Extracting and Using Music Audio Information

Dan Ellis

Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Engineering, Columbia University, NY USA

http://labrosa.ee.columbia.edu/

1. Motivation: Music Collections
2. Music Information
3. Music Similarity
4. Music Structure Discovery
LabROSA Overview

Information Extraction

Music

Environment

Recognition

Separation

Retrieval

Speech

Machine Learning

Signal Processing
1. Managing Music Collections

- A lot of music data available
  - e.g. 60G of MP3 ≈ 1000 hr of audio, 15k tracks

- Management challenge
  - how can computers help?

- Application scenarios
  - personal music collection
  - discovering new music
  - “music placement”
Learning from Music

• What can we infer from 1000 h of music?
  ○ common patterns
    sounds, melodies, chords, form
  ○ what is and what isn’t music

• Data driven musicology?

• Applications
  ○ modeling/description/coding
  ○ computer generated music
  ○ curiosity...
Music audio information - Ellis

The Big Picture

.. so far

Music audio

- Tempo and beat
- Key and chords
- Melody and notes
- Low-level features

Classification and Similarity

Music Structure Discovery

browsing discovery production

modeling generation curiosity
2. Music Information

- How to represent music audio?
- **Audio features**
  - spectrogram, MFCCs, bases
- **Musical elements**
  - notes, beats, chords, phrases
  - requires transcription
- Or something inbetween?
  - optimized for a certain task?
Transcription as Classification

• Exchange signal models for data
  ○ transcription as pure classification problem:

  **Training data and features:**
  • MIDI, multi-track recordings, playback piano, & resampled audio (less than 28 mins of train audio).
  • Normalized magnitude STFT.

  **Classification:**
  • N-binary SVMs (one for ea. note).
  • Independent frame-level classification on 10 ms grid.
  • Dist. to class bndy as posterior.

  **Temporal Smoothing:**
  • Two state (on/off) independent HMM for ea. note. Parameters learned from training data.
  • Find Viterbi sequence for ea. note.
Polyphonic Transcription

- Real music excerpts + ground truth

Frame-level transcription
Estimate the fundamental frequency of all notes present on a 10 ms grid

Note-level transcription
Group frame-level predictions into note-level transcriptions by estimating onset/offset

Precision Recall Acc Etot Esubs Emiss Efa

Precision Recall Ave. F-measure Ave. Overlap
Beat Tracking

• Goal: One feature vector per ‘beat’ (tatum)
  ○ for tempo normalization, efficiency
• “Onset Strength Envelope”
  ○ \( \text{sum}_t(\max(0, \text{diff}_t(\log |X(t,f)|))) \)

• Autocorr. + window → global tempo estimate

168.5 BPM
Beat Tracking

- **Dynamic Programming** finds beat times \( \{t_i\} \)
  - optimizes \( \sum_i O(t_i) + \alpha \sum_i W((t_{i+1} - t_i - \tau_p)/\beta) \)
  - where \( O(t) \) is onset strength envelope (local score)
    \( W(t) \) is a log-Gaussian window (transition cost)
  - \( \tau_p \) is the default beat period per measured tempo
  - incrementally find best predecessor at every time
  - backtrace from largest final score to get beats

\[
C^*(t) = \gamma O(t) + (1-\gamma)\max_{\tau}\left\{ W(\frac{\tau - \tau_p}{\beta})C^*(\tau) \right\}
\]

\[
P(t) = \arg\max_{\tau}\left\{ W(\frac{\tau - \tau_p}{\beta})C^*(\tau) \right\}
\]
Beat Tracking

- DP will **bridge gaps** (non-causal)
  - there is always a best path ...

2nd place in MIREX 2006 Beat Tracking
- compared to McKinney & Moelants human data
Chroma Features

• Chroma features convert spectral energy into musical weights in a canonical octave
  • i.e. 12 semitone bins

• Can resynthesize as “Shepard Tones”
  • all octaves at once
Key Estimation

- Covariance of chroma reflects **key**
- Normalize by **transposing** for best fit

- single Gaussian model of one piece
- find ML rotation of other pieces
- model all transposed pieces
- iterate until convergence
Chord Transcription

• “Real Books” give chord **transcriptions**
  ○ but no exact timing
  ○ .. just like speech transcripts

• Use **EM** to simultaneously
  learn and align chord models

# The Beatles - A Hard Day's Night
Bm Em Bm G Em C D G Cadd9 G F6 G Cadd9 G F6 G C D G C9 G
F6 G C D G C9 G D
G C7 G F6 G C7 G F6 G C D G C9 G Bm Em Bm
G Em C D
C9 G Cadd9 Fadd9

\[
\begin{align*}
\text{Model inventory} & \quad \text{Labelled training data} \\
\text{Initialization} & \quad \text{Uniform} \\
\text{parameters} & \quad \text{initialization} \\
\Theta_{init} & \quad \text{alignments} \\
\end{align*}
\]

Repeat until convergence

- **E-step:** probabilities of unknowns
  \[ p(q^i_n|X^N,\Theta_{old}) \]
- **M-step:** maximize via parameters
  \[ \Theta : \max E[\log p(X,Q | \Theta)] \]
Chord Transcription

Frame-level Accuracy

<table>
<thead>
<tr>
<th>Feature</th>
<th>Recog.</th>
<th>Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>8.7%</td>
<td>22.0%</td>
</tr>
<tr>
<td>PCP_ROT</td>
<td>21.7%</td>
<td>76.0%</td>
</tr>
</tbody>
</table>

**MFCCs are poor**
(can overtrain)  
**PCPs better**
(ROT helps generalization)

• Needed more training data...

Music Audio Information - Ellis  
2007-11-02  p. 15/42
3. Music Similarity

• The **most central problem**...
  ○ motivates extracting musical information
  ○ supports real applications (playlists, discovery)

• **But do we need** content-based similarity?
  ○ compete with collaborative filtering
  ○ compete with fingerprinting + metadata

• Maybe ... for the **Future of Music**
  ○ connect listeners directly to musicians
Discriminative Classification

- Classification as a **proxy** for similarity
- Distribution models...

• vs. SVM

---

Mandel & Ellis '05
Segment-Level Features

- Statistics of spectra and envelope define a point in feature space
  - for SVM classification, or Euclidean similarity...

![Diagram showing Segment-Level Features]

**Temporal Pipeline**
- Magnitude Bands
- Envelope Cepstrum
- DC on
- Rows
- Low Freq Modulation
- Stats
- Stack

**Spectral Pipeline**
- Mel Spectrum
- MFCCs
- DC on
- Cola
- Covariance
- Stack

**Final Features**
- Combined Features
- Normalize

---

Music Audio Information - Ellis

2007-11-02   p. 18/42
MIREX’07 Results

• One system for similarity and classification

Audio Music Similarity

Audio Classification

PS = Pohle, Schnitzer; GT = George Tzane-takis; LB = Barrington, Turnbull, Torres, Lanckriet; CB = Christoph Bastuck; TL = Lidy, Rauber, Pertusa, Iñesta; ME = Mandel, Ellis; BK = Bosteels, Kerre; PC = Paradzinets, Chen

IM = IMIRSEL M2K; ME = Mandel, Ellis; TL = Lidy, Rauber, Pertusa, Iñesta; GT = George Tzane-takis; KL = Kyogu Lee; CL = Laurier, Herrera; GH = Guaus, Herrera
Active-Learning Playlists

• SVMs are well suited to “active learning”
  ○ solicit labels on items closest to current boundary

• Automatic player with “skip”
  = Ground truth data collection
  ○ active-SVM automatic playlist generation
Cover Song Detection

• “Cover Songs” = reinterpretation of a piece
  ● different instrumentation, character
  ● no match with “timbral” features

Let It Be - The Beatles

Let It Be - Nick Cave

• Need a different representation!
  ● beat-synchronous chroma features
Beat-Synchronous Chroma Features

- Beat + chroma features / 30ms frames
  → average chroma within each beat
- compact; sufficient?
Matching: Global Correlation

- Cross-correlate *entire* beat-chroma matrices
  - ... at all possible *transpositions*
  - implicit *combination* of match quality and duration

- One good matching fragment is sufficient...?
MIREX 06 Results

- Cover song contest
  - 30 songs x 11 versions of each (!)
  - (data has not been disclosed)
  - # true covers in top 10
  - 8 systems compared (4 cover song + 4 similarity)

- Found 761/3300 = 23% recall
  - next best: 11%
  - guess: 3%
Cross-Correlation Similarity

- Use cover-song approach to find similarity
  - e.g. similar note/instrumentation sequence
  - may sound very similar to judges

- Numerous variants
  - try on chroma (melody/harmony) and MFCCs (timbre)
  - try full search (xcorr) or landmarks (indexable)
  - compare to random, segment-level stats

- Evaluate by subjective tests
  - modeled after MIREX similarity
Cross-Correlation Similarity

• Human web-based judgments
  ○ binary judgments for speed
  ○ 6 users x 30 queries x 10 candidate returns

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Similar count</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Xcorr, chroma</td>
<td>48/180 = 27%</td>
</tr>
<tr>
<td>(2) Xcorr, MFCC</td>
<td>48/180 = 27%</td>
</tr>
<tr>
<td>(3) Xcorr, combo</td>
<td>55/180 = 31%</td>
</tr>
<tr>
<td>(4) Xcorr, combo + tempo</td>
<td>34/180 = 19%</td>
</tr>
<tr>
<td>(5) Xcorr, combo at boundary</td>
<td>49/180 = 27%</td>
</tr>
<tr>
<td>(6) Baseline, MFCC</td>
<td>81/180 = 45%</td>
</tr>
<tr>
<td>(7) Baseline, rhythmic</td>
<td>49/180 = 27%</td>
</tr>
<tr>
<td>(8) Baseline, combo</td>
<td>88/180 = 49%</td>
</tr>
<tr>
<td>Random choice 1</td>
<td>22/180 = 12%</td>
</tr>
<tr>
<td>Random choice 2</td>
<td>28/180 = 16%</td>
</tr>
</tbody>
</table>

• Cross-correlation inferior to baseline...
  ○ ... but is getting somewhere, even with ‘landmark’
### Cross-Correlation Similarity

- Results are not overwhelming
  - ... but database is only a few thousand clips

---

<table>
<thead>
<tr>
<th>Song 1</th>
<th>Song 2</th>
<th>Similarity</th>
<th>Song 3</th>
<th>Similarity</th>
<th>Song 4</th>
<th>Similarity</th>
<th>Song 5</th>
<th>Similarity</th>
<th>Song 6</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Too Much</td>
<td>Too Much</td>
<td>-0.00</td>
<td>Erotica</td>
<td>Madonna</td>
<td>-0.00</td>
<td></td>
<td>Waiting</td>
<td>Madonna</td>
<td>Where Life Begins</td>
<td>Madonna</td>
</tr>
<tr>
<td>Hey Nineteen</td>
<td>Hey Nineteen</td>
<td>-0.00</td>
<td>Where Life Begins</td>
<td>Madonna</td>
<td>-0.00</td>
<td></td>
<td>Eroica</td>
<td>Madonna</td>
<td>Don't Tell Me</td>
<td>Madonna</td>
</tr>
<tr>
<td>Little 15</td>
<td>Little 15</td>
<td>-0.00</td>
<td>Don't Tell Me</td>
<td>Madonna</td>
<td>-0.00</td>
<td></td>
<td>Lolita</td>
<td>Suzanne Vega</td>
<td>Where Life Begins</td>
<td>Madonna</td>
</tr>
<tr>
<td>The Same Deep Water As You</td>
<td>The Same Deep Water As You</td>
<td>-0.00</td>
<td>Scarlet</td>
<td>U2</td>
<td>-0.00</td>
<td></td>
<td>Breathing in fumes</td>
<td>Depeche Mode</td>
<td>Where Life Begins</td>
<td>Madonna</td>
</tr>
<tr>
<td>Scarlet</td>
<td>Scarlet</td>
<td>-0.00</td>
<td>Breathing in fumes</td>
<td>U2</td>
<td>-0.00</td>
<td></td>
<td>Keep It Together</td>
<td>Madonna</td>
<td>Eroica</td>
<td>Madonna</td>
</tr>
<tr>
<td>Flying</td>
<td>Flying</td>
<td>-0.00</td>
<td>Keep It Together</td>
<td>Madonna</td>
<td>-0.00</td>
<td></td>
<td>Eroica</td>
<td>Madonna</td>
<td>I Wish U Heaven</td>
<td>Prince</td>
</tr>
<tr>
<td>Breathing in fumes</td>
<td>Breathing in fumes</td>
<td>-0.00</td>
<td>Eroica</td>
<td>Madonna</td>
<td>-0.00</td>
<td></td>
<td>I Wish U Heaven</td>
<td>Prince</td>
<td>Shiver And I'm Sorry</td>
<td>Prince</td>
</tr>
<tr>
<td>Bad Moon Rising</td>
<td>Bad Moon Rising</td>
<td>-0.00</td>
<td>Let's Pretend We're Married</td>
<td>Creedence Clearwater Revival</td>
<td>-0.00</td>
<td>Don't Look Now</td>
<td>Creedence Clearwater Revival</td>
<td>-0.00</td>
<td>Cry Baby</td>
<td>Madonna</td>
</tr>
<tr>
<td>Cry Baby</td>
<td></td>
<td></td>
<td>Fashion Victim</td>
<td>Green Day</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

*Note: The table above is a screenshot of a web page showing cross-correlation similarity results for a music database.*
• Acoustic features describe each song
  • .. but from a **signal**, not a **perceptual**, perspective
  • .. and not the **differences** between songs

• **Use genre classifiers to define new space**
  • prototype genres are “anchors”

---

**“Anchor Space”**

Berenzweig & Ellis ‘03
“Anchor Space”

- Frame-by-frame high-level categorizations
- properties in distributions? dynamics?

Cepstral Features

Anchor Space Features

Graphs and plots indicating comparisons of cepstral coefficients and anchor space features.
### ‘Playola’ Similarity Browser

#### Get Selections:
- 20 songs
- recently heard

#### Artist:
- Beatles

### Song Title | Artist | Time
--- | --- | ---
Baby You’re a Rich Man | Beatles | 3:03
Blue Jay Way | Beatles | 3:56
Penny Lane | Beatles | 3:03
Magical Mystery Tour | Beatles | 2:51
The Fool on the Hill | Beatles | 3:00
I Am the Walrus | Beatles | 4:37
Flying | Beatles | 2:17
Your Mother Should Know | Beatles | 2:29
Strawberry Fields Forever | Beatles | 4:10

#### Album: Magical Mystery Tour
- 9 songs

#### Album: Yellow Submarine
- 8 songs

#### Music-Space Browser

<table>
<thead>
<tr>
<th>Feature</th>
<th>Less</th>
<th>More</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative Rock</td>
<td></td>
<td></td>
</tr>
<tr>
<td>College Rock</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Country</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dance Rock</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metal Punk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Wave</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rap</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R&amp;B Soul</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singer-Songwriter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soft Rock</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trad Rock</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hi-Fi</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Similar Songs:

#### Song Title | Artist | Distance | Good Match?
--- | --- | --- | ---
Let It Be | The Beatles | 0.00 | Yes
Double Hockey Sticks | Adam The Gimbel | 0.06 | Yes
Light In Your Eyes | Blessed Union of Souls | 0.06 | Yes
Mori | Tranzas | 0.07 | Yes
Ground-truth data

• Hard to evaluate Playola’s ‘accuracy’
  - user tests...
  - ground truth?

• “Musicseer” online survey/game:
  - ran for 9 months in 2002
  - > 1,000 users,
    - > 20k judgments
“Semantic Bases”

• Describe segment in human-relevant terms
  ○ e.g. anchor space, but more so

• Need ground truth...
  ○ what words to people use?

• MajorMiner
  game:
  ○ 400 users
  ○ 7500 unique tags
  ○ 70,000 taggings
  ○ 2200 10-sec clips used

• Train classifiers...
3. Music Structure Discovery

• Use the many examples to map out the “manifold” of music audio
  • ... and hence define the subset that is music

• Problems
  • alignment/registration of data
  • factoring & abstraction
  • separating parts?
Eigenrhythms: Drum Pattern Space

• Pop songs built on repeating “drum loop”
  • variations on a few bass, snare, hi-hat patterns

• Eigen-analysis (or ...) to capture variations?
  • by analyzing lots of (MIDI) data, or from audio

• Applications
  • music categorization
  • “beat box” synthesis
  • insight
Aligning the Data

• Need to **align** patterns prior to modeling...

**tempo** (stretch): by inferring BPM & normalizing

**downbeat** (shift): correlate against ‘mean’ template
**Eigenrhythms (PCA)**

- Need 20+ Eigenvectors for good coverage of 100 training patterns (1200 dims)
- Eigenrhythms both **add** and **subtract**
**Posirhythms (NMF)**

- Nonnegative: only *adds* beat-weight
- Capturing some structure

---

Music Audio Information - Ellis

2007-11-02 p. 37/42
Eigenrhythm BeatBox

- Resynthesize rhythms from eigen-space

![Diagram of Eigenrhythm BeatBox](image-url)
Melody Clustering

• **Goal:** Find ‘fragments’ that recur in melodies
  • .. across large music database
  • .. trade data for model sophistication

- **Data sources**
  • pitch tracker, or MIDI training data

- **Melody fragment representation**
  • $\text{DCT}(1:20)$ - removes average, smoothes detail
**Melody Clustering**

- Clusters match underlying contour:
  - Some interesting matches:
    - e.g. Pink + Nsync
Beat-Chroma Fragment Codebook

• Idea: Find the very popular music fragments
  ○ e.g. perfect cadence, rising melody, ...?

• Clustering a large enough database should reveal these
  ○ but: registration of phrase boundaries, transposition

• Need to deal with really large datasets
  ○ e.g. 100k+ tracks, multiple landmarks in each
  ○ but: Locality Sensitive Hashing can help - quickly finds ‘most’ points in a certain radius

• Experiments in progress...
Conclusions

• Lots of data
  + noisy transcription
  + weak clustering
  ⇒ musical insights?