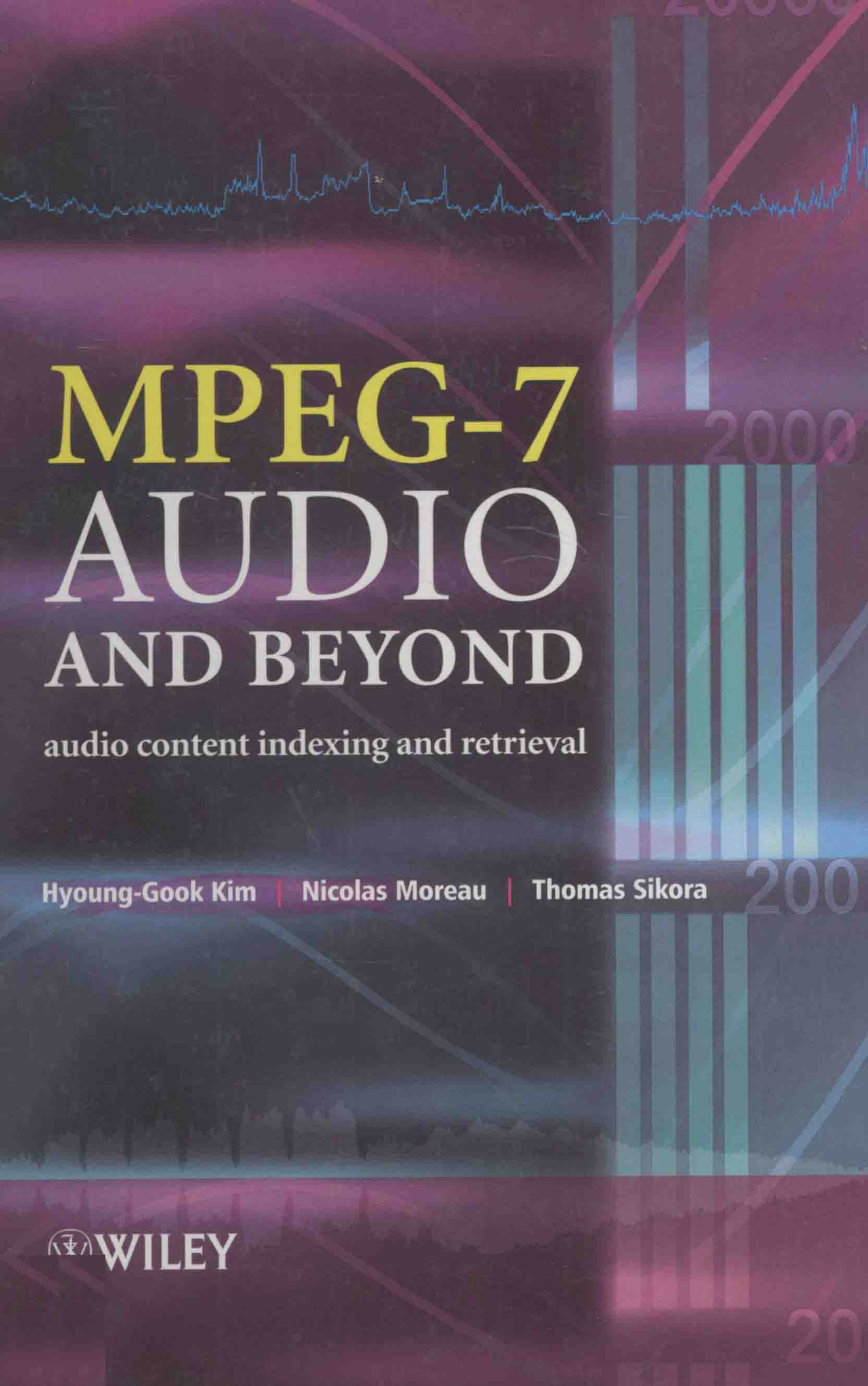

MPEG-7 AUDIO AND BEYOND

audio content indexing and retrieval

Hyoung-Gook Kim | Nicolas Moreau | Thomas Sikora

WILEY

MPEG-7 Audio and Beyond

Audio Content Indexing and Retrieval

Hyoung-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

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

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 lth frame
$P_l(k)$	power spectrum extracted from the lth 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 lth 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 hth harmonic
A_h	amplitude of the hth 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 lth frame
X_l	NASE vector of the lth 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
$u(i)$	ith coefficient of vector U
$v(j)$	jth 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 ith channel
$\Gamma_{si,sj}$	cross-correlation between channels i and j
P_i	mean power of the ith channel

Chapter 7

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

Contents