

An Introduction to Audio Content Analysis

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

IEEE Press Editorial Board
John B. Anderson, *Editor in Chief*

R. Abhari	G. W. Arnold	F. Canavero
D. Goldgof	B-M. Haemmerli	D. Jacobson
M. Lanzerotti	O. P. Malik	S. Nahavandi
T. Samad	G. Zobrist	

Kenneth Moore, *Director of IEEE Book and Information Services (BIS)*

Technical Reviewers
Roger B. Dannenberg, Carnegie Mellon University

An Introduction to Audio Content Analysis

Applications in Signal Processing and Music Informatics

Alexander Lerch

zplane.development, Berlin



IEEE PRESS



A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2012 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

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, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Lerch, Alexander.

Audio content analysis : an introduction / Alexander Lerch.

p. cm.

Includes bibliographical references and index.

ISBN 978-1-118-26682-3 (hardback)

1. Computer sound processing. 2. Computational auditory scene analysis. 3. Content analysis (Communication)—Data processing. I. Title.

TK7881.4.L485 2012

006.4'5—dc23

2012008107

Printed in the United States of America.

10 9 8 7 6 5 4 3 2

CONTENTS IN BRIEF

1	Introduction	1
2	Fundamentals	7
3	Instantaneous Features	31
4	Intensity	71
5	Tonal Analysis	79
6	Temporal Analysis	119
7	Alignment	139
8	Musical Genre, Similarity, and Mood	151
9	Audio Fingerprinting	163
10	Music Performance Analysis	169

CONTENTS

Preface	xiii
Acronyms	xv
List of Symbols	xix
1 Introduction	1
1.1 Audio Content	3
1.2 A Generalized Audio Content Analysis System	4
2 Fundamentals	7
2.1 Audio Signals	7
2.1.1 Periodic Signals	7
2.1.2 Random Signals	9
2.1.3 Sampling and Quantization	9
2.1.4 Statistical Signal Description	13
2.2 Signal Processing	14
2.2.1 Convolution	14
2.2.2 Block-Based Processing	18
2.2.3 Fourier Transform	20
2.2.4 Constant Q Transform	23
2.2.5 Auditory Filterbanks	24
2.2.6 Correlation Function	24
	vii

2.2.7	Linear Prediction	28
3	Instantaneous Features	31
3.1	Audio Pre-Processing	33
3.1.1	Down-Mixing	33
3.1.2	DC Removal	33
3.1.3	Normalization	34
3.1.4	Down-Sampling	34
3.1.5	Other Pre-Processing Options	35
3.2	Statistical Properties	35
3.2.1	Arithmetic Mean	36
3.2.2	Geometric Mean	36
3.2.3	Harmonic Mean	36
3.2.4	Generalized Mean	36
3.2.5	Centroid	37
3.2.6	Variance and Standard Deviation	37
3.2.7	Skewness	38
3.2.8	Kurtosis	39
3.2.9	Generalized Central Moments	40
3.2.10	Quantiles and Quantile Ranges	40
3.3	Spectral Shape	41
3.3.1	Spectral Rolloff	42
3.3.2	Spectral Flux	44
3.3.3	Spectral Centroid	45
3.3.4	Spectral Spread	47
3.3.5	Spectral Decrease	48
3.3.6	Spectral Slope	49
3.3.7	Mel Frequency Cepstral Coefficients	51
3.4	Signal Properties	54
3.4.1	Tonalness	54
3.4.2	Autocorrelation Coefficients	61
3.4.3	Zero Crossing Rate	62
3.5	Feature Post-Processing	63
3.5.1	Derived Features	64
3.5.2	Normalization and Mapping	65
3.5.3	Subfeatures	66
3.5.4	Feature Dimensionality Reduction	66
4	Intensity	71
4.1	Human Perception of Intensity and Loudness	71
4.2	Representation of Dynamics in Music	73
4.3	Features	73

4.3.1	Root Mean Square	73
4.3.2	Peak Envelope	76
4.3.3	Psycho-Acoustic Loudness Features	77
5	Tonal Analysis	79
5.1	Human Perception of Pitch	79
5.1.1	Pitch Scales	79
5.1.2	Chroma Perception	81
5.2	Representation of Pitch in Music	82
5.2.1	Pitch Classes and Names	82
5.2.2	Intervals	83
5.2.3	Root Note, Mode, and Key	83
5.2.4	Chords and Harmony	86
5.2.5	The Frequency of Musical Pitch	88
5.3	Fundamental Frequency Detection	91
5.3.1	Detection Accuracy	92
5.3.2	Pre-Processing	94
5.3.3	Monophonic Input Signals	97
5.3.4	Polyphonic Input Signals	103
5.4	Tuning Frequency Estimation	106
5.5	Key Detection	108
5.5.1	Pitch Chroma	108
5.5.2	Key Recognition	112
5.6	Chord Recognition	116
6	Temporal Analysis	119
6.1	Human Perception of Temporal Events	119
6.1.1	Onsets	119
6.1.2	Tempo and Meter	122
6.1.3	Rhythm	122
6.1.4	Timing	123
6.2	Representation of Temporal Events in Music	123
6.2.1	Tempo and Time Signature	123
6.2.2	Note Value	124
6.3	Onset Detection	124
6.3.1	Novelty Function	125
6.3.2	Peak Picking	127
6.3.3	Evaluation	128
6.4	Beat Histogram	133
6.4.1	Beat Histogram Features	134
6.5	Detection of Tempo and Beat Phase	135
6.6	Detection of Meter and Downbeat	136

7	Alignment	139
7.1	Dynamic Time Warping	139
7.1.1	Example	143
7.1.2	Common Variants	144
7.1.3	Optimizations	145
7.2	Audio-to-Audio Alignment	146
7.2.1	Ground Truth Data for Evaluation	147
7.3	Audio-to-Score Alignment	148
7.3.1	Real-Time Systems	148
7.3.2	Non-Real-Time Systems	149
8	Musical Genre, Similarity, and Mood	151
8.1	Musical Genre Classification	151
8.1.1	Musical Genre	152
8.1.2	Feature Extraction	154
8.1.3	Classification	155
8.2	Related Research Fields	156
8.2.1	Music Similarity Detection	156
8.2.2	Mood Classification	158
8.2.3	Instrument Recognition	161
9	Audio Fingerprinting	163
9.1	Fingerprint Extraction	164
9.2	Fingerprint Matching	165
9.3	Fingerprinting System: Example	166
10	Music Performance Analysis	169
10.1	Musical Communication	169
10.1.1	Score	169
10.1.2	Music Performance	170
10.1.3	Production	172
10.1.4	Recipient	172
10.2	Music Performance Analysis	172
10.2.1	Analysis Data	173
10.2.2	Research Results	177
A	Convolution Properties	181
A.1	Identity	181
A.2	Commutativity	181
A.3	Associativity	182
A.4	Distributivity	183
A.5	Circularity	183
B	Fourier Transform	185

B.1	Properties of the Fourier Transformation	186
B.1.1	Inverse Fourier Transform	186
B.1.2	Superposition	186
B.1.3	Convolution and Multiplication	186
B.1.4	Parseval's Theorem	187
B.1.5	Time and Frequency Shift	188
B.1.6	Symmetry	188
B.1.7	Time and Frequency Scaling	189
B.1.8	Derivatives	190
B.2	Spectrum of Example Time Domain Signals	190
B.2.1	Delta Function	190
B.2.2	Constant	191
B.2.3	Cosine	191
B.2.4	Rectangular Window	191
B.2.5	Delta Pulse	191
B.3	Transformation of Sampled Time Signals	192
B.4	Short Time Fourier Transform of Continuous Signals	192
B.4.1	Window Functions	193
B.5	Discrete Fourier Transform	195
B.5.1	Window Functions	196
B.5.2	Fast Fourier Transform	197
C	Principal Component Analysis	199
C.1	Computation of the Transformation Matrix	200
C.2	Interpretation of the Transformation Matrix	200
D	Software for Audio Analysis	201
D.1	Software Frameworks and Applications	202
D.1.1	Marsyas	202
D.1.2	CLAM	202
D.1.3	jMIR	203
D.1.4	CoMIRVA	203
D.1.5	Sonic Visualiser	203
D.2	Software Libraries and Toolboxes	204
D.2.1	Feature Extraction	204
D.2.2	Plugin Interfaces	205
D.2.3	Other Software	206
	References	207
	Index	243

PREFACE

The growing amount of audio and music data on the Internet and in user databases leads to an increasing need for intelligent browsing, retrieving, and processing of this data with automated methods. *Audio content analysis*, a subfield of the research field *music information retrieval*, aims at extracting (musical and perceptual) properties directly from the audio signal to support these tasks. Knowledge of these properties allows us to improve the interaction of humans or machines with digital audio signals. It enables new ways of assessing, processing, and visualizing music.

Although analysis of audio signals covers other research areas such as automatic speech recognition, we will restrict ourselves to the analysis of music signals in the context of this book.

When preparing classes on audio content analysis with a focus on music recordings it became quickly clear that — although there is a vast and growing amount of research literature available — there exists no introductory literature. This observation led to writing this book in the hope it might assist students, engineers, and developers who have basic knowledge of digital signal processing. The focus lies on the signal processing part of audio content analysis, but wherever it may improve the understanding of either algorithmic design choices or implementation details some basic characteristics of human perception, music theory, and notation as well as machine learning will be summarized.

Chapter 2 starts by introducing some definitions and offers a short reiteration of the most important tools of digital signal processing for the analysis of audio signals. The following chapters encompass the basic four technical content categories timbre, level, pitch, and rhythm. A fifth category is reserved for purely technical and statistical signal descriptions. Chapter 3 introduces low-level or short-term features that are widely used in systems for signal analysis. A large part of the chapter deals with timbre represen-

tations of a signal, accompanied by the introduction of statistical features. The chapter concludes with a summary of approaches to feature selection and post-processing. Chapter 4 focuses on intensity-related features. It covers envelope features and simple models of human loudness perception. The extraction of pitch-related information such as the detection of fundamental frequency, harmony, key, etc. is described in Chap. 5. Chapter 6 focuses on the temporal and rhythmic aspects of the audio signal. It explains the segmentation of audio signals into musical events and covers higher level information such as the detection of tempo and meter. The remaining chapters deal with analysis systems using combinations of timbre, loudness, onset, and pitch features to derive higher level information. Chapter 7 describes the automatic synchronization of two similar audio sequences or an audio and a score sequence. Musical genre classification, one of the most prominent research fields of audio content analysis, is explained in Chap. 8. Chapter 9 is about audio fingerprinting which is probably the commercially most successful application in audio content analysis. The concluding chapter, targeting classical music, covers the analysis of music performance. It is not a core field in audio content analysis but emphasizes the differentiation between performance aspects and musical aspects of recordings and elaborates on the manual and automated analysis methods used for musicological music performance analysis. The appendices provide details and derivations of some of the most important signal processing tools as well as a short survey on available software solutions for audio content analysis.

Downloadable MATLAB files are available at: <http://www.audiocontentanalysis.org>.

A. LERCH

Berlin

January, 2012

ACRONYMS

ACA	Audio Content Analysis
ACF	Autocorrelation Function
ADPCM	Adaptive Differential Pulse Code Modulation
AMDF	Average Magnitude Difference Function
ANN	Artificial Neural Network
AOT	Acoustic Onset Time
API	Application Programmer's Interface
BPM	Beats per Minute
CAMEL	Content-based Audio and Music Extraction Library
CASA	Computational Auditory Scene Analysis
CCF	Cross Correlation Function
CCIR	Comité Consultatif International des Radiocommunications
CD	Compact Disc
CiCF	Circular Correlation Function
CLAM	C++ Framework for Audio and Music
COG	Center of Gravity
CQT	Constant Q Transform

DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DP	Dynamic Programming
DTW	Dynamic Time Warping
EBU	European Broadcasting Union
ERB	Equivalent Rectangular Bandwidth
FEAPI	Feature Extraction Application Programmer's Interface
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FN	False Negative
FP	False Positive
FT	Fourier Transform
FWR	Full-Wave Rectification
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HPS	Harmonic Product Spectrum
HSS	Harmonic Sum Spectrum
HTK	HMM Toolkit
HWR	Half-Wave Rectification
IBI	Inter-Beat Interval
ICA	Independent Component Analysis
IDFT	Inverse Discrete Fourier Transform
IFT	Inverse Fourier Transform
IIR	Infinite Impulse Response
IO	Input/Output
IOI	Inter-Onset Interval
ISMIR	International Society for Music Information Retrieval
ITU	International Telecommunication Union
JNDL	Just Noticeable Difference in Level
KNN	K-Nearest Neighbor

LDA	Linear Discriminant Analysis
MA	Moving Average
MFCC	Mel Frequency Cepstral Coefficient
MIDI	Musical Instrument Digital Interface
MIR	Music Information Retrieval
MIREX	Music Information Retrieval Evaluation eXchange
MPA	Music Performance Analysis
MPEG	Motion Picture Experts Group
NOT	Note Onset Time
PAT	Perceptual Attack Time
PCA	Principal Component Analysis
PDF	Probability Density Function
POT	Perceptual Onset Time
PPM	Peak Program Meter
PSD	Peak Structure Distance
RBF	Radial Basis Function
RFD	Relative Frequency Distribution
RLB	Revised Low Frequency B Curve
RMS	Root Mean Square
ROC	Receiver Operating Curve
SIMD	Single Instruction Multiple Data
SNR	Signal-to-Noise Ratio
SOM	Self-Organizing Map
STFT	Short Time Fourier Transform
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
WEKA	Waikato Environment for Knowledge Analysis
YAAFE	Yet Another Audio Feature Extractor

LIST OF SYMBOLS

A	Amplitude
a	Filter Coefficient (recursive)
B	Number of Beats
b	Filter Coefficient (transversal)
β	Exponent
C	Number of (Audio) Channels
$\chi(\cdot)$	Center Clipping Function
C_{AB}	Cost Matrix for the Distance Matrix between Sequences A and B
\mathfrak{C}_{AB}	Overall Cost of a Path through the Cost Matrix
$c_x(\cdot)$	Cepstrum of the Signal x
D_{AB}	Distance Matrix between Sequences A and B
d	Distance Measure
Δ_Q	Quantization Step Size
e_P	Prediction Error
e_Q	Quantization Error
$\epsilon(f)$	Equivalent Rectangular Bandwidth

η	(Correlation) Lag
e_{Tfp}	(Spectral) Prediction Error
F	F -Measure
f	Frequency in Hz
f_0	Fundamental Frequency in Hz
f_s	Sample Rate
f_{A4}	Tuning Frequency in Hz
\mathcal{F}	Number of Features
f_l	Instantaneous Frequency in Hz
$\mathfrak{F}(\cdot)$	(Discrete) Fourier Transform
G	Threshold
$\mathbf{\Gamma}$	Chord Transformation Matrix
$\gamma_{x,\mathcal{O}}$	Central Moment of Order \mathcal{O} of Signal x
$H(\cdot)$	Transfer Function
$h(\cdot)$	Impulse Response
\mathcal{H}	Hop Size
i	Sample Index
\mathcal{J}	Impulse Response Length
j	Integer (Loop) Variable
\mathcal{K}	Block Size
k	Frequency Bin Index
κ	Percentage
$\Lambda(k, n)$	Tonalness Spectrum
λ	Weighting Factor
\mathcal{M}	Number of (Quantization) Steps
m	Key Index
$\mathbf{m}(f)$	Pitch (Mel)
μ_x	Arithmetic Mean of Signal x
\mathcal{N}	Number of Observations or Blocks
n	Block Index

\mathcal{O}	Order (e.g., Filter Order)
o_r	Block Overlap Ratio
ω	Angular Velocity ($\omega = 2\pi f$) in radians per second
O	Number of Onsets
P	Precision
p	Alignment Path
Φ_X	Phase Spectrum of the Signal x
$\varphi(\cdot)$	Gaussian Function
p	(MIDI) Pitch
ν	Pitch Chroma Vector/Key Profile
P_x	Power of the Signal x
$p_x(x)$	Probability Density Function of the Signal x
$\psi(\cdot)$	Chord Probability Vector
\mathcal{Q}	Quality Factor (Mid-Frequency divided by Bandwidth)
q	Evaluation Metric
$Q_x(\cdot)$	Quantile Boundary
R	Recall
$r_{xy}(\cdot)$	Correlation Function between the Signals x and y
r	Radius
\mathbf{R}	Covariance Matrix
σ_x	Standard Deviation of Signal x
σ_x^2	Variance of Signal x
SNR	Signal-to-Noise Ratio
T	Time Period in s
t	Time in s
T_0	Time Period of the Fundamental Frequency in s
\mathcal{T}	Number of (Chord) Templates
\mathfrak{T}	Tempo in BPM
\mathbf{T}	(PCA) Transformation Matrix
T_S	Sample Period in s
\mathbf{V}	Feature Matrix with dimensions $\mathcal{F} \times \mathcal{N}$
v_{ACF}^η	η th Autocorrelation Coefficient

v_C	Centroid
\mathcal{V}	Feature Set
v_K	Kurtosis
v_{MFCC}^j	j th MFCC
v_{Peak}	Peak Envelope
v_{PPM}	Peak Program Meter
v_{RMS}	RMS
v_{SC}	Spectral Centroid
v_{SD}	Spectral Decrease
v_{SF}	Spectral Flux
v_{SK}	Spectral Kurtosis
v_{Sk}	Skewness
v_{SR}	Spectral Rolloff
v_{SS}	Spectral Spread
v_{SSk}	Spectral Skewness
v_{SSI}	Spectral Slope
v_{Ta}	ACF Maximum
v_{Tf}	Spectral Flatness
v_{Tfp}	Spectral Predictivity
v_{Tp}	Predictivity Ratio
v_{Tpr}	Tonal Power Ratio
v_{Tsc}	Spectral Crest Factor
v_{ZC}	Zero Crossing Rate
w	Word Length in Bit
w_{AB}	Window Function with Alternative Blackman Shape
w_{B}	Window Function with Blackman Shape
w_{BH}	Window Function with Blackman-Harris Shape
w_{C}	Window Function with Cosine Shape
w_{H}	Window Function with von-Hann Shape
w_{Hm}	Window Function with Hamming Shape
w_{R}	Window Function with Rectangular Shape
w_{T}	Window Function with Bartlett Shape
$X(\cdot)$	Fourier Representation of the Signal x
$\mathfrak{x}(f)$	Normed Frequency Position on the Cochlea
$X^*(\cdot)$	Conjugate-Complex Spectrum of the Signal x
$\mathfrak{z}(f)$	Critical Band Rate (Bark)