Book Review

Computational Auditory Scene Analysis: Principles, Algorithms, and Applications—D. Wang and G. J. Brown, Eds. (New York, NY: Wiley/IEEE Press, 2006, pp. 395, ISBN: 0-471-74109-4). *Reviewed by J. Rouat*

Computational auditory scene analysis (CASA) is an interdisciplinary area derived from the field of auditory scene analysis (ASA). ASA describes and takes into account the perceptual organization of audition in the context of concurrent and/or simultaneous sounds (mixtures of sound). CASA systems attempt to implement and/or to study some of the ASA principles by using machinery for application designs. In fact, a remarkable feat of the auditory system is its ability to disentangle the acoustic mixtures and group the acoustic features from the same event or auditory source. In his seminal book [1], Bregman exposes the ASA principles and his view of auditory perception of mixtures of sound. Based on analogies with the analysis of visual scenes, Bregman proposes a complete framework for the analysis of auditory scenes [1].

The main audience to this book is a scientist or a graduate student anxious to grasp rapidly this multidisciplinary area. The book comprises chapters written by ten researchers in the field of auditory modeling, speech analysis and recognition, statistical signal processing, sound localization, reverberation processing, music processing, and networks of oscillatory or firing neurons. It is invaluable to the newcomer in the field who would like to start his own research or needs to know the main activities in CASA. To the senior researcher in CASA, the book can be used to track some of the main contributions in the field because of the rich bibliography. To some points, part of the material is strongly grounded into engineering and explores avenues that reach the same goal as the original ASA (that strongly integrates perception). Therefore, the book is addressed to a wide audience (engineers will learn about perception and specialists in perception will learn about engineering to process mixture of sounds).

A very interesting feature of the book is the existence of the website http://www.casabook.org/, from which it is possible to download sound examples or Matlab software for some chapters.

After the foreword by Bregman, the first chapter (44 pages) by Wang and Brown gives a brief overview of the human auditory system (considering a passive cochlea) and of the basic perceptual features and mechanisms in the context of auditory scene analysis. Then, they define CASA and give examples of cochleagram, correlogram, time-frequency masking, and resynthesis when using the perceptive approach to CASA. They finally discuss the difficulty to evaluate CASA systems and conclude with a brief history of CASA.

The second chapter (35 pages) is written by de Cheveigné who makes a good overview of the problem of estimating pitch for single and multiple F0 signals. The chapter recalls some of the fundamental works in single pitch estimation (spectral, temporal, and spectro-temporal approaches) and then enlarges the scope to multiple F0 signals. A full section is then devoted to the discussion of the limits and issues of current systems (time, frequency resolutions, estimation of the number of sources, performance evaluation, etc.).

In chapter three (34 pages), Wang first discusses and lists the main features that are used in a bottom-up process to characterize the signal cues (pitch and onset/offset detections, amplitude, and frequency modulations). He illustrates on examples of auditory scene segmentation (cross channel and temporal correlation/continuity). Then, two full sections are devoted to the difficult task of grouping the cues and associating them to the right auditory sources.

In chapter four (32 pages), Ellis introduces the reader to implementations of higher level auditory processing (*schema driven* according to [1]—it can be viewed as a way to integrate top-down knowledge in a CASA system) based essentially on Bayesian models. The hidden Markovian models of source separation are presented. Then, constraints, fitting models, training, signal representations, and various other issues (such as the complexity) are discussed. Relations to other statistical source separation systems are briefly discussed.

Chapter five (40 pages) is written by Stern, Brown, and Wang. It introduces the reader to the physical and physiological cues related to interaural time difference (ITD), interaural intensity difference (IID), and head related transfer functions (HRTF). Then, the examples of computer models with single or multiple sources localization are given and discussed.

Chapter six (21 pages) written by Feng and Jones first describes the conventional engineering array of microphone techniques for sound source localization and then the author's implementation of a binaural source localization (azimuth) system based on ITD. Finally, Feng and Jones discuss other models that also use IIDs.

Chapter seven (42 pages) written by Brown and Palomäki gives a good insight on the problems encountered in source localization and/or separation in presence of reverberation. After defining reverberation, the authors report results from the literature on the effects of reverberation on machines (speech recognition and source separation). They analyze the principal mechanisms underlying robustness to reverberation in humans. They conclude by analyzing few examples of CASA implementations in the context of reverberation.

Chapter eight (45 pages) written by Goto is first built around Goto's work on the predominant-F0 estimation method for musical signals, and then it describes related work from the literature. The second half of the chapter is devoted to the extraction of the acoustical features from mono and polyphonic music.

Chapter nine written by Baker is the longest chapter (54 pages) of the book with a very good grasp on the state-of-the-art speech recognition in relation with CASA and ASA. Section 2 positions ASA and Bregman's work in relation with ongoing works in the field of speech perception. Section 3 analyzes works in automatic speech recognition based on the usual engineering approach. Section 4 discusses the integration of CASA in the automatic speech recognizers (ASR) and details one solution based on the missing features approach. The author concludes by discussing the integration of the missing-data mask and the choice of the features for implementation in conventional speech recognizers.

Chapter ten (37 pages) written by Brown and Wang is entirely dedicated to the neural implementation of CASA systems. It introduces the reader to modelization of spiking or oscillatory neurons and to the implementation of binding with networks of neurons. They also link the neural network approach with auditory models in the context of CASA. They use the examples taken from their own work in the field.

The reviewer is with the NECOTIS, Electrical and Computer Engineering Department, Université de Sherbrooke, Sherbrooke, QC, J1K 2R1, Canada (e-mail: jean.rouat@ieee.org).

Digital Object Identifier 10.1109/TNN.2007.913988

References

[1] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.