An Audio Fingerprinting System for Live Version Identification using Image Processing Techniques

(Dr.?) Zafar Rafii
Northwestern University
EECS department
Acknowledgments

• Work done during internship at Gracenote, a company doing audio and video identification, with co-authors Bob Coover and Jinyu Han.

Context

• You are at a concert
  – You know the artist who is playing
  – You want to know about the song being played
  – You have a smart device (e.g., an iPhone)
Idea

• You can use a music identification system
  – You record an excerpt using your smart device
  – It is processed and compared against a database
  – You get information about the song (e.g., title)
Principle

• Audio fingerprinting systems
  – Transform the audio into a compact fingerprint
  – Compare the query against a database for a match
  – Typically index fingerprints to speed up matching
Limitations

• Does not work with cover versions (e.g., live)
  – Variations in tempo (e.g., faster renditions)
  – Variations in key (e.g., higher pitch)
  – Variations in instrumentations, etc.
Solution

• A novel system that can handle
  – Short excerpt quickly (i.e., less than 10 seconds)
  – Audio degradations (e.g., noise, encoding, etc.)
  – Audio variations (e.g., different tempo, key, etc.)
Approach

- **Fingerprinting stage**
  - Constant Q Transform
  - Adaptive thresholding

- **Matching stage**
  - Hamming similarity
  - Hough Transform
Fingerprinting

- **Constant Q Transform (CQT)**
  - We first transform the audio signal into a time-frequency representation using the CQT.
Fingerprinting

- Constant Q Transform (CQT)
  - The CQT has a log-frequency resolution, matching the notes of the chromatic scale (i.e., C, C#, etc.)

Each note has the frequency of the previous note multiplied by $\frac{12}{\sqrt{2}}$
Fingerprinting

• Constant Q Transform (CQT)
  – Unlike the FT, the CQT is more compact and better adapted to music (vertical shift = pitch shift)

Three notes played at different pitches in the FT-spectrogram (left) and the CQT-spectrogram (right)

https://ccrma.stanford.edu/~gautham/
Fingerprinting

• Adaptive thresholding
   – We transform the CQT-spectrogram into a binary image using an adaptive thresholding method
Fingerprinting

• Adaptive thresholding
  – For each bin in the spectrogram, we assign 1 if the bin is higher than the median of the neighborhood

Neighborhood at 27 seconds and C6 (1.0 kHz)

Neighborhood at 25 seconds and C4 (262 Hz)
Fingerprinting

• Adaptive thresholding
  – We get a fingerprint that reduces the spectrogram into 2 components, of locally low and high energy
Approach

• Fingerprinting stage
  – Constant Q Transform
  – Adaptive thresholding

• Matching stage
  – Hamming similarity
  – Hough Transform
Matching

• Hamming similarity
  – We then compute a similarity matrix between the fingerprints of a query and each of the references
Matching

- Hamming similarity
  - We use the Hamming similarity between all pairs of time frames (= percentage of bins that match)
Matching

- Hamming similarity
  - We compute the similarity matrix for different pitch shifts between the query and the references
Matching

• Hough Transform (HT)
  – We binarize the similarity matrix via a threshold to have pairs of time frames that match (1) or not (0)
Method

• Hough Transform (HT)
  – We use the HT to identify the best alignment between the query and the reference fingerprints

Best alignment: 64% of time frames matching
Matching

- Hough Transform (HT)
  - The HT helps to take into account potential tempo deviations, by trying different angles for a line

![Diagram of matching process](image)

- Angle of 41°: query faster than reference
- Angle of 45°: query as fast as reference
- Angle of 47°: query slower than reference
Evaluation

• References
  – 10 different artists of varied genres
  – 389 full tracks from studio albums
  – Durations from 01’04” to 11’06”

• Queries
  – 87 full tracks from live albums (experiment 1)
  – 87 audio tracks from smart devices (experiment 2)
  – 10 queries per tracks, 6 and 9 second length
# Data set

<table>
<thead>
<tr>
<th>artist</th>
<th>genre</th>
<th>#references</th>
<th>#queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC/DC</td>
<td>hard rock</td>
<td>36</td>
<td>60</td>
</tr>
<tr>
<td>Arcade Fire</td>
<td>indie rock</td>
<td>33</td>
<td>100</td>
</tr>
<tr>
<td>Bonobo</td>
<td>electronic</td>
<td>42</td>
<td>100</td>
</tr>
<tr>
<td>Eagles</td>
<td>rock</td>
<td>32</td>
<td>90</td>
</tr>
<tr>
<td>Foreigner</td>
<td>rock</td>
<td>29</td>
<td>100</td>
</tr>
<tr>
<td>Jefferson Airplane</td>
<td>psychedelic rock</td>
<td>65</td>
<td>40</td>
</tr>
<tr>
<td>Led Zeppelin</td>
<td>rock</td>
<td>40</td>
<td>80</td>
</tr>
<tr>
<td>Phoenix</td>
<td>alternative rock</td>
<td>38</td>
<td>100</td>
</tr>
<tr>
<td>Portishead</td>
<td>electronic</td>
<td>33</td>
<td>100</td>
</tr>
<tr>
<td>Suprême NTM</td>
<td>French hip hop</td>
<td>41</td>
<td>100</td>
</tr>
<tr>
<td><strong>all</strong></td>
<td>-</td>
<td><strong>389</strong></td>
<td><strong>870</strong></td>
</tr>
</tbody>
</table>
### Live albums (9 seconds)

<table>
<thead>
<tr>
<th>Top-k matches</th>
<th>k=1</th>
<th>k=2</th>
<th>k=3</th>
<th>k=4</th>
<th>k=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC/DC</td>
<td>0.92</td>
<td>0.95</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Arcade Fire</td>
<td>0.84</td>
<td>0.92</td>
<td>0.94</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>Bonobo</td>
<td>0.83</td>
<td>0.89</td>
<td>0.92</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td>Eagles</td>
<td>0.93</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Foreigner</td>
<td>0.88</td>
<td>0.93</td>
<td>0.93</td>
<td>0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>Jefferson Airplane</td>
<td>0.60</td>
<td>0.68</td>
<td>0.78</td>
<td>0.78</td>
<td>0.80</td>
</tr>
<tr>
<td>Led Zeppelin</td>
<td>0.74</td>
<td>0.81</td>
<td>0.84</td>
<td>0.85</td>
<td>0.90</td>
</tr>
<tr>
<td>Phoenix</td>
<td>0.88</td>
<td>0.92</td>
<td>0.93</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>Portishead</td>
<td>0.92</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>Suprême NTM</td>
<td>0.87</td>
<td>0.95</td>
<td>0.96</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>all</td>
<td>0.86</td>
<td>0.91</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
</tr>
</tbody>
</table>
## Smart devices (9 seconds)

<table>
<thead>
<tr>
<th>Top-k matches</th>
<th>k=1</th>
<th>k=2</th>
<th>k=3</th>
<th>k=4</th>
<th>k=5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AC/DC</strong></td>
<td>0.70</td>
<td>0.83</td>
<td>0.85</td>
<td>0.87</td>
<td>0.93</td>
</tr>
<tr>
<td><strong>Arcade Fire</strong></td>
<td>0.79</td>
<td><strong>0.86</strong></td>
<td>0.89</td>
<td>0.91</td>
<td>0.93</td>
</tr>
<tr>
<td><strong>Bonobo</strong></td>
<td>0.60</td>
<td>0.75</td>
<td>0.83</td>
<td>0.89</td>
<td>0.93</td>
</tr>
<tr>
<td><strong>Eagles</strong></td>
<td>0.70</td>
<td>0.77</td>
<td>0.88</td>
<td><strong>0.91</strong></td>
<td>0.91</td>
</tr>
<tr>
<td><strong>Foreigner</strong></td>
<td>0.68</td>
<td>0.83</td>
<td>0.86</td>
<td>0.86</td>
<td>0.88</td>
</tr>
<tr>
<td><strong>Jefferson Airplane</strong></td>
<td>0.40</td>
<td>0.53</td>
<td>0.55</td>
<td>0.60</td>
<td>0.63</td>
</tr>
<tr>
<td><strong>Led Zeppelin</strong></td>
<td>0.28</td>
<td>0.39</td>
<td>0.48</td>
<td>0.53</td>
<td>0.54</td>
</tr>
<tr>
<td><strong>Phoenix</strong></td>
<td>0.67</td>
<td>0.76</td>
<td>0.82</td>
<td>0.86</td>
<td>0.87</td>
</tr>
<tr>
<td><strong>Portishead</strong></td>
<td><strong>0.80</strong></td>
<td><strong>0.86</strong></td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td><strong>Suprême NTM</strong></td>
<td>0.30</td>
<td>0.42</td>
<td>0.45</td>
<td>0.51</td>
<td>0.55</td>
</tr>
<tr>
<td><strong>all</strong></td>
<td>0.61</td>
<td>0.71</td>
<td>0.76</td>
<td>0.79</td>
<td>0.81</td>
</tr>
</tbody>
</table>