Intelligent Audio Systems: A review of the foundations and applications of semantic audio analysis and music information retrieval
Administration

- Daily schedule
- Today’s schedule
- Introductions
  - A little about yourself
  - Your area of interest, background with DSP, coding, python, and any specific items of interest that you’d like to see covered.
  - And one interesting fact
Why MIR?

Organize Sound
- Intelligent metadata generation for media collections.
- Discovery and monetization of massive media archives.

Search Sound
- “Find me something that sounds like this” search engine for audio content (songs, real-world sounds, music loops, speech)
- Music-similarity search

Understand Sound
- Revolutionary workflows in consumer products through machine hearing. Automatic control of software, signal processing and mobile devices
Problems

1. Computers are deaf.
2. Content is overwhelming and unsearchable.

The state of the art is “search by text”

Text metadata only
(Name, Creator, User Tags)
Audio Visualization
Simplifying User Experience

Detection algorithms help musicians achieve their desired results quicker.
Harmony / Chord / Key Estimation
Mixed In Key

[Image of the Mixed In Key software interface]

- **Pryda**
  - Name: Layers (Original Mix)
  - Key Result: 7A
  - Tempo: 126
  - Energy: 8
  - Status: Completed

- **Axwell**
  - Name: Center of the Universe (Dyro Remix)
  - Key Result: 11B
  - Tempo: 129
  - Energy: 7
  - Status: Completed

- **DubVision**
  - Name: Redux (Original Mix)
  - Key Result: 10A
  - Tempo: 127
  - Energy: 8
  - Status: Completed

- **Hard Rock Sofa**
  - Name: Rasputin
  - Key Result: 4A
  - Tempo: 128
  - Energy: 8
  - Status: Completed

4 songs
BEFORE
Drum Transcription

Example 1

Example 2
Pitch Tracking / Query by Humming

QBH  - SoundHound

Madonna - https://soundcloud.com/madonnagr/secret-acapella-lead-vox

https://www.youtube.com/watch?v=QnRt_cWdV8c
https://www.youtube.com/watch?v=IArxakPsPE0
Discovery

Find me **sound effects** that sound like this?

Audio Search

*similarity search*

find similar
Sales and Marketing

BreakTweaker Launch

- Revenue through April: $440k
- Target: $379k
- 1,800 units sold
- 420 units / 22% new to iZotope

Compared to prior launches (first 90 days): 17
Music recommendation

<table>
<thead>
<tr>
<th>Name</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>M0112708_06 Overkill.wav</td>
<td>2:12</td>
</tr>
<tr>
<td>M0112707_05 Blood Lust.wav</td>
<td>2:12</td>
</tr>
<tr>
<td>M0112705_04 Napalm Blitz.wav</td>
<td>2:12</td>
</tr>
<tr>
<td>M0112711_08 Demolition Barbie.wav</td>
<td>2:23</td>
</tr>
<tr>
<td>M0112702_01 Axephetamine.wav</td>
<td>2:34</td>
</tr>
<tr>
<td>M0112703_02 Dimebag Damage.wav</td>
<td>2:25</td>
</tr>
<tr>
<td>Mermaid in Japan</td>
<td>5:06</td>
</tr>
<tr>
<td>M0112953_10 Hallowed By Thy Flame.wav</td>
<td>2:46</td>
</tr>
<tr>
<td>M0112713_09 Headlong Heracy.wav</td>
<td>3:01</td>
</tr>
<tr>
<td>M0112716_11 No Holds Barred.wav</td>
<td>2:37</td>
</tr>
<tr>
<td>M0112717_12 Billy Whizz.wav</td>
<td>2:40</td>
</tr>
<tr>
<td>M0112996_01 The Beast.wav</td>
<td>2:27</td>
</tr>
<tr>
<td>M0112704_03 Terrorize.wav</td>
<td>2:35</td>
</tr>
<tr>
<td>M0113007_07 Speed.wav</td>
<td>2:28</td>
</tr>
<tr>
<td>Bad Attraction – Earjamm Mix (Hipcola)</td>
<td>5:35</td>
</tr>
<tr>
<td>Show Me Fear</td>
<td>3:59</td>
</tr>
<tr>
<td>M0113004_05 Slow Death.wav</td>
<td>2:04</td>
</tr>
<tr>
<td>I Am</td>
<td>4:59</td>
</tr>
<tr>
<td>M0112544_15 Fastball Special.wav</td>
<td>3:51</td>
</tr>
<tr>
<td>Whispers and Knives (Yongan)</td>
<td>5:45</td>
</tr>
</tbody>
</table>
Rhythmic Similarity

Create by Example

Here’s my seed riff.

Find me a Bass loop that grooves with that.

Find me a Drum loop that grooves with that.

Play them all together.
Dynamic Media

James Brown – The Payback

Original

Groove

Remixes

Graphic: Clint Bajakian
Tagging

“distorted, guitar”
Tagging
Tagging

Machine learning
Content analysis

Audiol.wav
1011000111011000

Male Speech
Female Speech
Baby Cry
Music (Rock)
Applause

Sound signature
Smart chapter markers

Sound signature
Genre; Tags
Tempo; Beats; Groove
Intelligent FFWD
Smarter DSP
ROAD TRAFFIC NOISE
DeBreath

http://youtu.be/DZJfURZ12AM?t=7m15s
Auto-mixing

Before  After

For example: https://www.landr.com/#/
Why MIR?

- content-based querying and retrieval, indexing (tagging, similarity)
  - fingerprinting and digital rights management
  - music recommendation and playlist generation
- music transcription and annotation
  - score following and audio alignment
- automatic classification
- rhythm, beat, tempo, and form
- harmony, chords, and tonality
- timbre, instrumentation
- genre
- emotion, style, and mood analysis
  - music summarization
BASIC SYSTEM OVERVIEW
Basic system overview

Segmentation
(Frames, Onsets, Beats, Bars, Chord Changes, etc)
Basic system overview

Segmentation
  (Frames, Onsets, Beats, Bars, Chord Changes, etc)

Feature Extraction
  (Time-based, spectral energy, MFCC, etc)
Basic system overview

Segmentation
(Frames, Onsets, Beats, Bars, Chord Changes, etc)

Feature Extraction
(Time-based, spectral energy, MFCC, etc)

Analysis / Decision Making
(Classification, Clustering, etc)
TIMING AND SEGMENTATION
Timing and Segmentation

• Slicing up by fixed time slices...
  – 1 second, 80 ms, 100 ms, 20-40ms, etc.

• “Frames”
  – Different problems call for different frame lengths
FEATURE EXTRACTION
Timing and Segmentation

• Slicing up by fixed time slices…
  – 1 second, 80 ms, 100 ms, 20-40ms, etc.
• “Frames”
  – Different problems call for different frame lengths
• Onset detection
• Beat detection
  – Beat
  – Measure / Bar / Harmonic changes
• Segments
  – Musically relevant boundaries
  – Separate by some perceptual cue
FEATURE EXTRACTION
ZERO CROSSING RATE

FRAME 1

Zero crossing rate = 9
Frame 2

Zero crossing rate = 423
Features : SimpleLoop.wav

<table>
<thead>
<tr>
<th>Frame</th>
<th>ZCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>423</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>390</td>
</tr>
</tbody>
</table>

Warning: example results only - not actual results from audio analysis...
Heuristic Analysis

• Use basic thresholds or simple decision tree to form rudimentary transcription of kicks and snares.
• Time for more sophistication...
• Stairway / Stairway sorted
Example Feature Vector

<table>
<thead>
<tr>
<th></th>
<th>ZCR</th>
<th>Centroid</th>
<th>Bandwidth</th>
<th>Skew</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>205</td>
<td>982.0780</td>
<td>0.1452</td>
<td>1.3512e+03</td>
</tr>
<tr>
<td>2</td>
<td>150</td>
<td>621.0359</td>
<td>0.1042</td>
<td>296.0815</td>
</tr>
<tr>
<td>3</td>
<td>120.0000</td>
<td>361.6111</td>
<td>0.0607</td>
<td>263.7817</td>
</tr>
<tr>
<td>4</td>
<td>135</td>
<td>809.3978</td>
<td>0.1315</td>
<td>834.4116</td>
</tr>
<tr>
<td>5</td>
<td>220</td>
<td>634.7242</td>
<td>0.0906</td>
<td>274.5483</td>
</tr>
<tr>
<td>6</td>
<td>175</td>
<td>536.3318</td>
<td>0.0837</td>
<td>188.4155</td>
</tr>
<tr>
<td>7</td>
<td>190</td>
<td>567.0412</td>
<td>0.0953</td>
<td>253.0151</td>
</tr>
<tr>
<td>8</td>
<td>135</td>
<td>720.2892</td>
<td>0.1153</td>
<td>333.7646</td>
</tr>
<tr>
<td>9</td>
<td>195.0000</td>
<td>778.5310</td>
<td>0.1407</td>
<td>1.2328e+03</td>
</tr>
<tr>
<td>10</td>
<td>185</td>
<td>514.4315</td>
<td>0.0717</td>
<td>183.0322</td>
</tr>
</tbody>
</table>
ANALYSIS AND DECISION MAKING
INSTANCE-BASED CLASSIFIERS (K-NN)
Training...

TRAINING SET

“kick drum”       “not a kick drum”
k-NN

Explanation...

**Advantages:**
Training is trivial: just store the training samples very simple to implement and use

**Disadvantages**
Classification gets very complex with a lot of training data Must measure distance to all training samples; Euclidean distance becomes problematic in high-dimensional spaces; Can easily be “overfit”

We can improve computation efficiency by storing just the class prototypes.
**k-NN**

Steps:

– Measure distance to all points.
– Take the k closest
– Majority rules. (e.g., if k=5, then take 3 out of 5)

Fig. 2.15. k-nearest neighbours classification of two-dimensional data in the two-class case, with $k = 5$. The new datum $\mathbf{x}$ is represented by a non-filled circle. Elements of the training set $(X, Y)$ are represented with dots (those with label $-1$) and squares (those with label $+1$). The arrow lengths represent the Euclidean distance between $\mathbf{x}$ and its 5 nearest neighbours. Three of them are squares, which makes $\mathbf{x}$ have the label $y = +1$. 
k-NN

• Instance-based learning – training examples are stored directly, rather than estimate model parameters

• Generally choose k being odd to guarantee a majority vote for a class.