Bregman, A.S. (2005) Foreword to Divenyi, P. (Ed.). *Speech separation by humans and machines.* Norwell, MA: Kluwer Academic Publishers.

There is a serious problem in the recognition of sounds. It derives from the fact that they do not usually occur in isolation but in an environment in which a number of sound sources (voices, traffic, footsteps, music on the radio, and so on) are active at the same time. When these sounds arrive at the ear of the listener, the complex pressure waves coming from the separate sources add together to produce a single, more complex pressure wave that is the sum of the individual waves. The problem is how to form separate mental descriptions of the component sounds, despite the fact that the "mixture wave" does not directly reveal the waves that have been summed to form it.

The name *auditory scene analysis* (ASA) refers to the process whereby the auditory systems of humans and other animals are able to solve this mixture problem. The process is believed to be quite general, not specific to speech sounds or any other type of sounds, and to exist in many species other than humans. It seems to involve assigning spectral energy to distinct "auditory objects" and "streams" that serve as the mental representations of distinct sound sources in the environment and the patterns that they make as they change over time. How this energy is assigned will affect the perceived number of auditory sources, their perceived timbres, loudnesses, positions in space, and pitches. Indeed, every perceived property studied by psychoacoustics researchers seems to be affected by the partitioning of spectral energy. While the name ASA refers to the competence of humans and other animals, the name *computational* auditory scene analysis (CASA) refers to the attempt by scientists to program computers to solve the mixture problem.

In 2003, Pierre Divenyi put together an interdisciplinary workshop that was held in Montreal that autumn, a meeting focused on the topic of how to separate a speech signal from interfering sounds (including other speech). It is obvious why this topic is so important. Right now speech recognition by computers is a delicate process, easily derailed by the presence of interfering sounds. If methods could be evolved to focus recognition on just those components of the signal that came from a targeted source, recognition would be more robust and usable for human-computer interaction in a wide variety of environments. Yet, albeit of overwhelming importance, speech separation represents only a part of the more general ASA problem, the study of which may shed light on issues especially relevant to speech understanding in interference. It was therefore appropriate that Divenyi assembled members of a number of disciplines working on the problem of the separation of concurrent sounds: experimental psychologists studying how ASA was done by people, both for speech and non-speech sounds, neuroscientists interested in how the brain deals with sounds, as well as computer scientists and engineers developing computer systems to solve the problem. This book is a fascinating collection of their views and ideas on the problem of speech separation.

My personal interest in these chapters is that they bring to forefront the argument of special import to me as a cognitive psychologist. This argument, made by CASA researchers, is that since people can do sound separation quite well, a better understanding of *how* they do it will lead to better strategies for designing computer programs that can solve the same problem.

Others however, disagree with this argument, and want to accomplish sound segregation using any powerful signal-processing method that can be designed from scientific and mathematical principles, without regard for how humans do it. This difference in strategy leads one to ask the following question: Will one approach ultimately wipe out the other or will there always be a place for both? Maybe we can take a lesson from the ways in which humans and present-day computer systems are employed in the solving of problems. Humans are capable of solving an enormous variety of problems (including how to program computers to solve problems). However, they are slow, don't always solve the problems, and are prone to error. In contrast, a computer program is typically designed to carry out a restricted range of computations in a closed domain (e.g., statistical tests), but can do them in an error-free manner at blinding speeds. It is the "closedness" of the domain that permits a strict algorithmic solution, leading to the blinding speed and the absence of error. So we tend to use people when the problems reside in an "open" domain and computers when the problem domain is closed and well-defined. (It is possible that when computers become as all purpose and flexible in their thought as humans, they will be as slow and as subject to error as people are.)

The application of this lesson about general-purpose versus specialized computation to auditory scene analysis by computer leads to the conclusion that we should use general methods, resembling those of humans, when the situation is unrestricted – for example when both a robotic listener and a number of sound sources can move around, when the sound may be coming around a corner, when the component sounds may not be periodic, when substantial amounts of echo and reverberation exist, when objects can pass in front of the listener casting acoustic shadows, and so on. On the other hand, we may be able to use faster, more error-free algorithms when the acoustic situation is more restricted.

If we accept that specialized, algorithmic methods won't always be able to solve the mixture problem, we may want to base our general CASA methods on how people segregate sounds. If so, we need a better understanding of how human (and animal) nervous systems solve the problem of mixture. Achieving this understanding is the role that the experimental psychologists and the neuroscientists play in the CASA enterprise.

The present book represents the best overview of current work in the fields of ASA and CASA and should inspire researchers with an interest in sound to get involved in this exciting interdisciplinary area. Pierre Divenyi deserves our warmest thanks for his unstinting efforts in bringing together scientists of different orientations and for assembling their contributions to create the volume that you are now reading.

Albert S. Bregman
Professor Emeritus of Psychology
McGill University