Foreword to
DeLiang Wang & Guy Brown (Eds.)
"Computational auditory scene analysis: Principles, algorithms, and application."
IEEE / Wiley 2006

This book provides exactly what is needed by a young scientist or engineer who wants to get started in the exciting new field of computational auditory scene analysis (CASA). In describing what the book is about, it is helpful to start by considering a particular difficult problem that a human faces in everyday listening. A number of sounds are present at the same time, but the listener is interested in only one of them. How can the desired sound be isolated from the mixture?

When you ask people how they can listen to only one of the sounds in a mixture, a typical listener might answer, “I just listen to one and try not to be distracted by the others.” Notice that this answer presupposes that a number of separate sounds have already been discerned and the only problem is to listen to just one of them. It illustrates the limitations of introspection, because the scientist knows facts about sound that our typical listener does not. The first is that although each sound-producing event in the environment radiates its own pattern of pressure waves into the air, these individual sound-wave patterns are summed together on each of the two eardrums of the listener. Consequently each eardrum vibrates in a pattern that is the sum of all the wave patterns that reach it. Unfortunately, this summed waveform does not have written on it how many sounds have been added together to create it. There is an infinite number of possible ways in which simpler sounds could have been added together to make this sum. So the listener’s brain has to solve the difficult problem of deriving the actual waveforms that have contributed to the mixture – a process known as auditory scene analysis (ASA).

Years of research have shown that human listeners solve the ASA problem by exploiting the regularities of the world. Let me give an example. There are many *periodic* sounds in the world, among them the sounds of musical instruments, buzzes of insects, animal vocalizations, and many of the sounds present in human speech. For example, a vowel, produced on a certain pitch, is a complex tone which can be analyzed into a set of harmonics – pure tones whose frequencies are multiples of a fundamental frequency. When two vowels are heard at the same time, the ear encounters a set of components that consists of a mixture of the two sets of harmonics. This fact about harmonics is a regularity of the world, which does not yet have anything to do with human hearing. However, the human auditory system, in analyzing the incoming mixture, can look for sets of components that relate to different fundamentals. Suppose it finds that a majority of the incoming components can be assigned to two subsets, one whose fundamental is 110 Hz (110, 220, 330, 440, …) and a second whose fundamental is 160 Hz (160, 320, 480, …), how many distinct sounds should it conclude that it is hearing? The obvious answer is two. It is extremely unlikely that the frequency components present in a single sound or in several concurrent sounds would line up into exactly two sets of harmonics. This use of the regularity found in harmonic sounds is an example of exploiting the regularities of the world – inferring the number of acoustic sources from the properties of the acoustic input, under the assumption that the input is a mixture of natural sounds . However, not all sounds are harmonic; so we have to use other properties of a mixture that arise because it is the sum of a number of distinct natural sounds, probably separated in space. Over the years we have found out what many of these properties are and have studied how they are used by human listeners.

Our success in doing so has raised a number of interesting questions:

- Do animals other than humans perform ASA on auditory inputs, and, if so, do they exploit the same environmental regularities as humans do, or are there other regularities specific to the animals' own environments?
- What is the neurological basis of ASA?
- Do very young infants already perform ASA as adults do, or must this be learned?
- How do composers take advantage of the listener's ASA to control whether musical sounds will blend with, or stand out from, one another?
- When mixtures involve speech sounds, are they subject to the same ASA processes as mixtures of non-speech sounds?
- Can the use of sound in data display (sonification) profit from knowledge about human ASA?
- How can auditory prostheses help their users to segregate sounds?

The challenge of answering these questions has attracted the attention of researchers in a number of disciplines, and promising starts have already been made.

There is another question of tremendous practical importance: Is it possible to program computers to perform ASA? There have been many interesting approaches to this problem, and the evolving research field has been named "computational auditory scene analysis" (CASA). Solving this problem would have valuable applications. Computer programs for speech recognition tend to run into serious trouble when other sounds accompany the target speech. Solving the problem of mixtures would allow speech-recognition programs to operate in complex acoustic environments so that they could, for instance, interact verbally with a user, either face-to-face or by telephone, and respond to commercial orders, database queries, or other messages. Research on CASA can also serve theoretical purposes. For example, if computer systems utilize methods similar to those proposed as explanations of human ASA, their success or failure can serve as a test of the adequacy of these explanations, and can raise new questions for research on humans and other animals.

Many scientific meetings have been held on the subject of CASA, bringing together researchers from different institutions and countries. The present volume has been preceded by other books on the subject, but these have been reports of the research meetings on CASA, providing accounts of specific pieces of research or individual approaches to this problem. The present book is unique. Although written by leading researchers in the field, it is designed as a textbook, giving overall accounts of the different approaches to CASA with only as much detail as needed for clarity of exposition. The reader's access to the field as a whole is greatly facilitated by the comprehensiveness of the chapters and by the inclusion of extensive bibliographies (in the neighbourhood of 100 references per chapter). The broad coverage plus the abundant references provide everything that needed to rapidly come up to speed on CASA research. The editors and the other authors are to be congratulated on putting together a volume that should prove to be extremely useful in opening up this field to a large number of students and researchers.

Albert S. Bregman
December 4, 2005