Chapter 2
Modeling Grouping Cues for Auditory Scene Analysis Using a Spectral Clustering Formulation

Luis Gustavo Martins
Portuguese Catholic University, Portugal

Mathieu Lagrange
CNRS - Institut de Recherche et Coordination Acoustique Musique (IRCAM), France

George Tzanetakis
University of Victoria, Canada

ABSTRACT
Computational Auditory Scene Analysis (CASA) is a challenging problem for which many different approaches have been proposed. These approaches can be based on statistical and signal processing methods such as Independent Component Analysis or can be based on our current knowledge about human auditory perception. Learning happens at the boundary interactions between prior knowledge and incoming data. Separating complex mixtures of sound sources such as music requires a complex interplay between prior knowledge and analysis of incoming data. Many approaches to CASA can also be broadly categorized as either model-based or grouping-based. Although it is known that our perceptual-system utilizes both of these types of processing, building such systems computationally has been challenging. As a result most existing systems either rely on prior source models or are solely based on grouping cues. In this chapter the authors argue that formulating this integration problem as clustering based on similarities between time-frequency atoms provides an expressive yet disciplined approach to building sound source characterization and separation systems and evaluating their performance. After describing the main components of such an architecture, the authors describe a concrete realization that is based on spectral clustering of a sinusoidal representation. They show how this approach can be used to model both traditional grouping cues such as frequency and amplitude continuity as well as other types of information and prior knowledge such as onsets, harmonicity and timbre-models for specific instruments.

DOI: 10.4018/978-1-61520-919-4.ch002
Experiments supporting their approach to integration are also described. The description also covers issues of software architecture, implementation and efficiency, which are frequently not analyzed in depth for many existing algorithms. The resulting system exhibits practical performance (approximately real-time) with consistent results without requiring example-specific parameter optimization and is available as part of the Marsyas open source audio processing framework.

INTRODUCTION

Inspired by the classic book by Bregman on Auditory Scene Analysis (ASA) (Bregman, 1990) a variety of systems for Computational Auditory Scene Analysis (CASA) have been proposed (Wang and Brown, 2006). They can be broadly classified as bottom-up systems (or data-driven) where the flow of information is from the incoming audio signal to higher level representations or top-down systems (or model-based) where prior-knowledge about the characteristics of a particular type of sound source in the form of a model is utilized to assist the analysis. The human auditory system utilizes both of these types of processing. Although it has been argued that computational CASA systems should also utilize both types (Slaney, 1998) most existing systems fall into only one of the two categories. Another related challenge is the integration of several grouping cues that operate simultaneously into a single system. We believe that this integration becomes particularly challenging when the CASA system has a multiple stage architecture where each stage corresponds to a particular grouping cue or type of processing. In such architectures any errors in one stage propagate to the following stages and it is hard to decide what the ordering of stages should be. An alternative, which we advocate in this chapter, is to formulate the entire sound source formation problem from a complex sound mixture as a clustering based on similarities of time-frequency atoms across both time and frequency. That way all cues are taken into account simultaneously and new sources of information such as source models or other types of prior-knowledge can easily be taken into consideration using one unifying formulation.

Humans, even without any kind of formal music training, are typically able to extract, almost unconsciously, a great amount of relevant information from a musical signal. Features such as the beat of a musical piece, the main melody of a complex musical arrangement, the sound sources and events occurring in a complex musical mixture, the song structure (e.g. verse, chorus, bridge) and the musical genre of a piece, are just some examples of the level of knowledge that a naive listener is commonly able to extract just from listening to a musical piece. In order to do so, the human auditory system uses a variety of cues for perceptual grouping such as similarity, proximity, harmonicity, common fate, among others.

In the past few years interest in the emerging research area of Music Information Retrieval (MIR) has been steadily growing. It encompasses a wide variety of ideas, algorithms, tools, and systems that have been proposed to handle the increasingly large and varied amounts of musical data available digitally. Typical MIR systems for music signals in audio format represent statistically the entire polyphonic sound mixture (Tzanetakis and Cook, 2002). There is some evidence that this approach has reached a “glass ceiling” (Aucouturier and Pachet, 2004) in terms of retrieval performance. One obvious direction for further progress is to attempt to individually characterize the different sound sources comprising the polyphonic mixture. The predominant melodic voice (typically the singer in western popular music) is arguably the most important sound source and its separation has a large number of applications in Music Information Retrieval.