Music Information Retrieval
Theory and Applications

George Tzanetakis (gtzan@cs.uvic.ca)
Assistant Professor
Computer Science Department
(also in Music, ECE)
University of Victoria, Canada
Overview

- Introduction and history
- Audio and music-specific representations
- Analysis and retrieval tasks
- Tools, interfaces and applications
Music Retrieval through time
The not so far future of MIR

Library of all recorded music

organize    classify

search      retrieve

browse
MIR History

- Pre-history - before 2000 mostly symbolic based and scattered
- First ISMIR 2000
- Symposium -> conference -> society
- 10th ISMIR, Kobe, Japan 2009
- Recently formed Society for MIR
- Increasing presence in ICASSP, ICME, ACMM, TMM, TASLP, MMTA
## ISMIR Growth

<table>
<thead>
<tr>
<th>Year</th>
<th># papers</th>
<th># attendees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>35</td>
<td>88</td>
</tr>
<tr>
<td>2002</td>
<td>41</td>
<td>95</td>
</tr>
<tr>
<td>2003</td>
<td>57</td>
<td>184</td>
</tr>
<tr>
<td>2004</td>
<td>50</td>
<td>140</td>
</tr>
<tr>
<td>2005</td>
<td>105</td>
<td>194</td>
</tr>
<tr>
<td>2006</td>
<td>114</td>
<td>230</td>
</tr>
<tr>
<td>2007</td>
<td>95</td>
<td>158</td>
</tr>
<tr>
<td>2007</td>
<td>127</td>
<td>??</td>
</tr>
<tr>
<td>2008</td>
<td>105</td>
<td>??</td>
</tr>
<tr>
<td>2009</td>
<td>127</td>
<td>26</td>
</tr>
</tbody>
</table>
Community

- music-ir mailing list
- http://www.music-ir.org/
- http://www.ismir.net/
- Future webpage home of this tutorial (will send via email)
Industry

lost.fm

gracenote.

the echonest

Copyright 2009 G. Tzanetakis
Digital Music Distribution

Number of Songs Sold on the iTunes Store

<table>
<thead>
<tr>
<th>Billionth Song</th>
<th>Days taken</th>
<th>Songs purchased</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>1033</td>
<td>9,949,000,000</td>
</tr>
<tr>
<td>2nd</td>
<td>322</td>
<td>3,712,000,000</td>
</tr>
<tr>
<td>3rd</td>
<td>203</td>
<td>4,993,000,000</td>
</tr>
<tr>
<td>4th</td>
<td>169</td>
<td>5,997,000,000</td>
</tr>
<tr>
<td>5th</td>
<td>157</td>
<td>6,993,000,000</td>
</tr>
<tr>
<td>6th</td>
<td>202</td>
<td>4,949,000,000</td>
</tr>
</tbody>
</table>
MIR Dimensions

**BROWSING**

Playlists  Shuffle

**SPECIFICITY**

Fingerprinting  Remix  CoverSong  Artist  Genre

**INFORMATION**

Score  Melody  Structure  Harmony  Rhythm  Texture
Connections

Machine Learning  
Computer Science  

Signal Processing  
Information Science  

Psychology  
Human-Computer Interaction  

MUSIC
MIR @

- C5, Tuesday 10:30 (Audio and Music): “Comprehensive Query-Dependent Fusion using Regression-on-Folksonomies: A Case Study of Multimodal Retrieval” - Zhang, Xiang, Lu, Wang

- C2, Tuesday 12:30: “Improving Automatic Music Tag Annotation using Stacked Generalization of SVM outputs” - Ness, Theocharis, Martins, Tzanetakis

- “SVR-based Music Mood Classification and Context-based Music Recommendation” - Rho, Han, Hwang

- Open Source Competition, Thursday 16:00: “Marsyas”
Audio MIR Pipeline

Hearing Representation → Understanding Analysis → Acting Interaction

Signal Processing Feature Extraction → Machine Learning Analysis and Retrieval → Human Computer Interaction
Feature extraction

Feature Space

Feature vector

Copyright 2009 G. Tzanetakis
Audio Feature Extraction

- Sound and sine waves
- Timbral Features
  - STFT, DWT, MFCC, MP3
- Rhythm Analysis
- Pitch/Harmonic Analysis
Timbral Texture

Timbre = differentiate sounds of same pitch and loudness
Timbral Texture = differentiate mixtures of sounds (possibly with the same or similar rhythmic and pitch content)
Linear Systems and Sine Waves

Amplitude

Phase

Period = 1 / Frequency

True sine waves last forever

sine wave -> LTI -> new sine wave
Time-Frequency Analysis

Fourier Transform

\[ f(x) = \sum_{n=0}^{\infty} a_n \cos(n \pi x) + \sum_{n=0}^{\infty} b_n \sin(n \pi x) \]

\[ f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(\omega) e^{-i\omega t} \, dt \]

\[ f(\omega) = \int_{-\infty}^{\infty} f(t) e^{i\omega t} \, dt \]

\[ e^{i\theta} = \cos(\theta) + i \sin(\theta) \]

\[ e^{i\pi} = \cos(\pi) + i \sin(\pi) \]
Short Time Fourier Transform

Time-varying spectra
Fast Fourier Transform FFT
Demos (waterfall display + phasevocoder)
Short Time
Fourier Transform I

FT = global representation of frequency content

\[ S_f(u, \omega) = \int_{-\infty}^{\infty} f(t) g(t-u) e^{-i \omega t} dt \]

Time – Frequency

L2 Heisenberg uncertainty

\[ \sigma_t \sigma_\omega \geq 1/4 \]
STFT- Wavelets

Time – Frequency Heisenberg uncertainty

$\sigma_t \sigma_\omega \geq 1/4$

Copyright 2009 G.Tzanetakis
A filterbank view of STFT and DWT

STFT

Impulse input

| BW = 1KHz |
| CF = 500Hz |
| BW = 1KHz |
| CF = 1500Hz |
| BW = 1KHz |
| CF = 2500Hz |

DWT

Impulse input

| BW = 200 |
| BW = 400 |
| BW = 800 |
The Discrete Wavelet Transform

Octave filterbank

Analysis

Synthesis
Spectrum and Shape Descriptors

Centroid
Rolloff
Flux
Bandwidth
Moments
....

Feature vector = M

Feature Space

Centroid

25

Copyright 2009 G. Tzanetakis
Perception-based approaches

- Pitch perception
- Loudness perception
- Critical Bands
- Mel-Frequency Cepstral Coefficients
- Masking
- Perceptual Audio Compression (MPEG)
The Human Ear

Pinna
Auditory canal
Ear Drum
Stapes-Malleus-Incus (gain control)
Cochlea (freq. analysis)
Auditory Nerve

Wave travels to cutoff slowing down increasing in amplitude, power is absorbed
Each frequency has a position of maximum displacement
Masking

Two frequencies -> beats
-> harsh
-> separate

Inner Hair Cell excitation

Frequency Masking
Temporal Masking

Pairs of sine waves (one softer) – how much weaker in order to be masked?
(masking curves) wave of high frequency can not mask a wave of lower frequency.
Mel Frequency Cepstral Coefficients

Mel-scale
13 linearly-spaced filters
27 log-spaced filters

CF
CF / 1.0718
CF * 1.0718
CF-130
CF+130

Mel-filtering
Log
DCT
MFCCs
Discrete Cosine Transform

- Strong energy compaction
- For certain types of signals approximates KL transform (optimal)
- Low coefficients represent most of the signal - can throw high
- MFCCs keep first 13-20
- MDCT (overlap-based) used in MP3, AAC, Vorbis audio compression
MPEG Audio
Feature Extraction

Perceptual Audio Coding (slow encoding, fast decoding)
Psychoacoustic Model

- Each band is quantized
- Quantization introduces noise
- Adapt the quantization so that it is inaudible
- Take advantage of masking
  - Hide quantization noise where it is masked
- MPEG standardizes how the quantized bits are transmitted not the psychoacoustic model - (only recommended)
Analysis and Texture Windows

Running multidimensional Gaussian distribution (means, variances over texture window)

Speech
Orchestra
Piano

Analysis windows
Texture windows

20 milliseconds
40 analysis windows
Audio Feature Extraction

Waveform

Spectrogram

MFCCs

Texture Window

1 sec

Feature Vector

10 sec
Summary of Timbral Texture Features

- Time-Frequency analysis
- Signal processing (STFT, DWT)
- Perceptual (MFCC, MPEG)
- Statistics over “texture” window
Traditional Music Representations

GUIDO Noteserver. Powered by the SALIERI-Project. http://www.informatik.tu-darmstadt.de/AFS/SALIERI

Fast Latin Jazz (j = 120)

Manual signs for the tones of the scale, (From Carreño's "Chains of Cuba".)
Rhythm

- Rhythm = movement in time
- Origins in poetry (iamb, trochaic...)
- Foot tapping definition
- Hierarchical semi-periodic structure at multiple levels of detail
- Links to motion, other sounds
- Running versus global description (tempo induction, beat analysis, beat tracking, causal beat tracking)
Automatic Rhythm
Description

(A) Raw data (audio)

(A') Feature lists (e.g., onsets, frame energy)

(B) Metrical structure and timing features (e.g. gradually decreasing tempo)
Event-based

Event onset detection → IOI computation → Beat model

Single band
Multiple band
Thresholding
Chord changes

Successive onset pairs
All onset pairs
Quantization

Constant-tempo
Measure changes
Computational cost
Multiple hypotheses
Beat Histogram Calculation

Input Signal

DWT

Envelope Extraction
Full Wave Rectification - Low Pass Filtering - Normalization

Autocorrelation

Peak Picking

Beat Histogram

Copyright 2009 G.Tzanetakis
Beat Histograms

max(h(i)), argmax(h(i)) → Beat Histogram Features

CLASSICAL

HIP-HOP
Beat Spectrum

Figure 4. Beat spectrogram of Pink Floyd’s *Money* (excerpt), showing transition from 4/4 to 7/4 time.
Rhythmic content features

- Main tempo
- Secondary tempo
- Time signature
- Beat strength
- Regularity
Representations

- **Score**
  - Discrete, high level abstraction, explicit structure, no performance info

- **MIDI**
  - Discrete, medium level of abstraction, explicit time but less structure, targeted to keyboard performance

- **Audio**
  - Continuous, low level abstraction, timing and structure implicit
MIDI

- Musical Instrument Digital Interfaces
  - Hardware interface
  - File Format
- Note events
  - Duration, discrete pitch, "instrument"
- Extensions
  - General MIDI
  - Notation, OMR, continuous pitch
Pitch Perception I

- Pitch is not just fundamental frequency
- Periodicity or harmonicity or both?
- Human judgements (adjust sine method)
- 1924 Fletcher - harmonic partials missing fundamental (pitch is still heard)
  - Examples: phone, small radio
- Terhardt (1972), Licklider (1959)
  - duplex theory of pitch (virtual & spectral pitch)


Pitch Perception II

- One perception - two overlapping mechanisms
  - Counting cycles of period $< 800$Hz
  - Place of excitation along basilar membrane $> 1600$ Hz

![Diagram](FFT \rightarrow \text{Pick sinusoids} \rightarrow \text{weighting / masking} \rightarrow \text{candidate generation})

- most likely pitch
- common divisor
Pitch content

- Harmony, melody = pitch concepts
- Music Theory Score = Music
- Bridge to symbolic MIR
- Automatic music transcription
Pitch Detection

Autocorrelation
Peaks at multiple of the fundamental frequency
ZeroCrossings

\[ r_x = \sum_{n=0}^{N-1} x(n)x(n+l), l=0,1,\ldots L-1 \]

Rhythm -> ~20 Hz Pitch
(courtesy of R.Dannenberg – Nyquist)
Multiple Pitch Detection

Tolonen and Karjalainen, TSAP00

> 1kHz → Half-wave Rectifier LowPass

< 1kHz

Periodicity Detection

SACF Enhancer

Pitch Candidates

Tzanetakis et al, ISMIR 01

(7 * c ) mod 12

Circle of 5s

C G

Copyright 2009 G.Tzanetakis
Chroma - Pitch perception
Audio MIR Pipeline

Hearing Representation

Understanding Analysis

Acting Interaction

Signal Processing

Feature Extraction

Machine Learning Analysis and Retrieval

Human Computer Interaction
Analysis Overview

Musical Piece Trajectory

Point

Trajectory Musical Piece
Content-based Similarity Retrieval (or query-by-example)

Query:
Ranked list of similar audio files based on feature vector similarity:
TEXTURE, BEAT, FULL
Content-based MIR
(“Google” for music)

<table>
<thead>
<tr>
<th>Orchestral</th>
<th>Rock</th>
<th>Jazz</th>
<th>Opera</th>
<th>Reggae</th>
<th>Rap</th>
<th>World</th>
<th>Piano</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Orchestral" /></td>
<td><img src="image2" alt="Rock" /></td>
<td><img src="image3" alt="Jazz" /></td>
<td><img src="image4" alt="Opera" /></td>
<td><img src="image5" alt="Reggae" /></td>
<td><img src="image6" alt="Rap" /></td>
<td><img src="image7" alt="World" /></td>
<td><img src="image8" alt="Piano" /></td>
</tr>
</tbody>
</table>

Rank List

- Red
- Red
- Red
- Red
- Blue
- Blue
- Blue
- Blue

3500 clips (30 sec) - no Kenny G
NO METADATA USED
Some “interesting searches”
Analysis Overview

Musical Piece

Trajectory  Point
Partitioning of feature space
Generative vs discriminative models

\[ P(\square|\circ) = \frac{p(\circ|\square) \times P(\square)}{P(\circ)} \]

Decision boundary

- Music
- Speech
Non-parametric classifiers

\[ P(\text{\textbullet} | \text{\textcircled{\textbullet}}) = \frac{p(\text{\textbullet} | \text{\textcircled{\textbullet}}) \times P(\text{\textcircled{\textbullet}})}{p(\text{\textbullet})} \]

Nearest-neighbor classifiers (K-NN)
Parametric classifiers

Gaussian Classifier

Gaussian Mixture Models

\[ P(\bullet|\circ) = p(\bullet) \ast P(\circ) \]
Cross-validation

Overfitting

Training set
Testing set

Overfitting – generalization
Support Vector Machines

Decision boundary training error is not sufficient for choosing A, B or C

Large margin results in better generalization.
Supervised Learning

- Labeled data
  - Training set, testing set
  - Cross validation
- Classifiers
  - Gaussian + Gaussian Mixture Models
  - K Nearest Neighbors
  - Backpropagation Artificial Neural Network
  - Support Vector Machine
Unsupervised Learning
Clustering

Classify

k-means

Re-estimate
Automatic Musical Genre Classification

- Categorical music descriptions created by humans
  - Fuzzy boundaries
- Statistical properties
  - Timbral texture, rhythmic structure, harmonic content
- Automatic Musical Genre Classification
  - Evaluate musical content features
  - Structure audio collections
Classification Evaluation - 10 genres

Manual (52 subjects)
Perrot & Gjerdingen, M.Cognition 99

0.25 seconds 40% (demo)
3 seconds 70%

Automatic (different collection)
Tzanetakis & Cook, ISMIR 01, TSAP02

Gaussian Mixture Model (GMM)
Support Vector Machine (SVM)

Classification Accuracy

Random
GMM
SVM
M02-SVM

Manual  (52 subjects)
Perrot & Gjerdingen, M.Cognition 99

0.25 seconds 40% (demo)
3 seconds 70%

Automatic  (different collection)
Tzanetakis & Cook, ISMIR 01, TSAP02

Gaussian Mixture Model (GMM)
Support Vector Machine (SVM)

Classification Accuracy
Comparison of human and automatic genre classification

160 tracks
Representing western music consumption
Labels: classical, dance, pop, rap, rock, other
Voted by 27 human listeners

Histogram of votes for elected genre:

Human accuracy: range 57%-86%, avg 76%
Subset MAMI2: range 69%-98%, avg 90%

Accuracy calculated comparing to the majority vote for genre – no outside ground truth

MAMI2: the reliable songs

Lippens et al. ICASPP
Comparison of human and automatic genre classification

Lippens et al. ICASSP 04

Automatic classification performs as well as the "worst" human

"best" human is 20% more accurate

Difference is probably smaller today (better features, SVMs)
Glass ceiling

- Performance of MIR systems based on the "classic" approach has been reaching a plateau
- Approximately 75% for genre classification

<table>
<thead>
<tr>
<th>Participant</th>
<th>Avg. Hierarchical Classification Accuracy</th>
<th>Avg. Raw Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH</td>
<td>71.87%</td>
<td>62.89%</td>
</tr>
<tr>
<td>IM_knn</td>
<td>64.83%</td>
<td>54.87%</td>
</tr>
<tr>
<td>IM_svm</td>
<td>76.56%</td>
<td>68.29%</td>
</tr>
<tr>
<td>TL</td>
<td>75.57%</td>
<td>66.71%</td>
</tr>
<tr>
<td>ME</td>
<td>75.03%</td>
<td>66.60%</td>
</tr>
<tr>
<td>ME_spec</td>
<td>73.57%</td>
<td>65.50%</td>
</tr>
<tr>
<td>GT</td>
<td>74.15%</td>
<td>65.34%</td>
</tr>
</tbody>
</table>

MIREX 2007
10 genres
700 30-second clips / genre
GenreGram DEMO

Dynamic real time 3D display for classification of radio signals
Multi-tag annotation

- Free-form tags (female voice, woman singing)
- Multi-label classification problems with twists
- Issues: synonyms, subpart relations, sparse, noisy
- Cold start problem
- Typically each tag is treated independently as a classification problem
- Inverse also interesting (query-by-keywords)
Audio Segmentation

Segmentation: changes of sound texture

Music  Male Voice  Female Voice

News:

Segmentation = changes of sound texture
Segmentation

- Model-based
  - HMM
  - Fixed # of “textures”, no RMS

- Metric-based
  - Detect abrupt changes
  - Arbitrary # of “textures”, RMS
  - Sensitive to transients

- Hybrid

Aucouturier & Sandler, AES 01
Zang & Kuo, TSAP 01
Tzanetakis & Cook, WASPAA 99
HMM segmentation

Aucouturier & Sandler, AES 01
Multifeature Segmentation Methodology

- Time series of feature vectors $V(t)$

- $f(t) = d(V(t), V(t-1))$

  $d(x,y) = (x-y)C^{-1}(x-y)^\top$  \hspace{1cm} (Mahalanobis)

- $df/dt$ peaks correspond to texture changes
Locating singing voice segments

Berenzweig & Ellis, WASPAA

Multi-layer perceptron
2000 hidden units
54 phone classes

16 msec
p(phone class)

80% accuracy
Analysis Overview

Trajectory $\leftrightarrow$ Musical Piece $\rightarrow$ Point
POLYPHONIC AUDIO AND MIDI ALIGNMENT

Symbolic Representation
- easy to manipulate
- “flat” performance

Audio Representation
- expressive performance
- opaque & unstructured
POLYPHONIC AUDIO AND MIDI ALIGNMENT

Similarity Matrix for Beethoven’s 5th Symphony, first movement

Optimal Alignment Path

Oboe solo:
- Acoustic Recording
- Audio from MIDI

(Duration: 7:49)

(Duration: 6:17)
Polyphonic Transcription

Klapuri et al, DAFX

Mixture signal → Noise Suppression → Predominant pitch estimation → Remove detected sound

Estimate # voices iterate

Original

Transcribed

TAIVAS ON SININEN JA VALKONEN
“Classic” multi-stage approach

Time-Frequency representation

Grouping Cue 1

Grouping Cue 2

Short Time Fourier Transform
Discrete basis: windowed sine waves

Partial Tracking (McAuley & Quatieri)

Sound source formation:
  grouping of partials based on harmonicity

PROBLEMS: Difficult to decide ordering, brittle
Spectral Clustering

- Alternative to traditional point-based algorithms (k-means)
- Doesn’t assume convex shapes
- Doesn’t assume Gaussians
- Avoid multiple restarts
- Eigenstructure of
Sound Source Separation using Spectral Clustering
Comparison with partial tracking

MacAuly and Quatieri
Tracking of Partial

Proposed Approach
Synthetic Mixtures of Instruments

4-note mixture

Instrumentation detection based on timbral models

Martins, et al, ISMIR 07
“Real world” separation results

- Mirex database
- Live U2

More examples: [http://opihi.cs.uvic.ca/NormCutAudio](http://opihi.cs.uvic.ca/NormCutAudio)
Principal Component Analysis

Projection matrix

PCA
Eigenanalysis of collection correlation matrix
Self-Organizing Maps

\[ W_{v}(t + 1) = W_{v}(t) + \theta (v, t) \alpha (t)(D(t) - W_{v}(t)), \]
Self-Organizing Maps
Performance matching

Power plot

24d “pitch” vectors from FFT

Characteristic sequence

Nearest neighbor with Locality-Sensitive Hashing
Identical, different copy, different vocals, different performance (80%)

Yang, WASPAA 99

Dynamic programming
Symphonies

Foote, ISMIR 00

Copyright 2009 G.Tzanetakis
Playlist generation

$(s_1, s_2, s_3, \ldots, s_n)$ 20% slow songs, 80% fast, female jazz singers

Constraint-satisfaction problem
Smooth transitions

Technical attributes (artist, album, name)
Content attributes (jazz singer, brass)
Audio Thumbnailing

- Representative short summary of piece
  - Segmentation-based
  - Repetition-based
- Hard to evaluate
Segmentation-based Thumbnailing

- Begin and end times of a 2 second thumbnail that best represents the segment
  - 62% first two seconds of the segment
  - 92% two seconds within the first five seconds of the segment
- Automatic thumbnailing
  - first 5 sec + best effort about 80% "correct"
Structural Analysis

- Similarity matrix
- Representations
  - Notes
  - Chords
  - Chroma
- Greedy hill-climbing algorithm
  - Recognize repeated patterns
- Result = AABA (explanation)

Dannenberg & Hu, ISMIR 2002
Tzanetakis, Dannenberg & Hu, WIAMI
Similarity Matrices

Satin Doll - MIDI

Satin Doll – Audio-from-midi
An example - Naima
Repetition-based thumbnailing

Thumbnail = maximum repeated segment

Alternatives: Clustering, HMM

Logan, B., ICASSP 00
Bartch and Wakefield, WASPAA
Structure from similarity

Feature vector trajectory
Correlation at various time lags

ABAA'
Query-by-humming

- User sings a melody
- Computer searches database for song containing the melody
- Probably less useful than it sounds but interesting problem
- The challenge of difficult queries
The MUSART system

- Query preprocessing
  - Pitch contour extraction (audio)
  - Note segmentation (symbolic)
- Target preprocessing (symbolic)
  - Theme extraction
  - Model-forming, representation
- Search to find approximate match
  - Dynamic Programming, HMMs
Representations

- Pitch and tempo invariance
  - Quantized pitch intervals
  - Quantized IOI ratio
- Approximate matching
  - HMM
  - Dynamic programming
  - Time Series
Audio Fingerprinting and Watermarking

- Watermarking
  - Copyright protection
    - Proof of ownership
    - Usage policies
  - Metadata hiding
- Fingerprinting
  - Tracking
  - Copyright protection
  - Metadata linking
Watermarking

- Steganography (hiding information in messages - invisible ink)

Watermark Embedder
- watermark data
- signal repres.
- key

Watermark Extractor
- watermark data
- (original music)
- key

transmission attacks
Desired Properties

- Perceptually hidden (inaudible)
- Statistically invisible
- Robust against signal processing
- Tamper resistant
- Spread in the music, not in header
- Key dependent
Representations for Watermarking

- Basic Principles
  - Psychoacoustics
  - Spread Spectrum
    - redundant spread of information in TF plane

- Representations
  - Linear PCM
  - Compressed bitstreams
  - Phase, stereo
  - Parametric representations
Watermarking on parametric representations

Choose attack tolerance so perceptual distortion < t and lattice finding possible.
Problems with watermarking

- The security of the entire system depends on devices available to attackers
  - Breaks Kerckhoff's Criterion: A security system must work even if reverse-engineered

- Mismatch attacks
  - Time stretch audio - stretch it back (invertible)

- Oracle attacks
  - Poll watermark detector
Audio Fingerprinting

- Each song is represented as a fingerprint (small robust representation)
- Search database based on fingerprint
- Main challenges
  - highly robust fingerprint extraction
  - efficient fingerprint search strategy
- Information is summarized from the whole song - attacks degrade unlike watermarking
Hash functions

- $H(X) \rightarrow$ maps large $X$ to small hash value
- compare by comparing hash value
- Perceptual hash function?
  - impossible to get exact matching
- Perceptually similar objects result in similar fingerprints
- Detection/false alarm tradeoff
Properties

- Robustness
- Reliability
- Fingerprint size
- Granularity
- Search speed and scalability
Fraunhofer

- LLD Mpeg-7 framework (SFM)
- Vector quantization (k-means)
  - Codebook of representative vectors
- Database target signature is the codebook
- Query -> sequence of feature vectors
- Matching by finding “best” codebook
- Robust not very scalable (O(n) search))

Allamanche Ismir
2001
Philips Research

Haitsa & Kalker
Ismir 2002

- 32-bit subfingerprints for every 11.6 msec
- overlapping frames of 0.37 seconds (31/32 overlap)
- PSD -> logarithmic band spacing (bark)
- bits 0-1 sign of energy
- looks like a fingerprint
- assume one fingerprint perfect - hierarchical database layout (works ok)
Shazam Entertainment

- Pick landmarks on audio - calculate fingerprint
- Histogram of relative time differences for filtering
- Spectrogram peaks (time, frequency)
Spectrogram Peaks

Very robust – even over noisy cell phones
Audio Fingerprinting

Music piece → Signature → Matching → Signature Database

- 1 sec
- 300 bytes
- 20 msec
- 1.5 million

Copyright, metadata

Robustness

Copyright 2009 G. Tzanetakis

moodlogic.net
Evaluation

- Music Information Retrieval Evaluation Exchange (MIREX)
- 2009 5th running - hosted by J.S Downie and team at Univ. of Illinois UC
- A variety of tasks/evaluation metrics/teams
MIREX Tasks

- Audio Genre, Composer, Artist
- Audio Tag Classification
- Multiple FO
- Beat tracking
- QBH and QBT
- Audio Chord Detection
- Audio Melody Extraction
- Onset Detection
Some results (2008)

Audio Genre Classification (GenreMixed)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Participant</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tzanetakis, G. (stereo)</td>
<td>66.41%</td>
</tr>
<tr>
<td>2</td>
<td>Tzanetakis, G. (parallel)</td>
<td>65.62%</td>
</tr>
<tr>
<td>3</td>
<td>Mandel &amp; Ellis (1)</td>
<td>65.41%</td>
</tr>
<tr>
<td>4</td>
<td>Mandel &amp; Ellis (2)</td>
<td>65.30%</td>
</tr>
<tr>
<td>5</td>
<td>Mandel &amp; Ellis (3)</td>
<td>65.20%</td>
</tr>
<tr>
<td>6</td>
<td>Lidy, Rauber, Pertusa &amp; Ponce de León, Iñesta (1)</td>
<td>65.06%</td>
</tr>
<tr>
<td>7</td>
<td>Tzanetakis, G. (mono)</td>
<td>64.71%</td>
</tr>
<tr>
<td>8</td>
<td>Peeters, G.</td>
<td>63.90%</td>
</tr>
<tr>
<td>9</td>
<td>Cao &amp; Li (2)</td>
<td>63.39%</td>
</tr>
<tr>
<td>10</td>
<td>Lidy, Rauber, Pertusa &amp; Ponce de León, Iñesta (2)</td>
<td>62.26%</td>
</tr>
<tr>
<td>11</td>
<td>Cao &amp; Li (1)</td>
<td>62.04%</td>
</tr>
<tr>
<td>12</td>
<td>Lidy, Rauber, Pertusa &amp; Ponce de León, Iñesta (3)</td>
<td>60.84%</td>
</tr>
<tr>
<td>13</td>
<td>Lidy, Rauber, Pertusa &amp; Ponce de León, Iñesta (4)</td>
<td>60.46%</td>
</tr>
</tbody>
</table>

Audio Cover Song Identification

<table>
<thead>
<tr>
<th>Rank</th>
<th>Participant</th>
<th>Precis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Serrà, Gómez &amp; Herrera (1)</td>
<td>0.71</td>
</tr>
<tr>
<td>2</td>
<td>Serrà, Gómez &amp; Herrera (2)</td>
<td>0.61</td>
</tr>
<tr>
<td>3</td>
<td>Egorov &amp; Linetsky (2)</td>
<td>0.56</td>
</tr>
<tr>
<td>5</td>
<td>Egorov &amp; Linetsky (1)</td>
<td>0.51</td>
</tr>
<tr>
<td>6</td>
<td>Cao &amp; Li (1)</td>
<td>0.36</td>
</tr>
<tr>
<td>7</td>
<td>Cao &amp; Li (2)</td>
<td>0.36</td>
</tr>
<tr>
<td>8</td>
<td>Jensen, Christensen &amp; Jensen</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Audio Tag Classification

<table>
<thead>
<tr>
<th>Rank</th>
<th>Participant</th>
<th>Avg. F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Barrington, Turnbull &amp; Lanckriet</td>
<td>0.28</td>
</tr>
<tr>
<td>2</td>
<td>Mandel &amp; Ellis (2)</td>
<td>0.26</td>
</tr>
<tr>
<td>3</td>
<td>Mandel &amp; Ellis (3)</td>
<td>0.26</td>
</tr>
<tr>
<td>4</td>
<td>Mandel &amp; Ellis (1)</td>
<td>0.24</td>
</tr>
<tr>
<td>5</td>
<td>Bertin-Mahieux, Eck, Bengio &amp; Lamere (T/L boosting)</td>
<td>0.19</td>
</tr>
<tr>
<td>6</td>
<td>Bertin-Mahieux, Bengio &amp; Eck (nnet)</td>
<td>0.15</td>
</tr>
<tr>
<td>8</td>
<td>Bertin-Mahieux, Bengio &amp; Eck (knn)</td>
<td>0.06</td>
</tr>
<tr>
<td>9</td>
<td>Trohidis, Tsoumakas, Kalliris &amp; Vlahavas</td>
<td>0.04</td>
</tr>
<tr>
<td>10</td>
<td>Peeters, G. (1)</td>
<td>0.03</td>
</tr>
<tr>
<td>11</td>
<td>Peeters, G. (2)</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Fresh from the oven (2009): 73.3% for Audio Genre Classification (GenreMixe)
Datasets

- Several datasets some open and some closed
- Magnatagatune
  - http://tagatune.org/Magnatagatune.html
  - 25000 tracks
  - Tags collected and verified by TagaTune game with a purpose
Audio MIR Pipeline

Hearing Representation

Understanding Analysis

Acting Interaction

Signal Processing
Feature Extraction

Machine Learning
Analysis and Retrieval

Human Computer Interaction/Tools
Tools

- Applications
  - Sonic-Visualizer, BeatRoot

- Frameworks
  - Marsyas, CLAM, MIR Toolbox (matlab)

- Collections of Code
  - Dan Ellis MATLAB examples

- Web-APIs
  - Echonest, LastFM
BeatRoot
from pyechonest import artist

alist = artist.search_artists('weezer')
if (len(alist) > 0):
    print 'Artists similar to', alist[0].name
    for sim in alist[0].similar():
        print " ", sim.name
else:
    print "uh-oh, can't find weezer"
Software framework for audio analysis, synthesis and retrieval

- Efficient and extensible framework design
- Specific emphasis on Music Information Retrieval (MIR)
- C++, OOP
- Multiplatform (Linux, MS Windows®, MacOSX®, ...)

Copyright 2009 G.Tzanetakis
Interaction Outline

- Motivation
- Content & Context Aware UIs
  - Editors
  - Displays
- Query UIs
  - Audio-based
  - Midi-based
Content and Context

- Content ~ file
  - Genre, male voice, high frequency
- Context ~ file and collection
  - Similarity
  - Slow - fast
- Multiple visualizations
  - Same content, different context
Traditional Audio UI

Music production and recording

Waveform and Spectrogram Displays

Cut, paste, effects, etc

Limited content no context

CoolEdit
Frequency to color

Female-Male  Bass-Snare

Low frequencies darker  High frequencies lighter

Only content-sensitive
Timbregrams

Tzanetakis & Cook DAFX00, ICAD01

Content and context similarity and periodic structure using color

Principal Component Analysis
Enhanced Audio Editor Demo
Sonic Browser

Sonic Browser (Univ. Limerick)
Fernstrom & Brazil ICAD 01

Direct Sonification
Spatial-audio aura
Manual placement
drag-drop or simple attributes
TimbreSpace Browser

2D, 3D

Automatic coloring
Hierarchical zooming
Automatic positioning
Principal Component Analysis for dimensionality reduction

Tzanetakis & Cook DAFX00, I
Islands of Music

Pampalk, ISMIR

Automatic analysis

Feature vectors

Self-Organizing Map (SOM)
New Interfaces for Rhythm-based Retrieval

ISMIR 2005
Experiments

Table 1: “Chick” hit detection results

<table>
<thead>
<tr>
<th>Category</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rnb</td>
<td>0.844</td>
<td>0.878</td>
<td>0.861</td>
</tr>
<tr>
<td>Dnb</td>
<td>0.843</td>
<td>0.891</td>
<td>0.866</td>
</tr>
<tr>
<td>Dub</td>
<td>0.865</td>
<td>0.799</td>
<td>0.831</td>
</tr>
<tr>
<td>Hse</td>
<td>0.975</td>
<td>0.811</td>
<td>0.886</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.882</strong></td>
<td><strong>0.845</strong></td>
<td><strong>0.861</strong></td>
</tr>
<tr>
<td>Dadra</td>
<td>0.567</td>
<td>1.000</td>
<td>0.723</td>
</tr>
<tr>
<td>Rupak</td>
<td>0.662</td>
<td>1.000</td>
<td>0.797</td>
</tr>
<tr>
<td>Jhaptaal</td>
<td>0.713</td>
<td>1.000</td>
<td>0.833</td>
</tr>
<tr>
<td>Tintaal</td>
<td>0.671</td>
<td>0.981</td>
<td>0.727</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.653</strong></td>
<td><strong>0.995</strong></td>
<td><strong>0.787</strong></td>
</tr>
<tr>
<td>Various</td>
<td>0.699</td>
<td>0.554</td>
<td>0.618</td>
</tr>
<tr>
<td>Dance</td>
<td>0.833</td>
<td>0.650</td>
<td>0.730</td>
</tr>
<tr>
<td>Funk</td>
<td>0.804</td>
<td>0.621</td>
<td>0.701</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.779</strong></td>
<td><strong>0.609</strong></td>
<td><strong>0.683</strong></td>
</tr>
</tbody>
</table>

Table 2: “Boom” hit detection results

<table>
<thead>
<tr>
<th>Category</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rnb</td>
<td>0.791</td>
<td>0.956</td>
<td>0.866</td>
</tr>
<tr>
<td>Dnb</td>
<td>0.910</td>
<td>0.914</td>
<td>0.916</td>
</tr>
<tr>
<td>Dub</td>
<td>0.846</td>
<td>0.964</td>
<td>0.904</td>
</tr>
<tr>
<td>Hse</td>
<td>0.967</td>
<td>0.994</td>
<td>0.980</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.879</strong></td>
<td><strong>0.957</strong></td>
<td><strong>0.915</strong></td>
</tr>
<tr>
<td>Dadra</td>
<td>0.933</td>
<td>0.972</td>
<td>0.955</td>
</tr>
<tr>
<td>Rupak</td>
<td>1.000</td>
<td>0.763</td>
<td>0.865</td>
</tr>
<tr>
<td>Jhaptaal</td>
<td>0.947</td>
<td>0.981</td>
<td>0.960</td>
</tr>
<tr>
<td>Tintaal</td>
<td>0.843</td>
<td>0.965</td>
<td>0.900</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.931</strong></td>
<td><strong>0.920</strong></td>
<td><strong>0.920</strong></td>
</tr>
<tr>
<td>Various</td>
<td>0.745</td>
<td>0.803</td>
<td>0.772</td>
</tr>
<tr>
<td>Dance</td>
<td>0.823</td>
<td>0.864</td>
<td>0.746</td>
</tr>
<tr>
<td>Funk</td>
<td>0.863</td>
<td>0.820</td>
<td>0.842</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.810</strong></td>
<td><strong>0.829</strong></td>
<td><strong>0.815</strong></td>
</tr>
</tbody>
</table>
Musicream

ISMIR 2005

http://staff.aist.go.jp/m.goto/Musicream/
AudioScapes

http://audioscapes.snes
All the right buzzwords

- Personalized, collaborative, active, content and context aware music exploration
- Audio-Feature Extraction
- Self-organizing maps (SOM) ala Pampalk
- Twists
  - Tabletop interface, continuous playback, personalization, iTunes library import
- Self-Organizing Tag Clouds
Self-Organizing Tag Cloud

Positioning determined by weighted average of song position that have the tag on the self-organized map.

Coloring based on tags also
MarGrid

Available as part of Marsyas
Controllers
Assistive Music Browsing

Dan 19 years old cerebral palsy also blind users

Customized interfaces for each person
Computational Musicology -
recitation research

- Transition from oral to written transmission

- Study how diverse recitation traditions having their origin in primarily non-notated melodies, later became codified

- Hungarian siratok, torah cantillation, koran recitation, 10th century St. Gallen plainchant
Pitch Contour Extraction
Histogram-based contour abstraction
Dynamic-Time Warping

(a) F0 Contour of 11 Pashtas
(b) F0 Contour of 42 Pashtas
(c) F0 Contour of 18 Sof Pashtas
(d) F0 Contour of 11 Pashtas Doubled

(a) DTW of 11 Pashtas vs 11 Pashtas
(b) DTW of 11 Pashtas vs 42 Pashtas
(c) DTW of 11 Pashtas vs 18 Sof Pashtas
(d) DTW of 11 Pashtas vs 11 Pashtas
Cantillion

http://cantillion.sness.net
## Retrieval

<table>
<thead>
<tr>
<th>Gesture (Hungary)</th>
<th>Average Precision (Hungary)</th>
<th>Gesture (Morocco)</th>
<th>Average Precision (Morocco)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tipha</td>
<td>0.662</td>
<td>katon</td>
<td>0.453</td>
</tr>
<tr>
<td>pashta</td>
<td>0.647</td>
<td>mapah</td>
<td>0.347</td>
</tr>
<tr>
<td>mapah</td>
<td>0.641</td>
<td>tipha</td>
<td>0.303</td>
</tr>
<tr>
<td>katon</td>
<td>0.604</td>
<td>sofpasuq</td>
<td>0.285</td>
</tr>
<tr>
<td>etnachta</td>
<td>0.601</td>
<td>pashta</td>
<td>0.242</td>
</tr>
<tr>
<td>sofpasuq</td>
<td>0.591</td>
<td>merha</td>
<td>0.251</td>
</tr>
<tr>
<td>merha</td>
<td>0.537</td>
<td>etnachta</td>
<td>0.150</td>
</tr>
<tr>
<td>revia</td>
<td>0.372</td>
<td>zakef</td>
<td>0.125</td>
</tr>
<tr>
<td>zakef</td>
<td>0.201</td>
<td>revia</td>
<td>0.091</td>
</tr>
<tr>
<td>kada</td>
<td>0.200</td>
<td>kada</td>
<td>0.043</td>
</tr>
</tbody>
</table>
Retrieval at different levels of abstraction
Sensors and music performance
Digitizing North Indian Performance: Preservation and Extension

Ajay Kapur
E-Sitar: adding sensors to a traditional acoustic instrument
E-Sitar Fret Detection

- Network of resistors connecting each fret in series

- When finger depresses string to touch fret, circuit complete and unique voltage is read by the A/D converter.
Mahadevibot
## Results

<table>
<thead>
<tr>
<th>Signal</th>
<th>Tempo (BPM)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80</td>
<td>100</td>
<td>120</td>
<td>140</td>
</tr>
<tr>
<td>Audio</td>
<td>46%</td>
<td>85%</td>
<td>86%</td>
<td>80%</td>
</tr>
<tr>
<td>Fret</td>
<td>27%</td>
<td>27%</td>
<td>57%</td>
<td>56%</td>
</tr>
<tr>
<td>Thumb</td>
<td>35%</td>
<td>62%</td>
<td>75%</td>
<td>65%</td>
</tr>
<tr>
<td>WISP</td>
<td>50%</td>
<td>91%</td>
<td>69%</td>
<td>53%</td>
</tr>
</tbody>
</table>

**LATE FUSION:**

<table>
<thead>
<tr>
<th>Signal</th>
<th>Tempo (BPM)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio/WISP/Thumb/Fret</td>
<td>45%</td>
<td>83%</td>
<td>89%</td>
<td>84%</td>
</tr>
<tr>
<td>Audio/WISP/Thumb</td>
<td>55%</td>
<td>88%</td>
<td>90%</td>
<td>82%</td>
</tr>
<tr>
<td>Audio/ WISP</td>
<td>58%</td>
<td>88%</td>
<td>89%</td>
<td>72%</td>
</tr>
<tr>
<td>Audio/Thumb</td>
<td>57%</td>
<td>88%</td>
<td>90%</td>
<td>80%</td>
</tr>
<tr>
<td>WISP/Thumb</td>
<td>47%</td>
<td>95%</td>
<td>78%</td>
<td>69%</td>
</tr>
</tbody>
</table>
Learning patterns

\[ \text{Audio Input } f(t) \rightarrow \text{Discrete Wavelet Transform} \rightarrow 01010101 \]
Acceleration

![Graph showing acceleration over time](image-url)
Integration
Final thoughts

- MIR is not just:
  - timbral statistics
  - recommendation engines for pop music
  - a passive tool
  - retrieval
- MIR has the potential and will transform the way we create, distribute and consume music
THE END

- Perry Cook, Robert Gjerdingen, Ken Steiglitz
- Malcolm Slaney, Julius Smith, Richard Duda
- Georg Essl, John Forsyth
- Andreye Ermolinskiy, Doug Turnbull, George Tourtellot, Corrie Elder, Ajay Kapur, Stuart Bray, Adam Tindale, Antonin Stefannuti
- ISMIR, WASPAA, ICMC, DAFX, ICASSP, ACMMM