



INTRODUCTION TO

AUDIO ANALYSIS

A MATLAB APPROACH

THEODOROS GIANNAKOPOULOS | AGGELOS PIKRAKIS

MATLAB[®]
examples



Introduction to **AUDIO ANALYSIS:** A MATLAB Approach

THEODOROS GIANNAKOPOULOS

AGGELOS PIKRAKIS



Amsterdam • Boston • Heidelberg • London
New York • Oxford • Paris • San Diego
San Francisco • Singapore • Sydney • Tokyo

Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

First edition 2014

Copyright © 2014 Elsevier Ltd. All rights reserved.

MATLAB® is a registered trademarks of The MathWorks, Inc.

For MATLAB and Simulink product information, please contact:

The MathWorks, Inc.

3 Apple Hill Drive

Natick, MA, 01760-2098 USA

Tel: 508-647-7000

Fax: 508-647-7001

E-mail: info@mathworks.com

Web: mathworks.com

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email permissions@elsevier.com. Alternatively you can submit your request online by visiting the Elsevier website at <http://elsevier.com/locate/permissions>, and selecting *Obtaining permission to use Elsevier material*.

Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

ISBN: 978-0-08-099388-1

For information on all Academic Press publications
visit our web site at books.elsevier.com

Printed and bound in United States of America

14 15 16 17 18 10 9 8 7 6 5 4 3 2 1



Working together
to grow libraries in
developing countries

www.elsevier.com • www.bookaid.org

Introduction to
AUDIO ANALYSIS

PREFACE

This book attempts to provide a gentle introduction to the field of audio analysis using the MATLAB programming environment as the vehicle of presentation. Audio analysis is a multidisciplinary field, which requires the reader to be familiar with concepts from diverse research disciplines, including digital signal processing and machine learning. As a result, it is a great challenge to write a book that can provide sufficient coverage of the important concepts in the field of audio analysis and, at the same time, be accessible to readers who do not necessarily possess the required scientific background.

Our main goal has been to provide a standalone introduction, involving a balanced presentation of theoretical descriptions and reproducible MATLAB examples. Our philosophy is that readers with diverse scientific backgrounds can gain an understanding of the field of audio analysis, if they are provided with basic theory, in conjunction with reproducible experiments that can help them deal with the theory from a more practical perspective. In addition, this type of approach allows the reader to acquire certain technical skills that are useful in the context of developing real-world audio analysis applications. To this end, we also provide an accompanying software library which can be downloaded from the companion site and includes the MATLAB functions and related data files that have been used throughout the text.

We believe that this book is suitable for students, researchers, and professionals alike, who need to develop practical skills, along with a basic understanding of the field. The book does not assume previous knowledge of digital signal processing and machine learning concepts, as it provides introductory material for the necessary topics for both disciplines. We expect that, after reading this book, the reader will feel comfortable with various key processing stages of the audio analysis chain, including audio content creation, representation, feature extraction, classification, segmentation, sequence alignment and temporal modeling. Furthermore, we believe that the study of the presented case studies will provide further insight into the development of real-world applications.

This book is the product of several years of teaching and research and reflects our teaching philosophy, which has been shaped via our interaction with our students and colleagues, and to whom we are both grateful. We

hope that the will prove useful to all readers who are making their first steps in the field of audio analysis. Although we have made an effort to eliminate errors during the writing stage, we encourage the reader to contact us with any comments and suggestions for improvement, in either the text or the accompanying software library.

Theodoros Giannakopoulos and Aggelos Pikrakis
Athens, 2013

For access to the software library and other supporting materials, please visit the companion website at: <http://booksite.elsevier.com/9780080993881>

ACKNOWLEDGMENTS

This book has improved thanks to the support of a number of colleagues, students, and friends, who have provided generous feedback and constructive comments, during the writing process. Above all, T. Giannakopoulos would like to thank his wife, Maria, and his daughter, Eleni, for always being cheerful and supportive. A. Pikrakis would like to thank his family for their patience and generous support and dedicates this book to all the teachers who have shaped his life.

CONTENTS

<i>Preface</i>	<i>vii</i>
<i>Acknowledgments</i>	<i>ix</i>
<i>List of Tables</i>	<i>xi</i>
<i>List of Figures</i>	<i>xiii</i>

PART 1 – BASIC CONCEPTS, REPRESENTATIONS AND FEATURE EXTRACTION

1. Introduction	3
1.1 The MATLAB Audio Analysis Library	5
1.2 Outline of Chapters	5
1.3 A Note on Exercises	7
2. Getting Familiar with Audio Signals	9
2.1 Sampling	10
2.2 Playback	11
2.3 Mono and Stereo Audio Signals	13
2.4 Reading and Writing Audio Files	14
2.5 Reading Audio Files in Blocks	19
2.6 Recording Audio Data	20
2.7 Short-Term Audio Processing	24
2.8 Exercises	28
3. Signal Transforms and Filtering Essentials	33
3.1 The Discrete Fourier Transform	33
3.2 The Short-Time Fourier Transform	39
3.3 Aliasing in More Detail	43
3.4 The Discrete Cosine Transform	44
3.5 The Discrete-Time Wavelet Transform	46
3.6 Digital Filtering Essentials	49
3.7 Digital Filters in MATLAB	52
3.8 Exercises	56
4. Audio Features	59
4.1 Short-Term and Mid-Term Processing	60
4.2 Class Definitions	68

4.3 Time-Domain Audio Features	70
4.4 Frequency-Domain Audio Features	78
4.5 Periodicity Estimation and Harmonic Ratio	92
4.6 Exercises	97

PART 2 – AUDIO CONTENT CHARACTERIZATION

5. Audio Classification	107
5.1 Classification Fundamentals	109
5.2 Popular Classifiers	115
5.3 Implementation-Related Issues	129
5.4 Evaluation	134
5.5 Case Studies	142
5.6 Exercises	148
6. Audio Segmentation	153
6.1 Segmentation with Embedded Classification	154
6.2 Segmentation without Classification	169
6.3 Exercises	180
7. Audio Alignment and Temporal Modeling	185
7.1 Audio Sequence Alignment	186
7.2 Hidden Markov Modeling	193
7.3 The Viterbi Algorithm	196
7.4 The Baum-Welch Algorithm	198
7.5 HMM Training	199
7.6 Exercises	203

PART 3 – OTHER ISSUES

8. Music Information Retrieval	211
8.1 Music Thumbnailing	214
8.2 Music Meter and Tempo Induction	216
8.3 Music Content Visualization	219
8.4 Exercises	228
<i>Appendix A</i>	233
<i>Appendix B</i>	241
<i>Appendix C</i>	247
<i>Bibliography</i>	251
<i>Index</i>	259

LIST OF TABLES

Table 1.1	Difficulty Levels of the Exercises	7
Table 2.1	Execution Times for Different Loading Techniques	20
Table 2.2	Sound Recording Using the Data Acquisition Toolbox	20
Table 4.1	Class Descriptions for the Multi-Class Task of Movie Segments	69
Table 5.1	Classification Tasks and Files	131
Table 5.2	Row-Wise Normalized Confusion Matrix for the 8-Class Audio Segment Classification Task	144
Table 5.3	Row-Wise Normalized Confusion Matrix for the Speech vs Music Binary Classification Task	144
Table 5.4	Row-Wise Normalized Confusion Matrix for the 3-Class Musical Genre Classification Task	146
Table 5.5	Row-Wise Normalized Confusion Matrix for the Speech vs Non-Speech Classification Task	147
Table A.1	List of All Functions Included in the MATLAB Audio Analysis Library Provided with the Book	233
Table A.2	List of Data Files that are Available in the Library that Accompanies the Book	239
Table B.1	MATLAB Libraries—Audio and Speech	242
Table B.2	MATLAB Libraries—Pattern Recognition and Machine Learning	242
Table B.3	A List of Python Packages and Libraries that can be Used for Audio Analysis and Pattern Recognition Applications	244
Table B.4	Representative Audio Analysis and Pattern Recognition Libraries and Packages Written in C++	245
Table C.1	A Short List of Available Datasets for Selected Audio Analysis Tasks	247

LIST OF FIGURES

Figure 2.1	A synthetic audio signal.	12
Figure 2.2	A STEREO audio signal.	14
Figure 2.3	Short-term processing of an audio signal.	26
Figure 3.1	Plots of the magnitude of the spectrum of a signal consisting of three frequencies at 200, 500, and 1200 Hz.	38
Figure 3.2	A synthetic signal consisting of three frequencies is corrupted by additive noise.	40
Figure 3.3	The spectrogram of a speech signal.	41
Figure 3.4	Spectrograms of a synthetic, frequency-modulated signal for three short-term frame lengths.	42
Figure 3.5	Spectrum representations of (a) an analog signal, (b) a sampled version when the sampling frequency exceeds the Nyquist rate, and (c) a sampled version with insufficient sampling frequency. In the last case, the shifted versions of the analog spectrum are overlapping, hence the aliasing effect.	43
Figure 3.6	Spectral representations of the same three-tone (200, 500 and 3000 HZ) signal for two different sampling frequencies (8 kHz and 4 kHz).	44
Figure 3.7	Frequency response of a pre-emphasis filter for $a = -0.95$.	51
Figure 3.8	An example of the application of a lowpass filter on a synthetic signal consisting of three tones.	53
Figure 3.9	Example of a simple speech denoising technique applied on a segment of the <code>diarizationExample.wav</code> file, found in the data folder of the library of the book.	55
Figure 4.1	Mid-term feature extraction: each mid-term segment is short-term processed and statistics are computed based on the extracted feature sequence.	63
Figure 4.2	Plotting the results of <code>featureExtractionFile()</code> , using <code>plotFeaturesFile()</code> , for the six feature statistics drawn from the 6th adopted audio feature.	68
Figure 4.3	Histograms of the standard deviation by mean ratio ($\frac{\sigma^2}{\mu}$) of the short-term energy for two classes: music and speech.	72
Figure 4.4	Example of a speech segment and the respective sequence of ZCR values.	74
Figure 4.5	Histograms of the standard deviation of the ZCR for music and speech classes.	75

Figure 4.6	Sequence of entropy values for an audio signal that contains the sounds of three gunshots. Low values appear at the onset of each gunshot.	77
Figure 4.7	Histograms of the minimum value of the entropy of energy for audio segments from the genres of jazz, classical and electronic music.	78
Figure 4.8	Histograms of the maximum value of the sequence of values of the spectral centroid, for audio segments from three classes of environmental sounds: others1, others2, and others3.	81
Figure 4.9	Histograms of the maximum value of the sequences of the spectral spread feature, for audio segments from three music genres: classical, jazz, and electronic.	82
Figure 4.10	Histograms of the standard deviation of sequences of the spectral entropy feature, for audio segments from three classes: music, speech, and others1 (low-level environmental sounds).	83
Figure 4.11	Histograms of the mean value of the sequence of spectral flux values, for audio segments from two classes: music and speech.	85
Figure 4.12	Example of the spectral rolloff sequence of an audio signal that consists of four music excerpts. The first 5 s stem from a classical music track.	87
Figure 4.13	Frequency warping function for the computation of the MFCCs.	88
Figure 4.14	Histograms of the standard deviation of the 2nd MFCC for the classes of music and speech.	91
Figure 4.15	Chromagrams for a music and a speech segment.	92
Figure 4.16	Autocorrelation, normalized autocorrelation, and detected peak for a periodic signal.	94
Figure 4.17	Histograms of the maximum value of sequences of values of the harmonic ratio for two classes of sounds (speech and others1).	96
Figure 5.1	Generic diagram of the classifier training stage.	112
Figure 5.2	Diagram of the classification process.	113
Figure 5.3	Linearly separable classes in a two-dimensional feature space.	118
Figure 5.4	Decision tree for a classification task with 3-classes ($\omega_1, \omega_2, \omega_3$) and three features (x_1, x_2, x_3).	122
Figure 5.5	Decision tree for a 4-class task with Gaussian feature distributions in the two-dimensional feature space.	123
Figure 5.6	Decision tree for a musical genre classification task with two feature statistics (minimum value of the entropy of energy and mean value of the spectral flux).	124

Figure 5.7	SVM training for different values of the C parameter.	128
Figure 5.8	Classification accuracy on the training and testing dataset for different values of C .	129
Figure 5.9	Implementation of the k -NN classification procedure.	132
Figure 5.10	Binary classification task with Gaussian feature distributions and two different decision thresholds.	137
Figure 5.11	Performance of the k -NN classifier on an 8-class task, for different values of the k parameter and for two validation methods (repeated hold-out and leave-one-out).	143
Figure 5.12	Estimated performance for the 3-class musical genre classification task, for different values of the k parameter and for two evaluation methods (repeated hold-out and leave-one-out).	145
Figure 6.1	Post-segmentation stage: the output of the first stage can be (a) a sequence of hard classification decisions, $C_i, i = 1, \dots, N_{mt}$; or (b) a sequence of sets of posterior probability estimates, $P_i(j), i = 1, \dots, N_{mt}, j = 1, \dots, N_c$.	155
Figure 6.2	Fixed-window segmentation.	156
Figure 6.3	Fixed-window segmentation: naive merging vs Viterbi-based smoothing.	159
Figure 6.4	Example of the silence detection approach implemented in <code>silenceDetectorUtterance()</code> .	162
Figure 6.5	Speech-silence segmenter applied on a short-duration signal.	164
Figure 6.6	Fixed-window segmentation with an embedded 4-class classifier (silence, male speech, female speech, and music).	166
Figure 6.7	A sequence of segments in the dynamic programming grid.	168
Figure 6.8	Top: Signal change detection results from a TV program. Bottom: Ground truth.	171
Figure 6.9	A clustering example in the two-dimensional feature space.	173
Figure 6.10	Silhouette example: the average Silhouette measure is maximized when the number of clusters is 4.	176
Figure 6.11	Block diagram of the speaker diarization method implemented in <code>speakerDiarization()</code> .	178
Figure 6.12	Visualization of the speaker diarization results obtained by the <code>speakerDiarization()</code> function (visualization is obtained by calling the <code>segmentationPlotResults()</code> function).	179

Figure 8.1	Self-similarity matrix for the track 'True Faith' by the band New Order.	215
Figure 8.2	Approximation of the second derivative, D_2 of sequence B .	218
Figure 8.3	Visualization results for the three linear dimensionality reduction approaches, applied on the <code>musicSmallData.mat</code> dataset.	225
Figure 8.4	<code>gridtop</code> topology (5×5).	226
Figure 8.5	Visualization of selected nodes of the SOM of the data in the <code>musicLargeData.mat</code> dataset.	227