Chapter 7
Probabilistic Modeling Paradigms for Audio Source Separation

Emmanuel Vincent
INRIA, France

Maria G. Jafari
Queen Mary University of London, UK

Samer A. Abdallah
Queen Mary University of London, UK

Mark D. Plumbley
Queen Mary University of London, UK

Mike E. Davies
University of Edinburgh, UK

ABSTRACT

Most sound scenes result from the superposition of several sources, which can be separately perceived and analyzed by human listeners. Source separation aims to provide machine listeners with similar skills by extracting the sounds of individual sources from a given scene. Existing separation systems operate either by emulating the human auditory system or by inferring the parameters of probabilistic sound models. In this chapter, the authors focus on the latter approach and provide a joint overview of established and recent models, including independent component analysis, local time-frequency models and spectral template-based models. They show that most models are instances of one of the following two general paradigms: linear modeling or variance modeling. They compare the merits of either paradigm and report objective performance figures. They also, conclude by discussing promising combinations of probabilistic priors and inference algorithms that could form the basis of future state-of-the-art systems.

DOI: 10.4018/978-1-61520-919-4.ch007
INTRODUCTION

Many everyday sound scenes are produced by several concurrent sound sources: spoken communications are often obscured by background talkers, outdoor recordings feature a variety of environmental sounds, and most music recordings involve a group of several instruments. When facing such scenes, humans are able to perceive and listen to individual sources so as to communicate with other speakers, navigate in a crowded street or memorize the melody of a song (Wang and Brown, 2006). Source separation aims to provide machine listeners with similar skills by extracting the signals of individual sources from a given mixture signal. The estimated source signals may then be either listened to or further processed, giving rise to many potential applications such as speech enhancement for hearing aids, automatic speech and speaker recognition in adverse conditions, automatic indexing of large audio databases, 5.1 rendering of stereo recordings and music post-production.

Depending on the application, the notion of “source” may differ. For instance, musical instruments accompanying a singer may be considered as multiple sources or fused into a single source (Ozerov, Philippe, Bimbot, & Gribonval, 2007). Hence some minimal prior knowledge about the sources is always needed to address the separation task. In certain situations, information such as source positions, speaker identities or musical score may be known and exploited by informed source separation systems. In many situations however, only the mixture signal is available and blind source separation systems must be employed that do not rely on specific characteristics of the processed scene.

A first approach to audio source separation called computational auditory scene analysis (CASA) is to emulate the human auditory source formation process (Wang and Brown, 2006). Typical CASA systems consist of four processing stages. The signal is first transformed into a time-frequency-lag representation. Individual time-frequency bins are then clustered into small clusters, each associated with one source, by applying primitive auditory grouping and streaming rules. These rules state for example that sinusoidal sounds should be clustered together when they have harmonic frequencies, a smooth spectral envelope, similar onset and offset times, correlated amplitude and frequency modulations, and similar interchannel time and intensity differences. The resulting clusters are further processed using schema-based grouping rules implementing knowledge acquired by learning, such as the timbre of a known speaker or the syntax of a particular language, until a single cluster per source is obtained. The source signals are eventually extracted by associating each time-frequency bin with a single source and inverting the time-frequency transform, an operation known as binary time-frequency masking. Although some processing rules may explicitly or implicitly derive from probabilistic priors (Ellis, 2006), the overall process is deterministic: predefined rules implementing complementary knowledge are applied in a fixed precedence order. This bottom-up strategy allows fast processing, but relies on the assumption that each time-frequency bin is dominated by a single source. When this assumption is not satisfied, clustering errors might occur during early processing stages and propagate through subsequent stages.

An alternative approach to source separation is to rely on top-down generative models of the mixture signal that incorporate knowledge about the sound production process. The theory of Bayesian signal processing (Gelman, Carlin, Stern, & Rubin, 2003) provides an appropriate framework to build and exploit such models. The probabilistic distribution of a class of mixture signals is specified by a set of latent variables, including the source signals, and conditional prior distributions between these variables, that are either fixed or learned from training data. For a given mixture signal, the value of any variable may be estimated
Related Content

Hidden Markov Model Based Visemes Recognition, Part II: Discriminative Approaches
[www.igi-global.com/chapter/hidden-markov-model-based-visemes/31074?camid=4v1a](www.igi-global.com/chapter/hidden-markov-model-based-visemes/31074?camid=4v1a)

Issues and Challenges in Building Multilingual Information Access Systems
[www.igi-global.com/chapter/issues-challenges-building-multilingual-information/70068?camid=4v1a](www.igi-global.com/chapter/issues-challenges-building-multilingual-information/70068?camid=4v1a)

Hidden Markov Model Based Visemes Recognition, Part I: AdaBoost Approach
[www.igi-global.com/chapter/hidden-markov-model-based-visemes/31073?camid=4v1a](www.igi-global.com/chapter/hidden-markov-model-based-visemes/31073?camid=4v1a)

Dependency Parsing in Bangla
[www.igi-global.com/chapter/dependency-parsing-bangla/78475?camid=4v1a](www.igi-global.com/chapter/dependency-parsing-bangla/78475?camid=4v1a)