

MPEG-7 AUDIO AND BEYOND

audio content indexing and retrieval

Hyoungh-Gook Kim | Nicolas Moreau | Thomas Sikora

 WILEY

MPEG-7 Audio and Beyond

Audio Content Indexing and Retrieval

Hyung-Gook Kim

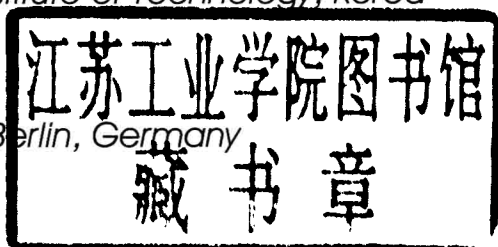
Samsung Advanced Institute of Technology, Korea

Nicolas Moreau

Technical University of Berlin, Germany

Thomas Sikora

Communication Systems Group, Technical University of Berlin, Germany



John Wiley & Sons, Ltd

Copyright © 2005

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk

Visit our Home Page on www.wiley.com

All Rights Reserved. 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 under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to permreq@wiley.co.uk, or faxed to (+44) 1243 770620.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging in Publication Data

Kim, Hyoung-Gook.

Introduction to MPEG-7 audio / Hyoung-Gook Kim, Nicolas Moreau, Thomas Sikora.
p. cm.

Includes bibliographical references and index.

ISBN-13 978-0-470-09334-4 (cloth: alk. paper)

ISBN-10 0-470-09334-X (cloth: alk. paper)

1. MPEG (Video coding standard) 2. Multimedia systems. 3. Sound—Recording and reproducing—Digital techniques—Standards. I. Moreau, Nicolas. II. Sikora, Thomas. III. Title.

TK6680.5.K56 2005

006.6'96—dc22

2005011807

British Library Cataloguing in Publication Data

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

ISBN-13 978-0-470-09334-4 (HB)

ISBN-10 0-470-09334-X (HB)

Typeset in 10/12pt Times by Integra Software Services Pvt. Ltd, Pondicherry, India

Printed and bound in Great Britain by TJ International Ltd, Padstow, Cornwall

This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

MPEG-7 Audio and Beyond

Acronyms

ADSR	Attack, Decay, Sustain, Release
AFF	Audio Fundamental Frequency
AH	Audio Harmonicity
AP	Audio Power
ASA	Auditory Scene Analysis
ASB	Audio Spectrum Basis
ASC	Audio Spectrum Centroid
ASE	Audio Spectrum Envelope
ASF	Audio Spectrum Flatness
ASP	Audio Spectrum Projection
ASR	Automatic Speech Recognition
ASS	Audio Spectrum Spread
AWF	Audio Waveform
BIC	Bayesian Information Criterion
BP	Back Propagation
BPM	Beats Per Minute
CASA	Computational Auditory Scene Analysis
CBID	Content-Based Audio Identification
CM	Coordinate Matching
CMN	Cepstrum Mean Normalization
CRC	Cyclic Redundancy Checking
DCT	Discrete Cosine Transform
DDL	Description Definition Language
DFT	Discrete Fourier Transform
DP	Dynamic Programming
DS	Description Scheme
DSD	Divergence Shape Distance
DTD	Document Type Definition
EBP	Error Back Propagation
ED	Edit Distance
EM	Expectation and Maximization
EMIM	Expected Mutual Information Measure

EPM	Exponential Pseudo Norm
FFT	Fast Fourier Transform
GLR	Generalized Likelihood Ratio
GMM	Gaussian Mixture Model
GSM	Global System for Mobile Communications
HCNN	Hidden Control Neural Network
HMM	Hidden Markov Model
HR	Harmonic Ratio
HSC	Harmonic Spectral Centroid
HSD	Harmonic Spectral Deviation
HSS	Harmonic Spectral Spread
HSV	Harmonic Spectral Variation
ICA	Independent Component Analysis
IDF	Inverse Document Frequency
INED	Inverse Normalized Edit Distance
IR	Information Retrieval
ISO	International Organization for Standardization
KL	Karhunen–Loève
KL	Kullback–Leibler
KS	Knowledge Source
LAT	Log Attack Time
LBG	Linde–Buzo–Gray
LD	Levenshtein Distance
LHSC	Local Harmonic Spectral Centroid
LHSD	Local Harmonic Spectral Deviation
LHSS	Local Harmonic Spectral Spread
LHSV	Local Harmonic Spectral Variation
LLD	Low-Level Descriptor
LM	Language Model
LMPS	Logarithmic Maximum Power Spectrum
LP	Linear Predictive
LPC	Linear Predictive Coefficient
LPCC	Linear Prediction Cepstrum Coefficient
LSA	Log Spectral Amplitude
LSP	Linear Spectral Pair
LVCSR	Large-Vocabulary Continuous Speech Recognition
mAP	Mean Average Precision
MCLT	Modulated Complex Lapped Transform
MD5	Message Digest 5
MFCC	Mel-Frequency Cepstrum Coefficient
MFFE	Multiple Fundamental Frequency Estimation
MIDI	Music Instrument Digital Interface
MIR	Music Information Retrieval
MLP	Multi-Layer Perceptron

M.M.	Metronom Mälzel
MMS	Multimedia Mining System
MPEG	Moving Picture Experts Group
MPS	Maximum Power Spectrum
MSD	Maximum Squared Distance
NASE	Normalized Audio Spectrum Envelope
NMF	Non-Negative Matrix Factorization
NN	Neural Network
OOV	Out-Of-Vocabulary
OPCA	Oriented Principal Component Analysis
PCA	Principal Component Analysis
PCM	Phone Confusion Matrix
PCM	Pulse Code Modulated
PLP	Perceptual Linear Prediction
PRC	Precision
PSM	Probabilistic String Matching
QBE	Query-By-Example
QBH	Query-By-Humming
RASTA	Relative Spectral Technique
RBF	Radial Basis Function
RCL	Recall
RMS	Root Mean Square
RSV	Retrieval Status Value
SA	Spectral Autocorrelation
SC	Spectral Centroid
SCP	Speaker Change Point
SDR	Spoken Document Retrieval
SF	Spectral Flux
SFM	Spectral Flatness Measure
SNF	Spectral Noise Floor
SOM	Self-Organizing Map
STA	Spectro-Temporal Autocorrelation
STFT	Short-Time Fourier Transform
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TA	Temporal Autocorrelation
TPBM	Time Pitch Beat Matching
TC	Temporal Centroid
TDNN	Time-Delay Neural Network
ULH	Upper Limit of Harmonicity
UM	Ukkonen Measure
UML	Unified Modeling Language
VCV	Vowel-Consonant-Vowel
VQ	Vector Quantization

VSM	Vector Space Model
XML	Extensible Markup Language
ZCR	Zero Crossing Rate

The 17 MPEG-7 Low-Level Descriptors:

AFF	Audio Fundamental Frequency
AH	Audio Harmonicity
AP	Audio Power
ASB	Audio Spectrum Basis
ASC	Audio Spectrum Centroid
ASE	Audio Spectrum Envelope
ASF	Audio Spectrum Flatness
ASP	Audio Spectrum Projection
ASS	Audio Spectrum Spread
AWF	Audio Waveform
HSC	Harmonic Spectral Centroid
HSD	Harmonic Spectral Deviation
HSS	Harmonic Spectral Spread
HSV	Harmonic Spectral Variation
LAT	Log Attack Time
SC	Spectral Centroid
TC	Temporal Centroid

Symbols

Chapter 2

n	time index
$s(n)$	digital audio signal
F_s	sampling frequency
l	frame index
L	total number of frames
$w(n)$	windowing function
L_w	length of a frame
N_w	length of a frame in number of time samples
$HopSize$	time interval between two successive frames
N_{hop}	number of time samples between two successive frames
k	frequency bin index
$f(k)$	frequency corresponding to the index k
$S_l(k)$	spectrum extracted from the l th frame
$P_l(k)$	power spectrum extracted from the l th frame
N_{FT}	size of the fast Fourier transform
ΔF	frequency interval between two successive FFT bins
r	spectral resolution
b	frequency band index
B	number of frequency bands
loF_b	lower frequency limit of band b
hiF_b	higher frequency limit of band b
$\Gamma_l(m)$	normalized autocorrelation function of the l th frame
m	autocorrelation lag
T_0	fundamental period
f_0	fundamental frequency
h	index of harmonic component
N_H	number of harmonic components
f_h	frequency of the h th harmonic
A_h	amplitude of the h th harmonic
V_E	reduced SVD basis
W	ICA transformation matrix

Chapter 3

X	feature matrix ($L \times F$)
L	total number of frames
l	frame index
F	number of columns in X (frequency axis)
f	frequency band index
E	size of the reduced space
U	row basis matrix ($L \times L$)
D	diagonal singular value matrix ($L \times F$)
V	matrix of transposed column basis functions ($F \times F$)
V_E	reduced SVD matrix ($F \times E$)
\hat{X}	normalized feature matrix
μ_f	mean of column f
μ_l	mean of row l
Γ_l	standard deviation of row l
χ_l	energy of the NASE
V	matrix of orthogonal eigenvectors
D	diagonal eigenvalue matrix
C	covariance matrix
C_P	reduced eigenvalues of D
C_E	reduced PCA matrix ($F \times E$)
P	number of components
S	source signal matrix ($P \times F$)
W	ICA mixing matrix ($L \times P$)
N	matrix of noise signals ($L \times F$)
\check{X}	whitened feature matrix
H	NMF basis signal matrix ($P \times F$)
G	mixing matrix ($L \times P$)
H_E	matrix H with $P = E(E \times F)$
x	coefficient vector
d	dimension of the coefficient space
λ	parameter set of a GMM
M	number of mixture components
$b_m(x)$	Gaussian density (component m)
μ_m	mean vector of component m
Σ_m	covariance matrix of component m
c_m	weight of component m
N_S	number of hidden Markov model states
S_i	hidden Markov model state number i
b_i	observation function of state S_i
a_{ij}	probability of transition between states S_i and S_j
π_i	probability that S_i is the initial state
θ	parameters of a hidden Markov model

w, b	parameters of a hyperplane
$d(w, b)$	distance between the hyperplane and the closest sample
α_i	Lagrange multiplier
$L(w, b, \alpha)$	Lagrange function
$K(\cdot, \cdot)$	kernel mapping
R_l	RMS-norm gain of the l th frame
X_l	NASE vector of the l th frame
Y	audio spectrum projection

Chapter 4

X	acoustic observation
w	word (or symbol)
W	sequence of words (or symbols)
λ_w	hidden Markov model of symbol w
S_i	hidden Markov model state number i
b_i	observation function of state S_i
a_{ij}	probability of transition between states S_i and S_j
D	description of a document
Q	description of a query
d	vector representation of document D
q	vector representation of query Q
t	indexing term
$q(t)$	weight of term t in q
$d(t)$	weight of term t in d
T	indexing term space
N_T	number of terms in T
$s(t_i, t_j)$	measure of similarity between terms t_i and t_j

Chapter 5

n	note index
$f(n)$	pitch of note n
F_s	sampling frequency
F_0	fundamental frequency
$scale(n)$	scale value for pitch n in a scale
$i(n)$	interval value for note n
$d(n)$	differential onset for note n
$o(n)$	time of onset of note n
C	melody contour
M	number of interval values in C
$m(i)$	interval value in C

$G(i)$	n -gram of interval values in C
Q	query representation
D	music document
Q_N	set of n -grams in Q
D_N	set of n -grams in D
c_d	cost of an insertion or deletion
c_m	cost of a mismatch
c_e	value of an exact match
U, V	MPEG-7 beat vectors
$u(i)$	i th coefficient of vector U
$v(j)$	j th coefficient of vector V
R	distance measure
S	similarity score
$\langle t, p, b \rangle$	time t , pitch p , beat b triplet
$\langle t_m, p_m, b_m \rangle$	melody segment m
$\langle t_q, p_q, b_q \rangle$	query segment q
n	measure number
S_n	similarity score of measure n
s_m	subsets of melody pitch p_m
s_q	subsets of query pitch p_q
i, j	contour value counters

Chapter 6

L_S	length of the digital signal in number of samples
N_{CH}	number of channels
$s_i(n)$	digital signal in the i th channel
$\Gamma_{si,sj}$	cross-correlation between channels i and j
P_i	mean power of the i th channel

Chapter 7

X_i	sub-sequence of feature vectors
μ_{X_i}	mean value of X_i
Σ_{X_i}	covariance matrix of X_i
N_{X_i}	number of feature vectors in X_i
R	generalized likelihood ratio
D	penalty

Contents

List of Acronyms	xi
List of Symbols	xv
1 Introduction	1
1.1 Audio Content Description	2
1.2 MPEG-7 Audio Content Description – An Overview	3
1.2.1 MPEG-7 Low-Level Descriptors	5
1.2.2 MPEG-7 Description Schemes	6
1.2.3 MPEG-7 Description Definition Language (DDL)	9
1.2.4 BiM (Binary Format for MPEG-7)	9
1.3 Organization of the Book	10
2 Low-Level Descriptors	13
2.1 Introduction	13
2.2 Basic Parameters and Notations	14
2.2.1 Time Domain	14
2.2.2 Frequency Domain	15
2.3 Scalable Series	17
2.3.1 Series of Scalars	18
2.3.2 Series of Vectors	20
2.3.3 Binary Series	22
2.4 Basic Descriptors	22
2.4.1 Audio Waveform	23
2.4.2 Audio Power	24
2.5 Basic Spectral Descriptors	24
2.5.1 Audio Spectrum Envelope	24
2.5.2 Audio Spectrum Centroid	27
2.5.3 Audio Spectrum Spread	29
2.5.4 Audio Spectrum Flatness	29
2.6 Basic Signal Parameters	32
2.6.1 Audio Harmonicity	33
2.6.2 Audio Fundamental Frequency	36

2.7	Timbral Descriptors	38
2.7.1	Temporal Timbral: Requirements	39
2.7.2	Log Attack Time	40
2.7.3	Temporal Centroid	41
2.7.4	Spectral Timbral: Requirements	42
2.7.5	Harmonic Spectral Centroid	45
2.7.6	Harmonic Spectral Deviation	47
2.7.7	Harmonic Spectral Spread	47
2.7.8	Harmonic Spectral Variation	48
2.7.9	Spectral Centroid	48
2.8	Spectral Basis Representations	49
2.9	Silence Segment	50
2.10	Beyond the Scope of MPEG-7	50
2.10.1	Other Low-Level Descriptors	50
2.10.2	Mel-Frequency Cepstrum Coefficients	52
	References	55
3	Sound Classification and Similarity	59
3.1	Introduction	59
3.2	Dimensionality Reduction	61
3.2.1	Singular Value Decomposition (SVD)	61
3.2.2	Principal Component Analysis (PCA)	62
3.2.3	Independent Component Analysis (ICA)	63
3.2.4	Non-Negative Factorization (NMF)	65
3.3	Classification Methods	66
3.3.1	Gaussian Mixture Model (GMM)	66
3.3.2	Hidden Markov Model (HMM)	68
3.3.3	Neural Network (NN)	70
3.3.4	Support Vector Machine (SVM)	71
3.4	MPEG-7 Sound Classification	73
3.4.1	MPEG-7 Audio Spectrum Projection (ASP) Feature Extraction	74
3.4.2	Training Hidden Markov Models (HMMs)	77
3.4.3	Classification of Sounds	79
3.5	Comparison of MPEG-7 Audio Spectrum Projection vs. MFCC Features	79
3.6	Indexing and Similarity	84
3.6.1	Audio Retrieval Using Histogram Sum of Squared Differences	85
3.7	Simulation Results and Discussion	85
3.7.1	Plots of MPEG-7 Audio Descriptors	86
3.7.2	Parameter Selection	88
3.7.3	Results for Distinguishing Between Speech, Music and Environmental Sound	91

3.7.4	Results of Sound Classification Using Three Audio Taxonomy Methods	92
3.7.5	Results for Speaker Recognition	96
3.7.6	Results of Musical Instrument Classification	98
3.7.7	Audio Retrieval Results	99
3.8	Conclusions	100
	References	101
4	Spoken Content	103
4.1	Introduction	103
4.2	Automatic Speech Recognition	104
4.2.1	Basic Principles	104
4.2.2	Types of Speech Recognition Systems	108
4.2.3	Recognition Results	111
4.3	MPEG-7 <i>SpokenContent</i> Description	113
4.3.1	General Structure	114
4.3.2	<i>SpokenContentHeader</i>	114
4.3.3	<i>SpokenContentLattice</i>	121
4.4	Application: Spoken Document Retrieval	123
4.4.1	Basic Principles of IR and SDR	124
4.4.2	Vector Space Models	130
4.4.3	Word-Based SDR	135
4.4.4	Sub-Word-Based Vector Space Models	140
4.4.5	Sub-Word String Matching	154
4.4.6	Combining Word and Sub-Word Indexing	161
4.5	Conclusions	163
4.5.1	MPEG-7 Interoperability	163
4.5.2	MPEG-7 Flexibility	164
4.5.3	Perspectives	166
	References	167
5	Music Description Tools	171
5.1	Timbre	171
5.1.1	Introduction	171
5.1.2	<i>InstrumentTimbre</i>	173
5.1.3	<i>HarmonicInstrumentTimbre</i>	174
5.1.4	<i>PercussiveInstrumentTimbre</i>	176
5.1.5	Distance Measures	176
5.2	Melody	177
5.2.1	<i>Melody</i>	177
5.2.2	<i>Meter</i>	178
5.2.3	<i>Scale</i>	179
5.2.4	<i>Key</i>	181

5.2.5	<i>MelodyContour</i>	182
5.2.6	<i>MelodySequence</i>	185
5.3	Tempo	190
5.3.1	<i>AudioTempo</i>	192
5.3.2	<i>AudioBPM</i>	192
5.4	Application Example: Query-by-Humming	193
5.4.1	Monophonic Melody Transcription	194
5.4.2	Polyphonic Melody Transcription	196
5.4.3	Comparison of Melody Contours	200
	References	203
6	Fingerprinting and Audio Signal Quality	207
6.1	Introduction	207
6.2	Audio Signature	207
6.2.1	Generalities on Audio Fingerprinting	207
6.2.2	Fingerprint Extraction	211
6.2.3	Distance and Searching Methods	216
6.2.4	MPEG-7-Standardized <i>AudioSignature</i>	217
6.3	Audio Signal Quality	220
6.3.1	<i>AudioSignalQuality</i> Description Scheme	221
6.3.2	<i>BroadcastReady</i>	222
6.3.3	<i>IsOriginalMono</i>	222
6.3.4	<i>BackgroundNoiseLevel</i>	222
6.3.5	<i>CrossChannelCorrelation</i>	223
6.3.6	<i>RelativeDelay</i>	224
6.3.7	<i>Balance</i>	224
6.3.8	<i>DcOffset</i>	225
6.3.9	<i>Bandwidth</i>	226
6.3.10	<i>TransmissionTechnology</i>	226
6.3.11	<i>ErrorEvent</i> and <i>ErrorEventList</i>	226
	References	227
7	Application	231
7.1	Introduction	231
7.2	Automatic Audio Segmentation	234
7.2.1	Feature Extraction	235
7.2.2	Segmentation	236
7.2.3	Metric-Based Segmentation	237
7.2.4	Model-Selection-Based Segmentation	242
7.2.5	Hybrid Segmentation	243
7.2.6	Hybrid Segmentation Using MPEG-7 ASP	246
7.2.7	Segmentation Results	250

7.3	Sound Indexing and Browsing of Home Video Using Spoken Annotations	254
7.3.1	A Simple Experimental System	254
7.3.2	Retrieval Results	258
7.4	Highlights Extraction for Sport Programmes Using Audio Event Detection	259
7.4.1	Goal Event Segment Selection	261
7.4.2	System Results	262
7.5	A Spoken Document Retrieval System for Digital Photo Albums	265
	References	266
	Index	271