

Machine Audition: Principles, Algorithms and Systems

Wenwu Wang
University of Surrey, UK

Information Science
REFERENCE

INFORMATION SCIENCE REFERENCE

Hershey • New York

Detailed Table of Contents

Preface	xv
Acknowledgment	xxi

Section 1

Audio Scene Analysis, Recognition and Modeling

Chapter 1

Unstructured Environmental Audio: Representation, Classification and Modeling.....	1
--	---

Selina Chu, University of Southern California, USA

Shrikanth Narayanan, University of Southern California, USA

C.-C. Jay Kuo, University of Southern California, USA

The goal of this chapter is on the characterization of unstructured environmental sounds for understanding and predicting the context surrounding of an agent or device. Most research on audio recognition has focused primarily on speech and music. Less attention has been paid to the challenges and opportunities for using audio to characterize unstructured audio. This chapter investigates issues in characterizing unstructured environmental sounds such as the development of appropriate feature extraction algorithms and learning techniques for modeling backgrounds of the environment.

Chapter 2

Modeling Grouping Cues for Auditory Scene Analysis using a Spectral Clustering Formulation	22
--	----

Luis Gustavo Martins, Portuguese Catholic University, Porto Portugal

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

George Tzanetakis, University of Victoria, Canada

Computational Auditory Scene Analysis (CASA) is challenging problem, to which many approaches can be broadly categorized as either model-based or grouping-based. 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 separa-

tion systems and evaluating their performance. They describe the main components of the architecture, its advantages, implementation details, and related issues.

Chapter 3

Cocktail Party Problem: Source Separation Issues and Computational Methods 61

Tariqullah Jan, University of Surrey, UK

Wenwu Wang, University of Surrey, UK

Cocktail party problem is a classical and challenging scientific problem that is still unsolved. Many efforts have been attempted by researchers to address this problem using different techniques. This chapter provides a review on recent progresses in several areas, such as independent component analysis, computational auditory scene analysis, model-based approaches, non-negative matrix factorization, sparse representation and compressed sensing. As an example, a multistage approach is also provided for addressing the source separation issue within this problem.

Chapter 4

Audition: From Sound to Sounds 80

Tjeerd C. Andringa, University of Groningen, The Netherlands

The demand to function in uncontrolled listening environments has severe implications for machine audition. The natural system has addressed this demand by adapting its function flexibly to changing task demands. This chapter addresses the functional requirements of auditory systems, both natural and artificial, to be able to deal with the complexities of uncontrolled real-world input. Signal processing methods that are needed for such scenarios are also discussed.

Section 2

Audio Signal Separation, Extraction and Localization

Chapter 5

A Multimodal Solution to Blind Source Separation of Moving Sources..... 107

Syed Mohsen Naqvi, Loughborough University, UK

Yonggang Zhang, Harbin Engineering University, China

Miao Yu, Loughborough University, UK

Jonathon A. Chambers, Loughborough University, UK

Machine separation of moving audio sources is a challenging problem. This chapter presents a novel multimodal solution to blind source separation (BSS) of moving sources, where the visual modality is utilized to facilitate the separation of moving sources. The movement of the sources is detected by a relatively simplistic 3-D tracker based on video cameras. The tracking process is based on particle filtering which provides robust tracking performance. Positions and velocities of the sources are obtained from the 3-D tracker and if the sources are moving, real time speech enhancement and separation of the sources are obtained by using a beamforming algorithm.

Chapter 6

Sound Source Localization: Conventional Methods and Intensity Vector Direction Exploitation..... 126

Banu Günel, University of Surrey, UK

Hüseyin Hacıhabiboğlu, King's College London, UK

Automatic sound source localization may refer to determining only the direction of a sound source, which is known as the direction-of-arrival estimation, or also its distance in order to obtain its coordinates. Many of the methods proposed previously use the time and level differences between the signals captured by each element of a microphone array. This chapter presents an overview of these conventional array processing methods and a discussion of the factors that affect their performance. The chapter also discusses an emerging source localization method based on acoustic intensity, and addresses two well-known problems, localization of multiple sources and localization of acoustic reflections.

Chapter 7

Probabilistic Modeling Paradigms for Audio Source Separation 162

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

Source separation aims to provide machine listeners with similar skills to humans by extracting the sounds of individual sources from a given audio scene. Existing separation systems operate either by emulating the human auditory system or 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, and they compare the merits of either paradigm, report objective performance figures and discuss promising combinations of probabilistic priors and inference algorithms.

Chapter 8

Tensor Factorization with Application to Convolutional Blind Source Separation of Speech..... 186

Saeid Sanei, Cardiff University, UK

Bahador Makkiabadi, Cardiff University, UK

In this chapter, Tensor factorization (TF) is introduced to the problem of separation of sound particularly speech sources from their corresponding convolutional mixtures. TF is flexible and can easily incorporate all possible parameters or factors into the separation formulation. As a consequence of that fewer assumptions (such as uncorrelatedness and independency) will be required. The new formulation allows further degree of freedom to the original parallel factor analysis (PARAFAC) problem in which the scaling and permutation problems of the frequency domain blind source separation (BSS) can be resolved.

Chapter 9

Multi-Channel Source Separation: Overview and Comparison of Mask-Based and Linear Separation Algorithms	207
<i>Nilesh Madhu, Ruhr-Universität Bochum, Germany</i>	
<i>André Gückel, Dolby Laboratories - Nürnberg, Germany</i>	

Machine-based multi-channel source separation in real life situations is a challenging problem, and has a wide range of applications, from medical to military. This chapter considers the specific application of a target speaker enhancement in the presence of competing speakers and background noise. It presents not only an exhaustive overview of state-of-the-art separation algorithms and the specific models they are based upon, but also the relations between these algorithms, where possible. In particular, it compares the performance difference between the mask-based techniques and the independent component analysis (ICA) techniques.

Chapter 10

Audio Source Separation using Sparse Representations	246
<i>Andrew Nesbit, Queen Mary University of London, UK</i>	
<i>Maria G. Jafari, Queen Mary University of London, UK</i>	
<i>Emmanuel Vincent, INRIA, France</i>	
<i>Mark D. Plumbley, Queen Mary University of London, UK</i>	

The authors address the problem of audio source separation based on the sparse component analysis framework. The overriding aim of this chapter is to demonstrate how this framework can be used to solve different problems in different mixing scenarios. To address the instantaneous and underdetermined mixing model, a lapped orthogonal transform is adapted to the signal by selecting a basis from a library of predetermined bases. In considering the anechoic and determined mixing case, a greedy adaptive transform is used based on orthogonal basis functions that are learned from the observed data. The chapter also demonstrates the good signal approximations and separation performance by these methods using experiments on mixtures of speech and music signals.

Section 3

Audio Transcription, Mining and Information Retrieval

Chapter 11

Itakura-Saito Nonnegative Factorizations of the Power Spectrogram for Music Signal Decomposition	266
<i>Cédric Févotte, TELECOM ParisTech, France</i>	

This chapter presents a nonnegative matrix factorization (NMF) technique for audio decomposition by considering factorization of the power spectrogram, with the Itakura-Saito (IS) divergence. The author shows that IS-NMF is connected to maximum likelihood inference of variance parameters in a well-defined statistical model of superimposed Gaussian components which is well suited to audio. The chapter further discusses the model order selection strategies and Markov regularization of the

activation matrix. Extensions of NMF to the multichannel case, in both instantaneous and convolutive recordings, possibly underdetermined, together with audio source separation results of a real stereo musical excerpt are also included.

Chapter 12

Music Onset Detection..... 297

Ruohua Zhou, Queen Mary University of London, UK

Josh D. Reiss, Queen Mary University of London, UK

The authors provide a comprehensive introduction to the design of music onset detection algorithms. First, it introduces the general scheme and commonly-used time-frequency analysis for onset detection. Then, it reviews many methods for onset detection in detail, such as energy-based, phase-based, pitch-based and supervised learning methods. The chapter also includes commonly used performance measures, onset annotation software, public database, and evaluation methods.

Chapter 13

On the Inherent Segment Length in Music 317

Kristoffer Jensen, Aalborg University Esbjerg, Denmark

This chapter presents automatic segmentation methods using different original representations of music, corresponding to rhythm, chroma, and timbre, and by calculating a shortest path through the self-similarity calculated from each time/feature representation. Each segmentation scale quality is analyzed through the use of the mean silhouette value, which permits automatic segmentation on different time scales and gives indication on the inherent segment sizes in the music analyzed. Different methods are employed to verify the quality of the inherent segment sizes, by comparing them to the literature (grouping, chunks), by comparing them among themselves, and by measuring the strength of the inherent segment sizes.

Chapter 14

Automatic Tagging of Audio: The State-of-the-Art..... 334

Thierry Bertin-Mahieux, Columbia University, USA

Douglas Eck, University of Montreal, Canada

Michael Mandel, University of Montreal, Canada and Columbia University, USA

A great deal of attention has been paid recently to the automatic prediction of tags for music and audio in general. In the case of music, social tags have become an important component of ``Web 2.0'' recommender systems. In an effort to better understand the task and also to help new researchers bring their insights to bear on this problem, this chapter provides a review of the state-of-the-art methods for addressing automatic tagging of audio. It is divided in the following sections: goal, framework, audio representation, labeled data, classification, evaluation, and future directions.

Chapter 15

Instantaneous vs. Convolutive Non-Negative Matrix Factorization: Models, Algorithms and Applications to Audio Pattern Separation 353

Wenwu Wang, University of Surrey, UK

Non-negative matrix factorization (NMF) has been shown recently to be a useful technique for audio decomposition. However, the instantaneous NMF model has difficulty in dealing with the audio signals whose frequencies change dramatically over time, which is nevertheless in practice a case for many real signals. This chapter intends to provide a brief overview of the models and algorithms for both instantaneous and convolutive NMF, with a focus on the theoretical analysis and performance evaluation of the convolutive NMF algorithms, and their applications to audio pattern separation problems.

Section 4
Audio Cognition, Modeling and Affective Computing

Chapter 16

Musical Information Dynamics as Models of Auditory Anticipation..... 371
Shlomo Dubnov, University of California in San Diego, USA

This chapter investigates the modeling methods for musical cognition. The author explores possible relations between cognitive measures of musical structure and statistical signal properties that are revealed through information dynamics analysis. The addressed questions include: (1) description of music as an information source, (2) modeling of music–listener relations in terms of communication channel, (3) choice of musical features and dealing with their dependencies, (4) survey of different information measures for description of musical structure and measures of shared information between listener and the music, and (5) suggestion of new approach to characterization of listening experience in terms of different combinations of musical surface and structure expectancies.

Chapter 17

Multimodal Emotion Recognition 398
Sanaul Haq, University of Surrey, UK
Philip J.B. Jackson, University of Surrey, UK

This chapter provides a survey of research efforts in emotion recognition using different modalities: audio, visual and audio-visual combined. It also describes fifteen audio, visual and audio-visual data sets, and the types of feature that researchers have used to represent the emotional content. Several important issues, such as feature selection and reduction, emotion classification, and methods for fusing information from multiple modalities are also discussed. The chapter concludes by pointing out interesting areas in this field for future investigation.

Chapter 18

Machine Audition of Acoustics: Acoustic Channel Modeling and Room Acoustic
Parameter Estimation 424
Francis F. Li, University of Salford, UK
Paul Kendrick, University of Salford, UK
Trevor J. Cox, University of Salford, UK

Propagation of sound from a source to a receiver in an enclosure can be modeled as an acoustic transmission channel. Objective room acoustic parameters are routinely used to quantify properties of such channels in the design and assessment of acoustically critical spaces. This chapter discusses a number of new methods and algorithms for determining room acoustic parameters using machine audition of naturally occurring sound sources, i.e. speech and music. In particular, reverberation time, early decay time and speech transmission index can be estimated from received speech or music signals using statistical machine learning or maximum likelihood estimation in a semi-blind or blind fashion.

Chapter 19

Neuromorphic Speech Processing: Objectives and Methods 447

Pedro Gómez-Vilda, Universidad Politécnica de Madrid, Spain

José Manuel Ferrández-Vicente, Universidad Politécnica de Madrid, Spain

Victoria Rodellar-Biarge, Universidad Politécnica de Madrid, Spain

Roberto Fernández-Baillo, Universidad Politécnica de Madrid, Spain

Agustín Álvarez-Marquina, Universidad Politécnica de Madrid, Spain

Rafael Martínez-Olalla, Universidad Politécnica de Madrid, Spain

Victor Nieto-Lluis, Universidad Politécnica de Madrid, Spain

Luis Miguel Mazaira-Fernández, Universidad Politécnica de Madrid, Spain

Cristina Muñoz-Mulas, Universidad Politécnica de Madrid, Spain

In speech perception and recognition, many hidden phenomena are not well understood yet, including the semantic gap going from spectral time-frequency representations to the symbolic translation into phonemes and words, and the construction of morpho-syntactic and semantic structures. This chapter is intended to explore some of these facts at a simplifying level under two points of view: that of top-down analysis provided from speech perception, and the symmetric from bottom-up synthesis provided by the biological architecture of auditory pathways. It also includes an application-driven design of a neuromorphic speech processing architecture and the simulation details provided by a parallel implementation of the architecture in a supercomputer.

Compilation of References 474

About the Contributors 515

Index 526