relative ease.
- **Synthesis:** The synthesis process involves the reconstruction of individual sound waveforms from the separated sound streams. While synthesis is an important process carried out in the brain (Warren, 1970; Warren, Obusek, & Ackroff, 1972; Bregman, 1990), the synthesis problem is primarily of interest to the machine CPP.

Note also that recognition does not require the analysis process to be perfect; and by the same token, an accurate synthesis does not necessarily mean having solved the analysis and recognition problems, although extra information might provide more hints for the synthesis process.

From an engineering viewpoint, in a loose sense, synthesis may be regarded as the inverse of the combination of analysis and recognition in that it attempts to uncover the relevant attributes of the speech production mechanism. The aim of synthesis is to build a machine that offers the capabilities of operating on a convolved mixture of multiple sources of sounds and to focus attention on the extraction from the convolved mixture a stream of sounds that is of particular interest to an observer; the convolution mentioned here refers to reverberation in a confined environment, which is a hallmark of real-life cocktail party phenomena.

The main theme of this review[4] is philosophical and didactic; hence, no detailed mathematical analysis is presented. The rest of the review is organized as follows. Section 2 discusses human auditory scene analysis. Section 3 discusses the impact of Marr's work in vision (Marr, 1982) on auditory scene analysis (Bregman, 1990). Section 4 presents an overview of computational approaches for solving the cocktail party problem, with an emphasis on independent component analysis, temporal binding and oscillatory correlation, and cortronic network. Section 5 discusses active audition. Discussion of some basic issues pertaining to the cocktail party problem in section 6 concludes the review.

[4] The early version of this review appeared as a presentation made by the first author (Haykin, 2003) and a more lengthy technical report by the second author (Chen, 2003). In this latter report, we presented a comprehensive overview of the cocktail party problem, including a historical account, auditory perceptual processes, relations to visual perception, and detailed descriptions of various related computational approaches.

2 Human Auditory Scene Analysis

Human auditory scene analysis (ASA) is a general process carried out by the auditory system of a human listener for the purpose of extracting information pertaining to a sound source of interest, which is embedded in a background of noise interference.

The auditory system is made up of two ears (constituting the organs of hearing) and auditory pathways. In more specific terms, it is a sophisticated information processing system that enables us to detect not only the frequency composition of an incoming sound but also to locate the sound sources (Kandel, Schwartz, & Jessell, 2000). This is all the more remarkable, given the fact that the energy in the incoming sound waves is exceedingly small and the frequency composition of most sounds is rather complicated.

In this review, our primary concern is with the cocktail party problem. In this context, a complete understanding of the hearing process must include a delineation of where the sounds are located, what sounds are perceived, as well as an explanation of how their perception is accomplished.

2.1 "Where" and "What." The mechanisms in auditory perception essentially involve two processes: sound localization ("where") and sound recognition ("what"). It is well known that (e.g., Blauert, 1983; Yost & Gourevitch, 1987; Yost, 2000) for localizing sound sources in the azimuthal plane, interaural time difference is the main acoustic cue for sound location at low frequencies, and for complex stimuli with low-frequency repetition, interaural level is the main cue for sound localization at high frequencies. Spectral differences provided by the head-related transfer function (HRTF) are the main cues used for vertical localization. Loudness (intensity) and early reflections are the probable cues for localization as a function of distance. In hearing, the precedence effect refers to the phenomenon that occurs during auditory fusion when two sounds of the same order of magnitude are presented dichotically and produce localization of the secondary sound waves toward the outer ear receiving the first sound stimulus (Yost, 2000); the precedence effect stresses the importance of the first wave in determining the sound location.

The "what" question mainly addresses the processes of sound segregation (streaming) and sound determination (identification). Although it has a critical role in sound localization, spatial separation is not considered a strong acoustic cue for streaming or segregation (Bregman, 1990). According to Bregman's studies, sound segregation consists of a two-stage process: feature selection and feature grouping. Feature selection invokes processing the auditory stimuli into a collection of favorable (e.g., frequency sensitive, pitch-related, temporal-spectral-like) features. Feature grouping is responsible for combining similar elements of incoming sounds according to certain principles into one or more coherent streams, with each stream corresponding to one informative sound source. Sound determination is more specific

than segregation in that it not only involves segmentation of the incoming sound into different streams, but also identifies the content of the sound source in question. We will revisit Bregman's viewpoint of human auditory scene analysis in section 3.

2.2 Spatial Hearing. From a communication perspective, our two outer ears act as receive-antennae for acoustic signals from a speaker or audio source. In the presence of one (or fewer) competing or masking sound source, the human ability to detect and understand the source of interest (i.e., target) is degraded. However, the influence of the masking source generally decreases when the target and masker are spatially separated, compared to when the target and masker are in the same location; this effect is credited to spatial hearing (filtering).

In his classic paper, Cherry (1953) suggested that spatial hearing plays a major role in the auditory system's ability to separate sound sources in a multiple-source acoustic environment. Many subsequent experiments have verified Cherry's conjecture. On the other hand, spatial hearing is viewed as one of the important cues that are exploited in solving the CPP (Yost & Gourevitch, 1987) and enhancing speech intelligibility (Hawley, Litovsky, & Colburn, 1999). Spatial separation of the sound sources is also believed to be more beneficial to localization than segregation (Bregman, 1990). But in some cases, spatial hearing is crucial to the sound determination task (Yost, 1991, 1997). Specifically, spatial unmasking produces three effects: (1) pure acoustic effects due to the way sound impinges on the listener's head and body, (2) binaural processing that improves the target signal-to-masker ratio, and (3) central attention whereby the listener can selectively focus attention on a source at a particular direction and block out the competing sources in the unattended directions. The classic book by Blauert (1983) presents a comprehensive treatment of the psychophysical aspect of human sound localization. Given multiple sound sources in an enclosed space (such as a conference room), spatial hearing helps the brain to take full advantage of the slight difference (timing, intensity) between the signals that reach the two outer ears to perform monaural (autocorrelation) and binaural (cross-correlation) processing for specific tasks (such as coincidence detection, precedence detection, localization, and fusion), based on which auditory events are identified and followed by higher-level auditory processing (e.g., attention, streaming, cognition). Figure 1 illustrates a functional diagram of the binaural spatial hearing process.

2.3 Binaural Processing. One of the key observations derived from Cherry's classic experiment (Cherry, 1953) was that it is easier to separate the sources heard binaurally than when they are heard monaurally. Quoting from Cherry and Taylor (1954): "One of the most striking facts about our ears is that we have two of them—and yet we hear one acoustic world; only one voice per speaker." We believe that nature gives us two ears for a

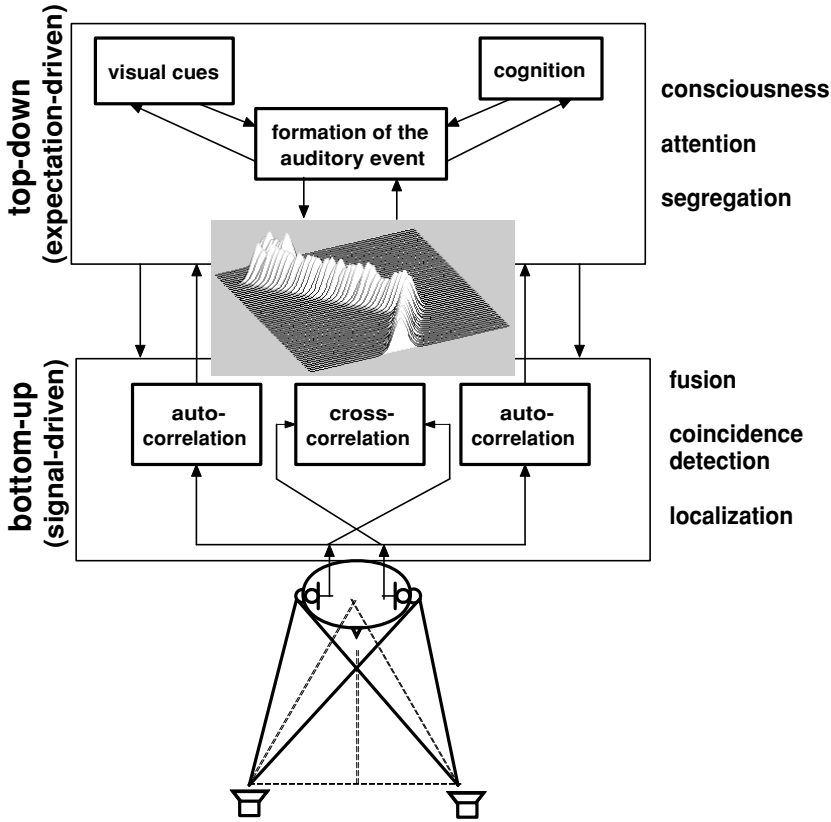

Figure 1: Functional diagram of binaural hearing, which consists of physical, psychophysical, and psychological aspects of auditory perception. (Adapted from Blauert, 1983, with permission.)

reason just like it gives us two eyes. It is the binocular vision (stereovision) and binaural hearing (stereausis) that enable us to perceive the dynamic outer world and provide the main sensory information sources. Binocular/binaural processing is considered to be crucial in certain perceptual activities (e.g., binocular/binaural fusion, depth perception, localization).[5] Given one sound source, the two ears receive slightly different sound patterns due to a finite delay produced by their physically separated locations. The brain is known to be extremely efficient in using varieties of acoustic cues, such as interaural time difference (ITD), interaural intensity difference

[5] We can view sound localization as binaural depth perception, representing the counterpart to binocular depth perception in vision.

(IID), and interaural phase difference (IPD), to perform specific audition tasks. The slight differences in these cues are sufficient to identify the location and direction of the incoming sound waves.

An influential binaural phenomenon is the so-called binaural masking (e.g., Durlach & Colburn, 1978; Moore, 1997; Yost, 2000). The threshold of detecting a signal masked in noise can sometimes be lower when listening with two ears compared to listening with only one, which is demonstrated by a phenomenon called binaural masking level difference (BMLD). It is known (Yost, 2000) that the masked threshold of a signal is the same when the stimuli are presented in a monotic or diotic condition; when the masker and the signal are presented in a dichotic situation, the signal has a lower threshold than in either monotic or diotic conditions. Similarly, many experiments have also verified that binaural hearing increases speech intelligibility when the speech signal and noise are presented dichotically. Another important binaural phenomenon is binaural fusion. Fusion is the essence of directional hearing; the fusion mechanism is often modeled as performing some kind of correlation analysis (Cherry & Sayers, 1956; Cherry, 1961), in which a binaural fusion model based on the autocorrelogram and cross-correlogram was proposed (see Figure 1).

2.4 Attention. Another function basic to human auditory analysis is that of attention, which is a dynamic cognitive process. According to James (1890), the effects of attention include five types of cognitive behavior: (1) perceive, (2) conceive, (3) distinguish, (4) remember, and (5) shorten the reaction time of perceiving and conceiving.

In the context of the cocktail party problem, attention pertains to the ability of a listener to focus attention on one channel while ignoring other irrelevant channels. In particular, two kinds of attention processes are often involved in the cocktail party phenomenon (Jones & Yee, 1993; Yost, 1997):

- Selective attention, in which the listener attends to one particular sound source and ignores other sources
- Divided attention, in which the listener attends to more than one sound source

Once the attention mechanism (either selective or divided) is initiated, a human subject is capable of maintaining attention for a short period of time, hence the term maintained attention.

Another intrinsic mechanism relating to attention is switched attention, which involves the ability of the human brain to switch attention from one channel to another; switched attention is probably mediated in a top-down manner by "gating" the incoming auditory signal (Wood & Cowan, 1995). In this context, a matter of particular interest is the fact that unlike the visual system, whose cortical top-down feedback goes only as far down as the thalamus, the cortical feedback in the auditory system exerts its effect

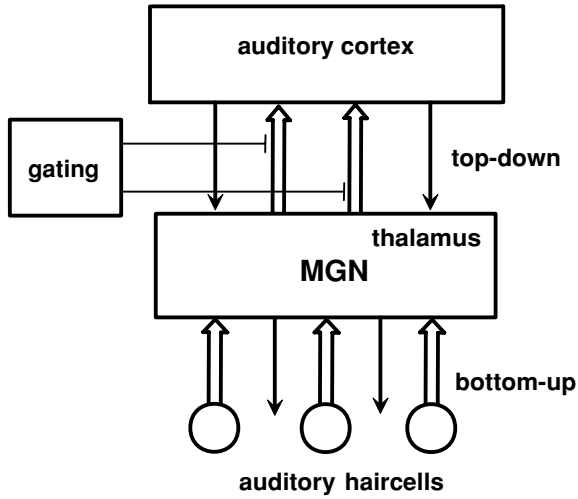

Figure 2: Schematic diagram of an auditory attention circuit.

all the way down to the outer hair cells in the cochlea via the midbrain structure (Wood & Cowan, 1995). Accordingly, the potential for the early selection process of a speech signal of interest in audition is large. Figure 2 is a schematic diagram of the auditory attention circuit. As depicted in the figure, the thalamus acts mainly as a relay station between the sensory hair cells and the auditory cortex.[6]

The bottom-up signals received from the hair cells are sent to medial geniculate nuclei (MGN) in the thalamus and farther up to the auditory cortex through the thalamocortical pathways. The top-down signals from the cortex are sent back to the hair cells through the corticothalamic pathways, to reinforce the signal stream of interest and maximize expectation through feedback.

In addition to auditory scene inputs, visual scene inputs are believed to influence the attention mechanism (Jones & Yee, 1993). For instance, lip-reading is known to be beneficial to speech perception. The beneficial effect is made possible by virtue of the fact that the attention circuit also encompasses cortico-cortical loops between the auditory and visual cortices.

2.5 Feature Binding. One other important cognitive process involved in the cocktail party phenomenon is that of feature binding, which refers to the problem of representing conjunctions of features. According to von der

[6] This is also consistent with the postulate of visual attention mechanism (Crick, 1984; Mumford, 1991, 1995).

Malsburg (1999), binding is a very general problem that applies to all types of knowledge representations, from the most basic perceptual representation to the most complex cognitive representation. Feature binding may be either static or dynamic. Static feature binding involves a representational unit that stands for a specific conjunction of properties, whereas dynamic feature binding involves conjunctions of properties as the binding of units in the representation of an auditory scene, an idea traced back to Treisman and Gelade (1980). The most popular dynamic binding mechanism is based on temporal synchrony, hence the reference to it as "temporal binding." König, Engel, and Singer (1996) suggested that synchronous firing of neurons plays an important role in information processing within the cortex. Rather than being a temporal integrator, the cortical neurons might serve as a coincidence detector evidenced by numerous physiological findings.[7]

Dynamic binding is closely related to the attention mechanism, which is used to control the synchronized activities of different assemblies of units and how the finite binding resource is allocated among the assemblies (Singer, 1993, 1995). Experimental evidence (especially in vision) has shown that synchronized firing tends to provide the attended stimulus with an enhanced representation. Temporal binding hypothesis is attractive, though not fully convincing, in interpreting the perceptual (Gestalt) grouping and sensory segmentation, which has also been evidenced by numerous neurophysiological data (Engel et al., 1991; von der Malsburg, 1999; see also the bibliographies in both works in the special issue on the binding problem).

2.6 Psychophysical and Psychoacoustic Perspectives.

2.6.1 Psychophysical Attributes and Cues. The psychophysical attributes of sound mainly involve three forms of information: spatial location, temporal structure, and spectral characterization. The perception of a sound signal in a cocktail party environment is uniquely determined by this kind of collective information; any difference in any of the three forms of information is believed to be sufficient to discriminate two different sound sources. In sound perception, many acoustic features (cues) are used to perform specific tasks. Table 1 summarizes the main acoustic features (i.e., the temporal or spectral patterns) used for a single-stream sound perception. Combination of a few or more of those acoustic cues is the key to conducting auditory scene analysis. Psychophysical evidence also suggests that significant cues may be provided by spectral-temporal correlations (Feng & Ratnam, 2000). It should be noted that the perception ability with respect to different sound objects (e.g., speech or music) may be different. The fundamental frequencies or tones of sounds are also crucial to perception sensitivity. Experimental results

[7] For detailed discussions of the coincidence detector, see the review papers (Singer, 1993; König & Engel, 1995; König, Engel, & Singer, 1996).

Table 1: The Features (Cues) Used in Sound Perception.

Feature or Cue	Domain	Task
Visual	Spatial	Where
Interaural time difference (ITD)	Spatial	Where
Interaural intensity difference (IID)	Spatial	Where
Intensity (volume), loudness	Temporal	Where + what
Periodicity, rhythm	Temporal	What
Onsets/offsets	Temporal	What
Amplitude modulation (AM)	Temporal	What
Frequency modulation (FM)	Temporal-spectral	What
Pitch	Spectral	What
Timbre, tone	Spectral	What
Hamonicity, formant	Spectral	What

have confirmed that difficulties occur more often in the presence of competing speech signals than in the presence of a single speech and other acoustic sources.

2.6.2 Room Acoustics. For auditory scene analysis, studying the effect of room acoustics on the cocktail party environment is important (Sabine, 1953; MacLean, 1959; Blauert, 1983). A conversation occurring in a closed room often suffers from the multipath effect—mainly echoes and reverberation, which are almost ubiquitous but are rarely consciously noticed. According to the acoustics of the room, a reflection from one surface (e.g., wall, ground) produces reverberation. In the time domain, the reflection manifests itself as smaller, delayed replicas (echoes) that are added to the original sound; in the frequency domain, the reflection introduces a comb-filter effect into the frequency response. When the room is large, echoes can sometimes be consciously heard. However, the human auditory system is so powerful that it can take advantage of binaural and spatial hearing to efficiently suppress the echo, thereby improving the hearing performance.

The acoustic cues listed in Table 1 that are spatially dependent, such as ITD and IID, are naturally affected by reverberation. The acoustic cues that are space invariant, such as common onset across frequencies and pitch, are less sensitive to reverberation. On this basis, it is conjectured that the auditory system has the ability to weight the different acoustic cues (prior to their fusion) so as to deal with a reverberant environment in an effective manner.

3 Insight from Marr's Vision Theory

Audition and vision, the most influential perception processes in the human brain, enable us to absorb the cyclopean information of the outer world in our daily lives. It is well known to neuroscientists that audition (hearing) and

vision (seeing) share substantial common features in the sensory processing principles as well as anatomical and functional organizations in higher-level centers in the cortex.

In his landmark book, David Marr first presented three levels of analysis of information processing systems (Marr, 1982, p. 25):

- Computation: What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?
- Representation: How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?
- Implementation: How can the representation and algorithm be realized physically?

In many perspectives, Marr's observations highlight the fundamental questions that need to be addressed in computational neuroscience, not only in vision but also audition. As a matter of fact, Marr's theory has provided many insights into auditory research (Bregman, 1990; Rosenthal & Okuno, 1998).

In a similar vein to visual scene analysis (e.g., Julesz & Hirsh, 1972), auditory scene analysis (Bregman, 1990) attempts to identify the content ("what") and the location ("where") of the sounds and speech in the environment. In specific terms, auditory scene analysis consists of two stages. In the first stage, the segmentation process decomposes a complex acoustic scene into a collection of distinct sensory elements; in the second stage, the grouping process combines these elements into a stream according to some principles; the streams can be interpreted by a higher-level process for recognition and scene understanding. Motivated by Gestalt psychology, Bregman (1990) has proposed five grouping principles for auditory scene analysis:

1. Proximity, which characterizes the distances between the auditory features with respect to their onsets, pitch, and intensity (loudness)
2. Similarity, which usually depends on the properties of a sound signal, such as timbre
3. Continuity, which features the smoothly varying spectra of a time-varying sound source
4. Closure, which completes fragmentary features that have a good Gestalt; the completion can be understood as an auditory compensation for masking
5. Common fate, which groups together activities (onset, glides, or vibrato) that are synchronous.

Moreover, Bregman (1990) has distinguished at least two levels of auditory organization: primitive streaming and schema-based segregation, with schemas provided by the collected phonetic, prosodic, syntactic, and semantic information. While being applicable to general sound (speech and music) scene analysis, Bregman's work focused mainly on primitive stream segregation. As discussed earlier, auditory scene analysis attempts to solve the analysis and recognition aspects of the CPP.

4 Computational Auditory Scene Analysis

Computational auditory scene analysis (CASA) relies on the development of a computational model of the auditory scene with one of two goals in mind, depending on the application of interest:

1. The design of a machine, which by itself is able to automatically extract and track a sound signal of interest in a cocktail party environment
2. The design of an adaptive hearing system, which automatically computes the perceptual grouping process missing from the auditory system of a hearing-impaired individual, thereby enabling that individual to attend to a sound signal of interest in a cocktail party environment.

Naturally, CASA is motivated by or builds on the understanding we have of human auditory scene analysis.

Following Bregman's seminal work, a number of researchers (Cooke, 1993; Brown, 1992; Ellis, 1996; Cooke & Brown, 1993; Brown & Cooke, 1994; Cooke & Ellis, 2001) have exploited the CASA.[8] In the literature, there are two representative kinds of CASA systems: data-driven system (Cooke, 1993) and prediction-driven system (Ellis, 1996). The common feature in these two systems is to integrate low-level (bottom-up, primitive) acoustic cues for potential grouping. The main differences between them are (Cooke, 2002) that data-driven CASA aims to decompose the auditory scene into time-frequency elements (so-called strands), and then runs the grouping procedure, while prediction-driven CASA regards prediction as the primary goal. It requires only a world model that is consistent with the stimulus; it contains integration of top-down and bottom-up cues and can deal with incomplete or masked data (i.e., speech signal with missing information). However, as emphasized by Bregman (1998), it is important for CASA modelers to take into account psychological data as well as the way humans carry out auditory scene analysis (ASA). For instance, to model the stability of human ASA, the computational system must allow different cues to collaborate and compete and must account for the propagation of constraints across the frequency-by-time field. It is noteworthy that the performances

[8] For review, see Rosenthal and Okuno (1998) and Cooke and Ellis, (2001).

of the CASA approaches are quite dependent on the conditions of noise or interference (such as the spatial location and overlapping time-frequency map), which may therefore be a practical limitation for solving a machine cocktail party problem.

In what follows, we restrict our attention on three major categorized computational approaches aimed at solving the cocktail party problem:[9] (1) blind source separation (BSS) and independent component analysis (ICA), (2) temporal binding and oscillatory correlation, and (3) cortronic network. The first approach has gained a great deal of popularity in the literature, and there is no doubt it will continue to play an important role in neuroscience and signal-processing research; however, the basic ICA approach is limited by its assumptions, and it is arguably biologically implausible in the context of CPP. The second approach is biologically inspired and potentially powerful. The third approach is biologically motivated and knowledge based, and it is configured to solve a realistic machine CPP in a real-life environment.

4.1 Independent Component Analysis and Blind Source Separation. The essence of independent component analysis (ICA) can be stated as follows: given an instantaneous linear mixture of signals produced by a set of sources, devise an algorithm that exploits a statistical discriminant to differentiate these sources so as to provide for the separation of the source signals in a blind (i.e., unsupervised) manner. From this statement, it is apparent that ICA theory and the task of blind source separation (BSS) are intrinsically related.

The earliest reference to this signal processing problem is the article by Jutten and Hérault (1991), which was motivated by Hebb's postulate of learning (1949). This was followed by Comon's article (1994) and that of Bell and Sejnowski (1995). Comon used some signal processing and information-theoretic ideas to formulate a mathematical framework for instantaneous linear mixing of independent source signals, in the course of which the notion of nongaussian ICA was clearly defined. Bell and Sejnowski developed a simple algorithm (Infomax) for BSS, which is inspired by Hebb's postulate of learning and the maximum entropy principle.

It is well known that if we are to achieve the blind separation of an instantaneous linear mixture of independent source signals, then there must be a

[9] It is noteworthy that our overview is by no means exclusive. In addition to the three approaches being reviewed here, several other approaches, some of them quite promising, have been discussed in the literature: Bayesian approaches (e.g., Knuth, 1999; Mohammad-Djafari, 1999; Rowe, 2002; Attias, 1999; Chan, Lee, & Sejnowski, 2003), time-frequency analysis approaches (e.g., Belouchrani & Amin, 1998; Rickard, Balan, & Rosca, 2001; Rickard & Yilmaz, 2002; Yilmaz & Rickard, 2004), and neural network approaches (e.g., Amari & Cichocki, 1998; Grossberg, Govindarajan, Wyse, & Cohen, 2004). Due to space limitation, we have not included these approaches in this article; the interested reader is referred to Chen (2003) for a more detailed overview.

characteristic departure from the simplest possible source model: an independently and identically distributed (i.i.d.) gaussian model. The departure can arise in three different ways, depending on which of the characteristic assumptions embodied in this simple source model is broken, as summarized here (Cardoso, 2001):

- Nongaussian i.i.d. model. In this route to BSS, the i.i.d. assumption for the source signals is retained but the gaussian assumption is abandoned for all the sources, except possibly for one of them. The Infomax algorithm due to Bell and Sejnowski (1995), the natural gradient algorithm due to Amari, Cichocki, and Yang (1996), Cardoso's JADE algorithm (Cardoso & Souloumiac, 1993; Cardoso, 1998), and the FastICA algorithm due to Hyvärinen and Oja (1997) are all based on the nongaussian i.i.d. model. These algorithms differ from each other in the way in which source information residing in higher-order statistics is exploited.
- Gaussian nonstationary model. In this second route to BSS, the gaussian assumption is retained for all the sources, which means that second-order statistics (i.e., mean and variance) are sufficient for characterizing each source signal. Blind source separation is achieved by exploiting the property of nonstationarity, provided that the source signals differ from each other in the ways in which their statistics vary with time. This approach to BSS was first described by Parra and Spence (2000) and Pham and Cardoso (2001). Whereas the algorithms focusing on the nongaussian i.i.d. model operate in the time domain, the algorithms that belong to the gaussian nonstationary model operate in the frequency domain, a feature that also makes it possible for these latter ICA algorithms to work with convolutive mixtures.
- Gaussian, stationary correlated-in-time model. In this third and final route to BSS, the blind separation of gaussian stationary source signals is achieved on the proviso that their power spectra are not proportional to each other. Recognizing that the power spectrum of a wide-sense stationary random process is related to the autocorrelation function via the Wiener-Khintchine theorem, spectral differences among the source signals translate to corresponding differences in correlated-in-time behavior of the source signals. It is this latter property that is available for exploitation. ICA algorithms that belong to this third class include those due to Tong, Soon, Huang, and Liu (1990), Belouchrani, Abed-Meraim, Cardoso, and Moulines (1997), Amari (2000), and Pham (2002).

In Cardoso (2001, 2003), Amari's information geometry is used to explore a unified framework for the objective functions that pertain to these three routes to BSS.

It is right and proper to say that in their own individual ways, Comon's 1994 article and the 1995 article by Bell and Sejnowski, have been the catalysts for the literature in ICA theory, algorithms, and novel applications.[10] Indeed, the literature is so extensive and diverse that in the course of ten years, ICA has established itself as an indispensable part of the ever-expanding discipline of statistical signal processing, and has had a great impact on neuroscience (Brown, Yamada, & Sejnowski, 2001). However, insofar as auditory phenomena are concerned, ICA algorithms do not exploit the merit of spatial hearing; this limitation may be alleviated by adding a spatial filter in the form of an adaptive beamformer (using an array of microphones) as the front-end processor to an ICA algorithm (e.g., Parra & Alvino, 2002). Most important, in the context of the cocktail party problem that is of specific interest to this review, we may pose the following question: Given the ability of the ICA algorithm to solve the BSS problem, can it also solve the cocktail party problem? Our short answer to this fundamental question is no; the rationale is discussed in section 6.

4.2 Temporal Binding and Oscillatory Correlation. Temporal binding theory was most elegantly illustrated by von der Malsburg (1981) in his seminal technical report, Correlation Theory of Brain Function, in which he suggested that the binding mechanism is accomplished by the correlation correspondence between presynaptic and postsynaptic activities, and the strengths of synapses follow the Hebbian postulate of learning. When the synchrony between the presynaptic and postsynaptic neurons is strong (weak), the strength would correspondingly increase (decrease) temporally. Such a synapse was referred to as the "Malsburg synapse" by Crick (1984). The synchronized mechanism allows the neurons to be linked in multiple active groups simultaneously and form a topological network. Moreover, von der Malsburg (1981) suggested a dynamic link architecture to solve the temporal binding problem by letting neural signals fluctuate in time and by synchronizing those sets of neurons that are to be bound together into a higher-level symbol/concept. Using the same idea, von der Malsburg and Schneider (1986) proposed a solution to the cocktail party problem. In particular, they developed a neural cocktail party processor that uses synchronization and desynchronization to segment the sensory inputs. Correlations are generated by an autonomous pattern formation process via neuron coupling and a new synaptic modulation rule. Though based on simple experiments (where von der Malsburg and Schneider used amplitude modulation and stimulus onset synchrony as the main features of the sound, in

[10] Several special issues have been devoted to ICA: Journal of Machine Learning Research (Dec. 2003), Neurocomputing (Nov. 1998, Dec. 2002), Signal Processing (Feb. 1999, Jan. 2004), Proceedings of the IEEE (Oct. 1998), and IEEE Transactions on Neural Networks (March 2004). For textbook treatments of ICA theory, see Hyvärinen, Karhunen, and Oja (2001), and Cichocki and Amari (2002).

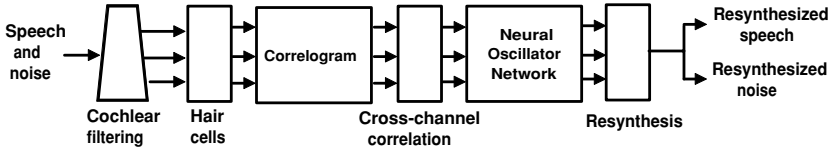

Figure 3: Neural correlated oscillator model from Wang and Brown (1999; adapted with permission).

line with Helmholtz's suggestion), the underlying idea is illuminating. Just as important, the model is consistent with anatomical and physiological observations. Mathematical details of the coupled neural oscillator model were later explored in von der Malsburg and Buhmann (1992). Note that correlation theory is also applicable to the feature binding problem in visual or sensor-motor systems (König & Engel, 1995; Treisman, 1996; von der Malsburg, 1995, 1999).

The idea of oscillatory correlation, as a possible basis for CASA, was motivated by the early work of von der Malsburg (von der Malsburg, 1981; von der Malsburg & Schneider, 1986), and it was extended to different sensory domains whereby phases of neural oscillators are used to encode the binding of sensory components (Wang, Bhumann, & von der Malsburg, 1990; Wang, 1996). Subsequently, Brown and Wang (1997) and Wang and Brown (1999) developed a two-layer oscillator network (see Figure 3) that performs stream segregation based on oscillatory correlation. In the oscillatory correlation-based model, a stream is represented by a population of synchronized relaxation oscillators, each of which corresponds to an auditory feature, and different streams are represented by desynchronized oscillator populations. Lateral connections between oscillators encode the harmonicity and proximity in time and frequency. The aim of the model is to achieve "searchlight attention" by examining the temporal cross-correlation between the activities of pairs (or populations) of neurons:

$$C = \frac{\sum x(t)y(t)}{\sqrt{\sum x^2(t) \sum y^2(t)}},$$

where $x(t)$ and $y(t)$ are assumed to be two zero-mean observable time series.

The neural oscillator model depicted in Figure 3 comprises two layers: a segmentation layer and a grouping layer. The first layer acts as a locally excitatory, globally inhibitory oscillator, and the second layer essentially performs auditory scene analysis. Preceding the oscillator network, there is an auditory periphery model (cochlear and hair cells) as well as a middle-level auditory representation stage (correlogram). As reported by Wang and Brown (1999), the model is capable of segregating a mixture of voiced speech and different interfering sounds, thereby improving the signal-to-noise

ratio (SNR) of the attended speech signal. The correlated neural oscillator is arguably biologically plausible (Wang & Brown, 1999). In specific terms, the neural oscillator acts as a source extraction functional block by treating the attended signal as a "foreground" stream and putting the remaining segments into a "background" stream. However, the performance of the neural oscillator appears to deteriorate significantly in the presence of multiple competitive sources. Recently, van der Kouwe, Wang, and Brown (2001) compared their neural oscillator model with representative BSS techniques for speech segregation in different scenarios; they reported that the performance of the oscillator model varied from one test to another, depending on the time-frequency characteristics of the sources. Under most of the noise conditions, the BSS technique more or less outperformed the oscillator model; however, the BSS techniques worked quite poorly when applied to sources in motion or gaussian sources due to the violation of basic ICA assumptions.

4.3 Cortronic Network. The idea of a so-called cortronic network was motivated by the fact that the human brain employs an efficient sparse coding scheme to extract the features of sensory inputs and accesses them through associative memory. Using a cortronic neural network architecture proposed by Hecht-Nielsen (1998), a biologically motivated connectionist model has been recently developed to solve the machine CPP (Sagi et al., 2001). In particular, Sagi et al. view the CPP as an aspect of the human speech recognition problem in a cocktail party environment, and thereby regard the solution as an attended source identification problem. Only one microphone is used to record the auditory scene; however, the listener is assumed to be familiar with the language of conversation of interest and ignoring other ongoing conversations. All the subjects were chosen to speak the same language and have the same voice qualities. The goal of the cortronic network is to identify one attended speech of interest.

The learning machine described in Sagi et al. (2001) is essentially an associative memory neural network model. It consists of three distinct layers (regions): sound-input representation region, sound processing region, and word processing region. For its operation, the cortronic network rests on two assumptions:

- The network has knowledge of the speech signals (e.g., language context) used.
- The methodology used to design the network resides within the framework of associative memory and pattern identification.

In attacking the machine CPP, the cortronic network undertakes three levels of association (Sagi et al., 2001): (1) sound and subsequent sound, (2) sequence of sounds and the token (i.e., information unit) that is sparsely coded, and (3) a certain word and the word that follows it in the language.

In terms of performance described in Sagi et al., it appears that the cortronic network is quite robust with respect to the variations of speech, speaker, and noise, even under a −8 dB SNR. From one microphone, it can extract one attended speech source in the presence of four additive speech interferences (R. Hecht-Nielsen, personal communication, Sept. 2003). Different from the other computational approaches proposed to solve the CPP, the cortronic network exploits the knowledge context of speech and language; it is claimed to address a theory of how the brain thinks rather than how the brain listens.

5 Active Audition

The three approaches to computational auditory scene analysis described in section 4 share a common feature: the observer merely listens to the environment but does not interact with it (i.e., the observer is passive). In this section, we briefly discuss the idea of active audition in which the observer (human or machine) interacts with the environment. This idea is motivated by the fact that human perception is not passive but active. Moreover, there are many analogies between the mechanisms that go on in auditory perception and their counterparts in visual perception.[11] In a similar vein, we may look to active vision (on which much research has been done for over a decade) for novel ideas in active audition as the framework for an intelligent machine to solve the cocktail party problem.

According to Varela, Thompson, and Rosch (1991) and Sporns (2003), embodied cognitive models rely on cognitive processes that emerge from interactions between neural, bodily, and environment factors. A distinctive feature of these models is that they use the world as their own model. For example, in active vision (also referred to as animated vision), proposed by Bajcsy (1988) and Ballard (1991), among others, it is argued that vision is best understood in the context of visual behaviors. The key point here is that the task of vision is not to build the model of a surrounding real world as originally postulated in Marr's theory, but rather to use visual information in the service of the real world in real time, and do so efficiently and inexpensively (Clark & Eliasmith, 2003). In effect, the active vision paradigm gives "action" a starring role (Sporns, 2003).

With this brief background on active vision, we may now propose a framework for active audition, which may embody four specific functions:

1. Localization, the purpose of which is to infer the directions of incoming sound signals. This function may be implemented by using an

[11] The analogy between auditory and visual perceptions is further substantiated in Shamma (2001), where it is argued that they share certain processing principles: lateral inhibition for edge and peak enhancement, multiscale analysis, and detection mechanisms for temporal coincidence and spatial coincidence.

adaptive array of microphones, whose design is based on direction of arrival (DOA) estimation algorithms developed in the signal-processing literature (e.g., Van Veen & Buckley, 1997; Doclo & Moonen, 2002).

2. Segregation and focal attention, where the attended sound stream of interest (i.e., target sound) is segregated and the sources of interference are ignored, thereby focusing attention on the target sound source. This function may be implemented by using several acoustic cues (e.g., ITD, IID, onset, and pitch) and then combining them in a fusion algorithm.[12]

3. Tracking, the theoretical development of which builds on a state-space model of the auditory environment. This model consists of a process equation that describes the evolution of the state with time and a measurement equation that describes the dependence of the observables on the state. More specifically, the state is a vector defined by the acoustic cues (features) characterizing the target sound stream and its direction.[13] By virtue of its very design, the tracker provides a one-step prediction of the underlying features of the target sound. We may therefore view tracking as a mechanism for dynamic feature binding.

4. Learning, the necessity of which in active audition is the key function that differentiates an intelligent machine from a human brain. Audition is a sophisticated, dynamic information processing task performed in the brain, which inevitably invokes other tasks simultaneously (such as vision and action). It is this unique feature that enables the human to survive in a dynamic environment. For the same reason, it is our belief that an intelligent machine that aims at solving a cocktail party problem must embody a learning capability to adapt itself to an ever-changing dynamic environment. The learning ability must also be of a kind that empowers the machine to take "action" whenever changes in the environment call for it.

Viewed together, these four functions provide the basis for building an embodied cognitive machine that is capable of human-like hearing in an active fashion. The central tenet of active audition embodying such a machine is that an observer may be able to understand an auditory environment more

[12] This approach to segregation and focal attention is currently being pursued by the first author, working with his research colleague Rong Dong at McMaster University.

[13] In the context of tracking, Nix, Kleinschmidt, and Hohmann (2003) used a particle filter as a statistical method for integrating temporal and frequency-specific features of a target speech signal.

effectively and efficiently if the observer interacts with the environment rather than is a passive observer.[14]

6 Discussion

We conclude the article by doing two things. First, we present a philosophical discussion that in the context of the cocktail party phenomenon, ICA algorithm addresses an entirely different signal processing problem. Second, in an attempt to explain how the brain solves the CPP, we postulate a biologically motivated correlated firing framework for single-stream extraction based on our own recent work.

6.1 Brain Theory and ICA. In the course of over ten years, the idea of ICA has blossomed into a new field that has enriched the discipline of signal processing and neural computation. However, when the issue of interest is a viable solution to the cocktail party problem, the ICA/BSS framework has certain weaknesses:

- Most ICA/BSS algorithms require that the number of sources not be fewer than the number of independent signal sources.[15] In contrast, the human auditory system requires merely two outer ears to solve the cocktail party problem, and it can do so with relative ease.
- The matter of independent signal sources is assumed to remain constant in the ICA/BSS framework. This is an unrealistic assumption in a neurobiological context. For instance, it is possible for an auditory environment to experience a varying number of speakers (i.e., sound sources) or a pulse-like form of noise (e.g., someone laughing or coughing), yet the human capability to solve the cocktail party problem remains essentially unaffected by the variations in the auditory scene.
- Last but by no means least, the ICA/BSS framework usually requires the separation of all source signals. With both outer ears of the human auditory systems focused on a signal speaker of interest in a complex auditory scene, the cocktail party problem is solved by extracting that speaker's speech signal and practically suppressing all other forms of noise or interference.

[14] In this concluding statement on active audition, we have paraphrased the essence of active vision (Blake & Yuille, 1992), or more generally, that of active perception (Bajcsy, 1988).

[15] In the literature, there are also some one-microphone BSS approaches that attempt to exploit either the acoustic features and masking technique (Roweis, 2000), the splitting of time-domain disjoint/orthogonal subspaces (Hopgood & Rayner, 1999), or the prior knowledge of the source statistics (Jang & Lee, 2003).

Simply put, viewed in the context of the cocktail party problem, ICA/BSS algorithms seem not to be biologically plausible. Rather, we say that for a computational auditory scene analysis framework to be neurobiologically feasible, it would have to accommodate the following ingredients:

- Ability to operate in a nonstationary convolutive environment, where the speech signal of interest is corrupted by an unknown number of competing speech signals or sources of noise.
- Ability to switch the focus of attention from one speech signal of interest to another and do so with relative ease.
- Working with a pair of sensors so as to exploit the benefit of binaural hearing. However, according to Sagi et al. (2001), it is claimed that a single sensor (reliance on monaural hearing) is sufficient to solve the cocktail party problem. There is no contradiction here between these two statements, as demonstrated by Cherry and coworkers over 50 years ago: simply put, binaural hearing is more effective than monaural hearing by working with a significantly smaller SNR.
- Parallelism, which makes it possible for the incoming signal to be worked on by a number of different paths, followed by a fusion of their individual outputs.

6.2 Correlative Neural Firing for Blind Single-Source Extraction. Correlation theory has played an influential role on memory recall, coincidence detection, novelty detection, perception, and learning, which cover most of the intelligent tasks of a human brain (Cook, 1991). Eggermont (1990), in his insightful book, has presented a comprehensive investigation of correlative activities in the brain; in that book, Eggermont argued that correlation, in one form or another, is performed in 90 percent of the human brain. Eggermont (1993) also investigated the neural correlation mechanisms (such as coincidence detection and tuned-delay mechanisms) in the auditory system and showed that synchrony is crucial in sound localization, pitch extraction of music, and speech coding (such as intensity-invariant representation of sounds, suppressing less salient features of speech sound, and enhancing the representation of speech formant).

Recently, we (Chen, 2005; Chen & Haykin, 2004) proposed a stochastic-correlative firing mechanism and an associated learning rule for solving a simplified form of CPP. The idea of correlative firing mechanism is similar to that of synchrony and correlation theory (von der Malsburg, 1981; Eggermont, 1990) and, interestingly enough, it is motivated by some early work on vision (Harth, Unnikrishnan, & Pandya, 1987; Mumford, 1995). In our proposed correlative firing framework, the auditory cortex implements a certain number of parallel circuits, each responsible for extracting the attended sound source of interest. By analogy with figure-ground segregation, the circuit extracts the "figure" from a complex auditory scene. To model the selective attention in the thalamus, a gating network is proposed for

deciding or switching the attention of the segregated sources, as depicted in the schematic diagram of Figure 2. Specifically, the proposed stochastic correlative learning rule (Chen, 2005; Chen & Haykin, 2004) can be viewed as a variant of the ALOPEX (ALgorithm Of Pattern EXtraction), an optimization procedure that was originally developed in vision research (Harth & Tzanakou, 1974; Tzanakou, Michalak, & Harth, 1979). The stochastic correlative learning rule is temporally asymmetric and Hebbian-like. The most attractive aspect of this learning rule lies in the fact that it is both gradient free and model independent, which make it potentially applicable to any neural architecture that has hierarchical or feedback structures (Haykin, Chen, & Becker, 2004); it also allows us to develop biologically plausible synaptic rules based on the firing rate of the spiking neurons (Harth et al., 1987). Another appealing attribute of this learning rule is its parallelism in that synaptic plasticity allows a form of synchronous firing, which lends itself to ease of hardware implementation. The proposed correlative firing framework for the CPP assumes a time-invariant instantaneous linear mixing of the sources as in the ICA theory. However, unlike ICA, our method is aimed at extracting the "figure" (i.e., one single stream), given two sensors (corresponding to two ears), from the "background" (auditory scene) containing more than two nongaussian sound sources; additionally, our interpretation of mixing coefficients in the mixing matrix as neurons' firing rates (instead of acoustic mixing effect) is motivated from the notion of the firing-rate stimulus correlation function in neuroscience (Dayan & Abbott, 2001), and the learning rule aimed at extracting a single source that bears the highest synchrony in terms of neurons' firing rate. In simulated experiments reported in Chen and Haykin, (2004), we have demonstrated a remarkable result: the algorithm is capable of extracting one single source stream given only two sensors and more than four sources (including one gaussian source). This is indeed an intractable task for ICA algorithms. Our results emphasize that the main goal of the solution to the CPP is not to separate out the competing sources but rather to extract the source signal of interest (or "figure") in a complex auditory scene that includes competing speech signals and noise.

To conclude, our proposed stochastic correlative neural firing mechanism, embodying its own learning rule, is indeed aimed at blind signal source extraction. Most important, it provides a primitive yet arguably convincing basis for how the human auditory system solves the cocktail party problem; that is, it addresses question 2 of the introductory section. However, in its present form, the learning rule is not equipped to deal with question 3 pertaining to computational auditory scene analysis.

Acknowledgments

The work reported in this review is supported by the Natural Sciences and Engineering Research Council of Canada. It was carried out under a joint

Canada-Europe BLISS project. We thank the editor and two anonymous reviewers for constructive suggestions and valuable comments that have helped us reshape the review into its present form. We also thank P. Divenyi, S. Grossberg, S. Shamma, S. Becker, and R. Dong for helpful comments on various early drafts of this review and J. J. Eggermont, R. Hecht-Nielsen, and W. A. Yost for kindly supplying some references pertinent to the cocktail party problem. We also greatly appreciate certain authors and publishers for permission to reproduce or adapt Figures 1 and 3 in the review.

References

Amari, S. (2000). Estimating functions of independent component analysis for temporally correlated signals. *Neural Computation, 12*(9), 2083–2107.

Amari, S., & Cichocki, A. (1998). Adaptive blind signal processing—neural network approaches. *Proceedings of the IEEE, 86*(10), 2026–2048.

Amari, S., Cichocki, A., & Yang, H. H. (1996). A new learning algorithm for blind signal separation. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems, 8* (pp. 757–763). Cambridge, MA: MIT Press.

Arons, B. (1992). A review of the cocktail party effect. *Journal of the American Voice I/O Society, 12*, 35–50.

Attias, H. (1999). Independent factor analysis. *Neural Computation, 11*, 803–851.

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE, 76*, 996–1005.

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence, 48*, 57–86.

Bell, A., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*, 1120–1159.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., & Moulines, E. (1997). A blind source separation technique based on second order statistics. *IEEE Transactions on Signal Processing, 45*(2), 434–444.

Belouchrani, A., & Amin, M. G. (1998). Blind source separation technique based on time-frequency representations. *IEEE Transactions on Signal Processing, 46*(11), 2888–2897.

Blake, A., & Yuille, A. (Eds.). (1992). *Active vision*. Cambridge, MA: MIT Press.

Blauert, J. (1983). *Spatial hearing: The psychophysics of human sound localization* (rev. ed.). Cambridge, MA: MIT Press.

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

Bregman, A. S. (1998). Psychological data and computational ASA. In D. F. Rosenthal & H. G. Okuno (Eds.), *Computational auditory scene analysis*. Mahwah, NJ: Erlbaum.

Bronkhorst, A. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker condition. *Acoustica, 86*, 117–128.

Brown, G. J. (1992). *Computational auditory scene analysis: A representational approach*. Unpublished doctoral dissertation, University of Sheffield.

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

Brown, G. J., & Wang, D. L. (1997). Modelling the perceptual segregation of concurrent vowels with a network of neural oscillation. *Neural Networks, 10*(9), 1547–1558.

Brown, G. D., Yamada, S., & Sejnowski, T. J. (2001). Independent component analysis at the neural cocktail party. *Trends in Neuroscience, 24*, 54–63.

Cardoso, J.-F. (1998). Blind signal separation: Statistical principles. *Proceedings of the IEEE, 86*(10), 2029–2025.

Cardoso, J.-F. (2001). The three easy routes to independent component analysis: Contrasts and geometry. In *Proc. ICA2001*. San Diego.

Cardoso, J.-F. (2003). Dependence, correlation and gaussianity in independent component analysis. *Journal of Machine Learning Research, 4*, 1177–1203.

Cardoso, J.-F., & Souloumiac, A. (1993). Blind beamforming for non-Gaussian signals. *IEE Proceedings–F, 140*(6), 362–370.

Chan, K., Lee, T.-W., & Sejnowski, T. J. (2003). Variational Bayesian learning of ICA with missing data. *Neural Computation, 15*(8), 1991–2011.

Chen, Z. (2003). *An odyssey of the cocktail party problem* (Tech. Rep.). Hamilton, Ontario: Adaptive Systems Lab, McMaster University. Available online: http://soma.crl.mcmaster.ca/~zhechen/download/cpp.ps

Chen, Z. (2005). Stochastic correlative firing for figure-ground segregation. *Biological Cybernetics*, *92*(3), 192–198.

Chen, Z., & Haykin, S. (2004). *Figure-ground segregation in sensory perception using a stochastic correlative learning rule* (Tech. Rep.). Hamilton, Ontario: Adaptive Systems Lab, McMaster University. Available online: http://soma.crl.mcmaster.ca/~zhechen/download/scl.ps.

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

Cherry, E. C. (1957). *On human communication: A review, survey, and a criticism*. Cambridge, MA: MIT Press.

Cherry, E. C. (1961). Two ears—but one world. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 99–117). New York: Wiley.

Cherry, E. C., & Sayers, B. (1956). Human "cross-correlation"—A technique for measuring certain parameters of speech perception. *Journal of the Acoustical Society of America, 28*, 889–895.

Cherry, E. C., & Sayers, B. (1959). On the mechanism of binaural fusion. *Journal of the Acoustical Society of America, 31*, 535.

Cherry, E. C., & Taylor, W. K. (1954). Some further experiments upon the recognition of speech, with one and, with two ears. *Journal of the Acoustical Society of America, 26*, 554–559.

Cichocki, A., & Amari, S. (2002). *Adaptive blind signal and image processing*. New York: Wiley.

Clark, A., & Eliasmith, C. (2003). Philosophical issues in brain theory and connectionism. In M. Arbib (Ed.), *Handbook of brain theory and neural networks* (2nd ed., pp. 886–888). Cambridge, MA: MIT Press.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing, 36*, 287–314.

Cook, J. E. (1991). Correlated activity in the CNS: A role on every timescale? *Trends in Neuroscience, 14*, 397–401.

Cooke, M. (1993). *Modelling auditory processing and organization*. Cambridge: Cambridge University Press.

Cooke, M. P. (2002, December). *Computational auditory scene analysis in listeners and machines*. Tutorial at NIPS2002, Vancouver, Canada.

Cooke, M. P., & Brown, G. J. (1993). Computational auditory scene analysis: Exploiting principles of perceived continuity. *Speech Communication, 13*, 391–399.

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

Crick, F. (1984). Function of the thalamic reticular complex: The searchlight hypothesis. *Proc. Natl. Acad. Sci. USA, 81*, 4586–4590.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: MIT Press.

Doclo, S., & Moonen, M. (2002). Robust adaptive time-delay estimation for speaker localization in noisy and reverberant acoustic environments. *EURASIP Journal of Applied Signal Processing, 5*, 2230–2244.

Divenyi, P. (Ed.). (2004). *Speech separation by humans and machines*. Berlin: Springer.

Durlach, N. I., & Colburn, H. S. (1978). Binaural phenomena. In E. C. Cartrette & M. P. Friedman (Eds.), *Handbook of perception*. New York: Academic Press.

Eggermont, J. J. (1990). *The correlative brain: Theory and experiment in neural interaction*. New York: Springer-Verlag.

Eggermont, J. J. (1993). Function aspects of synchrony and correlation in the auditory nervous system. *Concepts in Neuroscience, 4*(2), 105–129.

Ellis, D. (1996). *Prediction-driven computational auditory scene analysis*. Unpublished doctoral dissertation, MIT.

Engel, A. K., König, P., & Singer, W. (1991). Direct physiological evidence for scene segmentation by temporal coding. *Proc. Natl. Acad. Sci. USA, 88*, 9136–9140.

Feng, A. S., & Ratnam, R. (2000). Neural basis of hearing in real-world situations. *Annual Review of Psychology, 51*, 699–725.

Grossberg, S., Govindarajan, K., Wyse, L. L., & Cohen, M. A. (2004). ARTSTREAM: A neural network model of auditory scene analysis and source segregation. *Neural Networks, 17*(4), 511–536.

Harth, E., & Tzanakou, E. (1974). Alopex: A stochastic method for determining visual receptive fields. *Vision Research, 14*, 1475–1482.

Harth, E., Unnikrishnan, K. P., & Pandya, A. S. (1987). The inversion of sensory processing by feedback pathways: A model of visual cognitive functions. *Science, 237*, 184–187.

Hawley, M. L., Litovsky, R. Y., & Colburn, H. S. (1999). Speech intelligibility and localization in a multisource environment. *Journal of the Acoustical Society of America, 105*, 3436–3448.

Haykin, S. (2003, June). *The cocktail party phenomenon*. Presentation at the ICA Workshop, Berlin, Germany.

Haykin, S., Chen, Z., & Becker, S. (2004). Stochastic correlative learning algorithms. *IEEE Transactions on Signal Processing, 52*(8), 2200–2209.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.

Hecht-Nielsen, R. (1998). A theory of cerebral cortex. In *Proc. 1998 Int. Conf. Neural Information Processing, ICONIP'98* (pp. 1459–1464). Burke, VA: IOS Press.

Hopgood, E., & Rayner, P. (1999). Single channel signal separation using linear time-varying filters: Separability of non-stationary stochastic signals. In *Proc. ICASSP* (Vol. 3, pp. 1449–1452). Piscataway, NJ: IEEE Press.

Hyvärinen, A., Karhunen, V., & Oja, E. (2001). *Independent component analysis*. New York: Wiley.

Hyvärinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation, 9*, 1483–1492.

James, W. (1890). *Psychology (briefer course)*. New York: Holt.

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

Jones, M., & Yee, W. (1993). Attending to auditory events: The role of temporal organization. In S. McAdams & E. Bigand (Eds.), *Thinking in sound* (pp. 69–106). Oxford: Clarendon Press.

Julesz, B., & Hirsh, I. J. (1972). Visual and auditory perception—an essay in comparison. In E. David & P. B. Denes (Eds.), *Human communication: A unified view*. New York: McGraw-Hill.

Jutten, C., & Hérault, J. (1991). Blind separation of sources, part I–III. *Signal Processing, 24*, 1–29.

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (Eds.). (2000). Hearing. In *Principles of neural science* (4th ed.). New York: McGraw-Hill.

Knuth, K. H. (1999). A Bayesian approach to source separation. In *Proc. ICA'99* (pp. 283–288). Aussois, France.

König, P., & Engel, A. K. (1995). Correlated firing in sensory-motor systems. *Current Opinions in Neurobiology, 5*, 511–519.

König, P., Engel, A. K., & Singer, W. (1996). Integrator or coincidence detector? The role of the cortical neuron revisited. *Trends in Neuroscience, 19*(4), 130–137.

MacLean, W. R. (1959). On the acoustics of cocktail parties. *Journal of the Acoustical Society of America, 31*(1), 79–80.

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

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

Mohammad-Djafari, A. (1999). A Bayesian approach to source separation. In *Proc. 19th Int. Workshop on Maximum Entropy and Bayesian Methods (MaxEnt99)*. Boise, ID.

Moore, B. C. J. (1997). *An introduction to the psychology of hearing* (4th ed.). San Diego: Academic Press.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology, 27*, 56–60.

Mumford, D. (1991). On the computational architecture of the neocortex. Part I: The role of the thalamocortical loop. *Biological Cybernetics, 65*, 135–145.

Mumford, D. (1995). Thalamus. In M. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 981–984). Cambridge, MA: MIT Press.

Nix, J., Kleinschmidt, M., & Hohmann, V. (2003). Computational auditory scene analysis of cocktail-party situations based on sequential Monte Carlo methods. In *Proc. 37th Asilomar Conference on Signals, Systems and Computers* (pp. 735–739). Redondo Beach, CA: IEEE Computer Society Press.

Parra, L., & Alvino, C. (2002). Geometric source separation: Merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing, 10*(6), 352–362.

Parra, L., & Spence, C. (2000). Convolutive blind source separation of nonstationary sources. *IEEE Transactions on Speech and Audio Processing, 8*(3), 320–327.

Pham, D. T. (2002). Mutual information approach to blind separation of stationary sources. *IEEE Transactions on Information Theory, 48*(7), 1935–1946.

Pham, D. T., & Cardoso, J. F. (2001). Blind separation of instantaneous mixtures of non-stationary sources. *IEEE Transactions on Signal Processing, 49*(9), 1837–1848.

Rickard, S., Balan, R., & Rosca, J. (2001). Real-time time-frequency based blind source separation. In *Proc. ICA2001* (pp. 651–656). San Diego, CA.

Rickard, S., & Yilmaz, O. (2002). On the approximate *W*-disjoint orthogonality of speech. In *Proc. ICASSP2002* (pp. 529–532). Piscataway, NJ: IEEE Press.

Rosenthal, D. F., & Okuno, H. G. (Eds.). (1998). *Computational auditory scene analysis*. Mahwah, NJ: Erlbaum.

Rowe, D. B. (2002). A Bayesian approach to blind source separation. *Journal of Interdisciplinary Mathematics, 5*(1), 49–76.

Roweis, S. (2000). One microphone source separation. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems, 13* (pp. 793–799). Cambridge, MA: MIT Press.

Sabine, H. J. (1953). Room acoustics. *Transactions of IRE, 1*, 4–12.

Sagi, S., Nemat-Nasser, S. C., Kerr, R., Hayek, R., Downing, C., & Hecht-Nielsen, R. (2001). A biologically motivated solution to the cocktail party problem. *Neural Computation, 13*, 1575–1602.

Sayers, B., & Cherry, E. C. (1957). Mechanism of binaural fusion in the hearing of speech. *Journal of the Acoustical Society of America, 31*, 535.

Schultz, S. R., Golledge, H. D. R., & Panzeri, S. (2001). Synchronization, binding and the role of correlated firing in fast information transmission. In S. Wermter, J. Austin, & D. Willshaw (Eds.), *Emergent neural computational architectures based on neuroscience*. Berlin: Springer-Verlag.

Shamma, S. (2001). On the role of space and time in auditory processing. *Trends in Cognitive Sciences, 5*(8), 340–348.

Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology, 55*, 349–374.

Singer, W. (1995). Synchronization of neural responses as putative binding mechanism. In M. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 960–964). Cambridge, MA: MIT Press.

Sporns, O. (2003). Embodied cognition. In M. Arbib (Ed.), *Handbook of brain theory and neural networks* (2nd edition, pp. 395–398). Cambridge, MA: MIT Press.

Tong, L., Soon, V., Huang, Y., & Liu, R. (1990). AMUSE: A new blind identification problem. In *Proc. ICASSP* (pp. 1784–1787). Piscataway, NJ: IEEE Press.

Treisman, A. M. (1996). The binding problem. *Current Opinion in Neurobiology, 6*, 171–178.

Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97–136.

Tzanakou, E., Michalak, R., & Harth, E. (1979). The Alopex process: Visual receptive fields by response feedback. *Biological Cybernetics, 35*, 161–174.

van der Kouwe, A. J. 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*, 189–195.

Van Veen, B., & Buckley, K. (1997). Beamforming techniques for spatial filtering. In V. K. Madisetti & D. B. Williams (Eds.), *Digital signal processing handbook*. Boca Raton, FL: CRC Press.

Varela, F., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.

von der Malsburg, C. (1981). *The correlation theory of brain function*. (Internal Rep. 81-2). Göttingen: Department of Neurobiology, Max-Plank-Institute for Biophysical Chemistry.

von der Malsburg, C. (1995). Binding in models of perception and brain function. *Current Opinion in Neurobiology, 5*, 520–526.

von der Malsburg, C. (1999). The what and why of binding: The modeler's perspective. *Neuron, 24*, 95–104.

von der Malsburg, C., & Buhmann, J. (1992). Sensory segmentation with coupled neural oscillators. *Biological Cybernetics, 67*, 233–242.

von der Malsburg, C., & Schneider, W. (1986). A neural cocktail-party processor. *Biological Cybernetics, 54*, 29–40.

Wang, D. L. (1996). Primitive auditory segregation based on oscillatory correlation. *Cognitive Science, 20*(3), 409–456.

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., Buhmann, J., & von der Malsburg, C. (1990). Pattern segmentation in associative memory. *Neural Computation, 2*, 94–106.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167*, 392–393.

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

Wood, N. L., & Cowan, N. (1995). The cocktail party phenomenon revisited: Attention and memory in the classic listening selective procedure of Cherry (1953). *Journal of Experimental Psychology, General, 124*, 243–262.

Yilmaz, O., & Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing, 52*, 1830–1847.

Yost, W. A. (1991). Auditory image perception and analysis. *Hearing Research, 46*, 8–18.

Yost, W. A. (1997). The cocktail party problem: Forty years later. In R. Gilkey & T. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments* (pp. 329–348). Ahwah, NJ: Erlbaum.

Yost, W. A. (2000). *Fundamentals of hearing: An introduction* (4th ed.). San Diego: Academic Press.

Yost, W. A., & Gourevitch, G. Eds. (1987). *Directional hearing*. New York: Springer-Verlag.

Received April 1, 2004; accepted February 23, 2005.