8. Musical Genre Classification
Goal

• Learn about
  – Task of musical genre classification
  – Feature extraction
  – K-nearest neighbor classifier

Lecture based on:
“Musical Genre Classification of Audio Signals”
By George Tzanetakis and Perry Cook
IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 10, NO. 5, JULY 2002
Motivation

• Provide means
  • to access music by verbal description
  • to access music by example
  • to structure music data bases
Audio Classification

Classical

Country

Rock
Audio Classification Hierarchy

Music
- Classical
- Country
- Disco
- HipHop
- Jazz
- Rock
- Blues
- Reggae
- Pop
- Metal

Speech
- Choir
- Orchestra
- Piano
- String Quartet
- BigBand
- Cool
- Fusion
- Piano
- Quartet
- Swing

Male
Female
Sports
Automatic Musical Genre Classification

• **Issues:**
  • Genre boundaries are
    • Fuzzy
    • Arguable
Classification Procedure

Raw audio → Digitally encode → Extract features

Training

Build classifier models

Decide class

Raw audio → Digitally encode → Extract features

Classification
Extract Features

- Mel-scaled cepstral coefficients (MFCCs)
- Musical surface features
- Rhythm Features
- Others…
Musical surface features

- Represents characteristics of music
  - Texture
  - Timbre
  - Instrumentation
- Statistics over spectral distribution
  - Centroid
  - Rolloff
  - Flux
  - Zero Crossings
  - Low Energy
Calculating Surface Features

Signal → Divide into windows → FFT over window → Calculate feature for window → Calculate mean and std. dev. over windows

\[ \sum \ldots \]
Surface Features: Centroid

- Centroid: Measures spectral brightness
  \[ C_t = \frac{\sum_{f=1}^{N} f M_t[f]}{\sum_{f=1}^{N} M_t[f]} \]
- Measures amount of high frequency components

\[ M[f] = \text{magnitude of FFT at frequency bin } f \text{ over } N \text{ bins} \]
Surface Features

• Rolloff: Spectral Shape

Determine R such that:

\[ \sum_{f=1}^{R} M[f] = 0.85 \times \sum_{f=1}^{N} M[f] \]
Surface Features

- **Flux**: Spectral change

\[ F_t = \| N_t[f] - N_{t-1}[f] \| \]

Where, \( N_t[f] \) is the per frame normalized magnitude
Surface Features

- Zero Crossings:

\[ Z_t = \frac{1}{2} \sum_{n=t-N/2}^{t+N/2} |\text{sign}(x[n]) - \text{sign}(x[n-1])| \]

- \text{sign}: 1 is argument is positive, 0 else
- Provides a measure of noisiness of the signal
Rhythm Features

Discrete Wavelet Transform Octave Frequency Bands

Full Wave Rectification

Low Pass Filtering

Downsampling

Mean Removal

Envelope Extraction

Envelope Extraction

Envelope Extraction

Autocorrelation

Multiple Peak Picking

Beat Histogram
Examples of Beat Histograms

CLASSICAL

ROCK

JAZZ

HIP-HOP
Experimental Setup

- Songs collected from radio, CDs and Web
- 50 samples for each class, 30 sec. Long
- 15 genres
- Features: MFCC, surface and rhythm features
- Classifiers: GMM and kNN
- 10 fold cross validation
## Classification Results

<table>
<thead>
<tr>
<th></th>
<th>Genres (10)</th>
<th>Classical (4)</th>
<th>Jazz (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>10</td>
<td>25</td>
<td>16</td>
</tr>
<tr>
<td>RT GS</td>
<td>44 ± 2</td>
<td>61 ± 3</td>
<td>53 ± 4</td>
</tr>
<tr>
<td>GS</td>
<td>59 ± 4</td>
<td>77 ± 6</td>
<td>61 ± 8</td>
</tr>
<tr>
<td>GMM (2)</td>
<td>60 ± 4</td>
<td>81 ± 5</td>
<td>66 ± 7</td>
</tr>
<tr>
<td>GMM (3)</td>
<td>61 ± 4</td>
<td>88 ± 4</td>
<td>68 ± 7</td>
</tr>
<tr>
<td>GMM (4)</td>
<td>61 ± 4</td>
<td>88 ± 5</td>
<td>62 ± 6</td>
</tr>
<tr>
<td>GMM (5)</td>
<td>61 ± 4</td>
<td>88 ± 5</td>
<td>59 ± 6</td>
</tr>
<tr>
<td>KNN (1)</td>
<td>59 ± 4</td>
<td>77 ± 7</td>
<td>57 ± 6</td>
</tr>
<tr>
<td>KNN (3)</td>
<td>60 ± 4</td>
<td>78 ± 6</td>
<td>58 ± 7</td>
</tr>
<tr>
<td>KNN (5)</td>
<td>56 ± 3</td>
<td>70 ± 6</td>
<td>56 ± 6</td>
</tr>
</tbody>
</table>
## Confusion Matrix

<table>
<thead>
<tr>
<th>Actual Genre</th>
<th>Classic</th>
<th>Country</th>
<th>Disco</th>
<th>Hiphop</th>
<th>Jazz</th>
<th>Rock</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classic</td>
<td>86</td>
<td>2</td>
<td>0</td>
<td>4</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>Country</td>
<td>1</td>
<td>57</td>
<td>5</td>
<td>1</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>Disco</td>
<td>0</td>
<td>6</td>
<td>55</td>
<td>4</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Hiphop</td>
<td>0</td>
<td>15</td>
<td>28</td>
<td>90</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>Jazz</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>37</td>
<td>12</td>
</tr>
<tr>
<td>Rock</td>
<td>6</td>
<td>19</td>
<td>11</td>
<td>0</td>
<td>27</td>
<td>48</td>
</tr>
</tbody>
</table>
## Confusion Matrix within Classical

<table>
<thead>
<tr>
<th></th>
<th>Choral</th>
<th>Orchestral</th>
<th>Piano</th>
<th>String-Quar.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choral</td>
<td>99</td>
<td>10</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>Orchestral</td>
<td>0</td>
<td>53</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Piano</td>
<td>1</td>
<td>20</td>
<td>75</td>
<td>3</td>
</tr>
<tr>
<td>String-Quar.</td>
<td>0</td>
<td>17</td>
<td>7</td>
<td>80</td>
</tr>
</tbody>
</table>
Importance of Features

Achtung: das Resultat bedeutet nicht, dass der Rhythmus das wichtigste Merkmal ist!
Summary

• Audio retrieval is a relatively new field
• Various different types of features
  • MFCC
  • Surface
  • Rhythm
• kNN and GMM Classifier