Chapter 10

Audio Source Separation Using Sparse Representations

Andrew Nesbit
Queen Mary University of London, UK

Maria G. Jafari
Queen Mary University of London, UK

Emmanuel Vincent
INRIA, France

Mark D. Plumbley
Queen Mary University of London, UK

ABSTRACT

The authors address the problem of audio source separation, namely, the recovery of audio signals from recordings of mixtures of those signals. The sparse component analysis framework is a powerful method for achieving this. Sparse orthogonal transforms, in which only few transform coefficients differ significantly from zero, are developed; once the signal has been transformed, energy is apportioned from each transform coefficient to each estimated source, and, finally, the signal is reconstructed using the inverse transform. The overriding aim of this chapter is to demonstrate how this framework, as exemplified here by two different decomposition methods which adapt to the signal to represent it sparsely, can be used to solve different problems in different mixing scenarios. To address the instantaneous (neither delays nor echoes) and underdetermined (more sources than mixtures) mixing model, a lapped orthogonal transform is adapted to the signal by selecting a basis from a library of predetermined bases. This method is highly related to the windowing methods used in the MPEG audio coding framework. In considering the anechoic (delays but no echoes) and determined (equal number of sources and mixtures) mixing case, a greedy adaptive transform is used based on orthogonal basis functions that are learned from the observed data, instead of being selected from a predetermined library of bases. This is found to encode the signal characteristics, by introducing a feedback system between the bases and the observed data. Experiments on mixtures of speech and music signals demonstrate that these methods give good signal approximations and separation performance, and indicate promising directions for future research.

DOI: 10.4018/978-1-61520-919-4.ch010
Audio Source Separation Using Sparse Representations

INTRODUCTION

The problem of audio source separation involves recovering individual audio source signals from a number of observed mixtures of those simultaneous audio sources. The observations are often made using microphones in a live recording scenario, or can be taken, for example, as the left and right channels of a stereo audio recording. This is a very challenging and interesting problem, as evidenced by the multitude of techniques and principles used in attempts to solve it. Applications of audio source separation and its underlying principles include audio remixing (Woodruff, Pardo, & Dannenberg, 2006), noise compensation for speech recognition (Benaroya, Bimbot, Gravier, & Gribonval, 2003), and transcription of music (Bertin, Badeau, & Vincent, 2009). The choice of technique used is largely governed by certain constraints on the sources and the mixing process. These include the number of mixture channels, number of sources, nature of the sources (e.g., speech, harmonically related musical tracks, or environmental noise), nature of the mixing process (e.g., live, studio, using microphones, echoic, anechoic, etc), and whether or not the sources are moving in space.

The type of mixing process that generates the observed sources is crucially important for the solution of the separation problem. Typically, we distinguish between instantaneous, anechoic and convolutive mixing. These correspond respectively to the case where the sources are mixed without any delays or echoes, when delays only are present, and when both echoes and delays complicate the mixing. Source separation for the instantaneous mixing case is generally well understood, and satisfactory algorithms have been proposed for a variety of applications. Conversely, the anechoic and convolutive cases present bigger challenges, although they often correspond to more realistic scenarios, particularly for audio mixtures recorded in real environments. Algorithms for audio source separation can also be classified as blind or semi-blind, depending on whether a priori information regarding the mixing. Blind methods assume that nothing is known about the mixing, and the separation must be carried out based only on the observed signals. Semi-blind methods incorporate a priori knowledge of the mixing process (Jafari et al., 2006) or the sources’ positions (Hesse & James, 2006).

The number of mixture channels relative to the number of sources is also very important in audio source separation. The problem can be overdetermined, when more mixtures than sources exist, determined, with equal number of mixtures and sources, and underdetermined, when we have more sources than mixtures. Since the overdetermined problem can be reduced to a determined problem (Winter, Sawada, & Makino, 2006), only the determined and underdetermined situations have to be considered. The latter is particularly challenging, and conventional separation methods alone cannot be applied. An overview of established, statistically motivated, model-based separation approaches are presented elsewhere in this book (Vincent et al., 2010), which can also serve as an introduction to audio source separation for the non-expert reader. Another useful introduction is the review article by O’Grady, Pearlmutter, & Rickard (2005).

A widely used class of model-based source separation algorithms that exploits the sparsity of the source signals in some time-frequency (TF) transform domain is sparse component analysis (SCA). It entails transforming the signals into a domain in which they are sparse, estimating the mixing matrix from the transform coefficients, estimating the source representations, and finally, inverting the transform representation of the estimated sources. A sparse signal representation is one which conveys the information within the signal using only a few elementary components, denoted as atoms, which are selected from a dictionary to form a sparse signal decomposition. This often helps to uncover hidden structure in the analysed signal by characterising the original signal using only a small number of large coef-
17 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:

www.igi-global.com/chapter/audio-source-separation-using-sparse/45488?camid=4v1

This title is available in InfoSci-Intelligent Technologies, InfoSci-Books, Science, Engineering, and Information Technology, InfoSci-Computer Science and Information Technology, InfoSci-Select. Recommend this product to your librarian:

www.igi-global.com/e-resources/library-recommendation/?id=16

Related Content

Audio and Speech Watermarking and Quality Evaluation
www.igi-global.com/chapter/audio-speech-watermarking-quality-evaluation/4686?camid=4v1a

Audition: From Sound to Sounds
www.igi-global.com/chapter/audition-sound-sounds/45482?camid=4v1a

Differentially Fed Artificial Neural Networks for Speech Signal Prediction
www.igi-global.com/chapter/differentially-fed-artificial-neural-networks/4690?camid=4v1a

Applying NLP Metrics to Students' Self-Explanations
www.igi-global.com/chapter/applying-nlp-metrics-students-self/61053?camid=4v1a