Machine Audition: Principles, Algorithms and Systems

Edited by: Wenwu Wang, University of Surrey, UK

Preface

Outline and Subject of this Book

Machine audition is the field of the study of algorithms and systems for the automatic analysis and understanding of sound by machine. It plays an important role in many applications, such as automatic audio indexing for internet searching, robust speech recognition in un-controlled natural environment, untethered audio communication within an intelligent office scenario, and speech enhancement for hearing aids and cochlear implants, etc. It has recently attracted increasing interest within several research communities, such as signal processing, machine learning, auditory modelling, perception and cognition, psychology, pattern recognition, and artificial intelligence. However, the developments made so far are fragmented within these disciplines, lacking connections and incurring potentially overlapping research activities in this subject area. The proposed book intends to bring together the advances in recent algorithmic developments, bridge the gaps between the methodologies adopted by the various disciplines, and overlook future directions in this subject.

Objectives, Missions and the Scholarly Value

This book aims to provide algorithmic developments, theoretical frameworks and empirical and experimental research findings in the area of machine audition. It could be useful for professionals who want to improve their understanding about how to design algorithms for performing automatic analysis of audio signals, how to construct a computing system that could understand sound sources around us, and how to build advanced human-computer interactive systems. The book covers the existing and the emerging algorithms and frameworks for processing sound mixtures, the practical approaches for implementing machine audition systems, as well as the relationship between human and machine audition. It will provide professionals, academic researchers, students, consultants and practitioners with a good overview of how the sound might be understood by a machine based on algorithmic operation, and how the machine audition approaches might be useful for solving practical engineering problems in daily life.

The book is the first of its kind that describes the theoretical, algorithmic and systematic results from the area of machine audition. It intends to promote “machine audition” as a subject area that is equally attractive to the popular subject of “computer vision”. The book treats audition in the context of general audio, rather than for specific data, such as speech in some existing literature. It contains many new approaches and algorithms, most recent numerical and experimental results, which could foster a better understanding of the state of the art of the subject and

ultimately motivate novel ideas and thinking in the research communities. A unique characteristic about the book is that it brings together the fragments of the research findings in machine audition research across several disciplines, which could potentially promote cutting-edge research in this subject area.

Target Audience

The contents of this book are expected to be attractive to professionals, researchers, students and practitioners working in the fields of machine audition, audio engineering and signal processing. Researchers from the field of computer sciences, information technology and psychology will also be the audience of the book. The proposed book will be a precious reference for these audience who wish to have better understanding about the subject, to contribute to research of the subject, and to implement their new ideas and to provide technical consultancy in the field.

The potential uses of the book include library reference, upper-level course supplement, resource for instructors, reference for researchers, reference book for policy makers, reference book for businessman, studying material for undergraduate or postgraduate students, resource for practitioners, resource for consultants, etc

Organisation of the Book

This book has nineteen chapters divided into four broader areas as follows:

- Audio Scene Analysis, Recognition and Modeling
- Audio Signal Separation, Extraction and Localization
- Audio Transcription, Mining and Information Retrieval
- Audio Cognition, Modeling and Affective Computing

We briefly summarize the contents of each section and the main contributions of each chapter, based on the abstracts and details of the chapters provided by the authors. To be as much consistent with the original contributions as possible, the following summaries for each chapter are direct quotations of the descriptions provided by the authors, with some moderations.

Section I: Audio Scene Analysis, Recognition and Modeling

This section focuses on the computational principles and algorithms for audio scene analysis, recognition and modeling. It includes four chapters in several aspects of audio scene analysis, such as environmental audio recognition, computational auditory scene analysis, cocktail party problem, and the functional requirements of auditory systems for uncontrolled natural environments. The key issue that machine audition attempts to address is on the automatic analysis (understanding by computers) of the audio scenes using algorithm-based operations. From this aspect, progresses in this area are likely to have significant impact on this subject.

Chapter I, “Unstructured Environmental Audio: Representation, Classification and Modeling” by Chu, Narayanan, and Jay Kuo, discusses 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 II, "Modeling Grouping Cues for Auditory Scene Analysis using a Spectral Clustering Formulation" by Martins, Lagrange, and Tzanetakis, proposes to formulate the integration problem in scene analysis as clustering based on similarities between time-frequency atoms and this provides an expressive yet disciplined approach to building sound source characterization and separation systems and evaluating their performance. The authors describe the main components of the architecture, its advantages, implementation details, and related issues.

Chapter III: "Cocktail Party Problem: Source Separation Issues and Computational Methods" by Jan and Wang, provides a review on recent progresses for cocktail party problem 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. The chapter also discusses the applications of cocktail party processing and its potential research directions for the future.

Chapter IV: "Audition: from Sound to Sounds" by Andringa, 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. The discussions are based on the demand to function in uncontrolled listening environments and their implications for machine audition.

Section II: Audio Signal Separation, Extraction and Localization

Source separation, extraction and localization play a central role in automatic auditory scene analysis. This section collects six recent contributions in this area, such as a multimodal approach for moving source separation, source separation based on probabilistic modeling or sparse representation, tensor factorization for source separation, multichannel source separation, and sound source localization based on intensity vector directions. Source separation problems have been studied extensively in the past two decades. It has widespread applications in, for example, robust speech recognition, teleconferencing, human-computer interaction and so on.

Chapter V: "A Multimodal Solution to Blind Source Separation of Moving Sources" by Naqvi, Zhang, Yu, and Chambers, proposes 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 VI: “Sound Source Localization: Conventional Methods and Intensity Vector Direction Exploitation” by Günel and Hacıhabiboğlu, presents an overview of the conventional array processing methods for sound source localization and a discussion of the factors that affect their performance. The chapter then 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 VII: “Probabilistic Modeling Paradigms for Audio Source Separation” by Vincent, Jafari, Abdallah, Plumbley, and Davies, focuses on the audio source separation methods by inferring the parameters of probabilistic sound models. The authors 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 VIII: “Tensor Factorization with Application to Convolutive Blind Source Separation of Speech” by Sanei and Makkiabadi, introduces the Tensor factorization (TF) technique for the separation of sound particularly speech sources from their corresponding convolutive 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 IX: “Multi-Channel Source Separation: Overview and Comparison of Mask-based and Linear Separation Algorithms” by Madhu and Gückel, 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 X: “Audio Source Separation using Sparse Representations” by Nesbit, Jafari, Vincent, and Plumbley, addresses the problem of audio source separation based on the sparse component analysis framework. The overriding aim 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 III: Audio Transcription, Mining and Information Retrieval

This section includes five contributions on different aspects of audio transcription, mining and information retrieval, such as music decomposition based on machine learning techniques, music onset detection, music segmentation, and automatic tagging of audio. All these are important topics in machine audition and they attract increasing research interests recently. Research outputs in this area are likely to have strong impact in audio coding, compression, and indexing.

Chapter XI: “Itakura-Saito Nonnegative Factorizations of the Power Spectrogram for Music Signal Decomposition” by Févotte, 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 XII: “Music Onset Detection” by Zhou and Reiss, provides 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 XIII: “On the Inherent Segment Length in Music” by Jensen, 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 XIV: “Automatic Tagging of Audio: The State-of-the-Art” by Bertin-Mahieux, Eck, and Mandel, provides a review of the state-of-the-art methods for addressing automatic tagging of audio. 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. The chapter is devoted as an effort to better understand the task and also to help new researchers bring their insights to bear on this problem. It is divided in the following sections: goal, framework, audio representation, labeled data, classification, evaluation, and future directions.

Chapter XV: “Instantaneous versus Convolutive Non-negative Matrix Factorization: Models, Algorithms and Applications to Audio Pattern Separation” by Wang, presents an overview of the models and algorithms for instantaneous and convolutive non-negative matrix factorization (NMF), with a focus on the convolutive NMF algorithms and their performance. The chapter

discusses the limitations of the instantaneous model and the advantages of the convolutive model in addressing such limitations. The chapter also provides application examples of both models and algorithms in audio pattern separation and onset detection. A theoretical analysis of the convolutive NMF algorithms is also included.

Section IV: Audio Cognition, Modeling and Affective Computing

This section is a collection of four contributions in the area of audio cognition, modeling and affective computing, such as the modeling methods for music cognition, in particular, music anticipation, emotion recognition from audio, video, or audio-visual data, acoustic channel modeling and parameter estimation from speech or music signals recorded in a room, and using semantic and symbolic information from speech perception for the design of speech processing systems. The topics in this section bring together knowledge from several subjects including signal processing, psychology, computer science, and statistics. Many of these topics are emerging areas in the field.

Chapter XVI: "Musical Information Dynamics as Models of Auditory Anticipation" by Dubnov, 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 XVII: "Multimodal Emotion Recognition" by Haq and Jackson, 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 XVIII: "Machine Audition of Acoustics - Acoustic Channel Modeling and Room Acoustic Parameter Estimation" by Li, Kendrick, and Cox, 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 XIX: "Neuromorphic Speech Processing: Objectives and Methods" by Gómez-Vilda, Ferrández-Vicente, Rodellar-Biarge, Fernández-Baíllo, Álvarez-Marquina, Martínez-Olalla, Nieto-Lluis, Mazaira-Fernández, and Muñoz-Mulas, is intended to explore some of the hidden phenomena in speech perception and recognition, 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, for the design of a neuromorphic speech processing architecture. These facts are considered in 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.

Acknowledgements

I wish to thank all the people who were involved in different phases during the preparation of this book, including all the authors who have submitted their important contributions to this book and also provided assistance in reviewing the submitted chapters, all the advisory editorial board members who have helped in the review of the submissions and given useful suggestions on the structure of the book, and the executive editors at IGI, in particular, Mr Joel A. Gamon, who has provided many helpful suggestions and answered each question that I raised when preparing this book. My thanks also go to Clive Cheong Took and Hector Perez-Meana for their assistance in the review process. Finally, I would like to take this opportunity to thank my wife and lovely daughter for their consistent support and love. Without the unselfish assistance of all these people, the successful publication of the book would not have been possible.

Wenwu Wang
University of Surrey