Introduction

- The members of a particular genre share certain characteristics
- Automatic musical genre classification
  - Music information retrieval
  - Developing and evaluating features that can be used in similarity retrieval, classification, segmentation, and audio thumbnailing
Related Work

- Audio classification has a long history originating from speech recognition
  - Classify audio signals into music, speech, and environmental sounds
  - Classify musical instrument sounds and sound effects
- The features they used are not adequate for automatic musical genre classification
Feature Extraction

- Timbral Texture Features
  - spectral centroid, spectral rolloff, spectral flux, zero-crossing rate, MFCC, energy
- Rhythmic Content Features
- Pitch Content Features
Spectral Centroid

- The center of gravity of the magnitude spectrum of short-time Fourier transform (STFT)

\[
C_t = \frac{\sum_{n=1}^{N} M_t[n] \cdot n}{\sum_{n=1}^{N} M_t[n]}
\]

- A measure of spectral shape and higher centroid values correspond to “brighter” textures with high frequencies

\(M_t[n]\) is the magnitude of the Fourier transform at frame \(t\) and frequency bin \(n\)
Spectral Rolloff

- The frequency $R_t$ such that
  $$\sum_{n=1}^{R_t} M_t[n] = 0.85 \sum_{n=1}^{N} M_t[n]$$

- A measure of the “skewness” of the spectral shape

- It is used to distinguish voiced from unvoiced speech and music. (unvoiced speech has a high proportion of energy contained in the high-freq. range of the spectrum)
Spectral Flux

- Squared difference between the normalized magnitudes of successive spectral distributions

\[ F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2 \]

- \( N_t[n] \) and \( N_{t-1}[n] \) are the normalized magnitude of the Fourier transform at frames \( t \) and \( t-1 \)

- A measure of the amount of local spectral change
Zero-Crossing Rate

- A measure of the noisiness of the signal

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

*sign* function is 1 for positive arguments and 0 for negative arguments

\( x[n] \) is the time domain signal for frame \( t \)

- Unvoiced speech has a low volume but a high ZCR
Mel-Frequency Cepstral Coefficients (MFCC)

- First five coefficients provide the best genre classification performance

\[ X_a[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N}, \quad 0 \leq k < N \]

\[ S[m] = \ln \left[ \sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k] \right], \quad 0 < m \leq M \]

\[ c[n] = \sum_{m=0}^{M-1} S[m] \cos(\pi n (m-1/2)/M), \quad 0 \leq n < M \]

*M*: the number of filters

*N*: the size of the FFT
Examples of Audio Features

Clip-level Zero-Crossing Rate

Speech  Music
Analysis and Texture Window (1/2)

- For short-time audio analysis, small audio segments are processed (*analysis window*).
- To capture the long term nature of sound “texture”, means and variances of features over a number of *analysis windows* are calculated (*texture windows*).
- For each texture window, multidimensional Gaussian distribution of features are estimated.
Analysis and Texture Window (2/2)

23ms

(audio)

... analysis window ...

Feature values

(512 samples at 22050 Hz sampling rate)

1s

(audio)

... texture window ...

Means and variance of features

(43 analysis windows)
Low-Energy Feature

- Based on the texture window
- The percentage of analysis windows that have less energy than the average energy across the texture window.
- Ex: vocal music with silences have large low-energy value
Rhythmic Content Features

- Characteristics: the regularity of the rhythm, the relation of the main beat to the subbeats, and the relative strength of subbeats to the main beat

- Steps of a common automatic beat detector
  - 1. Filterbank decomposition
  - 2. Envelope extraction
  - 3. Periodicity detection algorithm used to detect the lag at which the signal’s envelope is most similar to itself

- Similar to pitch detection but with larger periods: approximately 0.5 to 1.5 s for beat vs. 2 ms to 50 ms for pitch
Rhythmic Content Features

- Based on discrete wavelet transform (DWT)
  - Overcome the resolution problems (people percept differently in different freq. bands)
  - The DWT can be viewed as a computationally efficient way to calculate an octave decomposition of the signal in frequency.
  - DAUB4 filters are used.
- Find the rhythmic structure: detect the most salient periodicities of the signal
Rhythmic Content Features

- Beat detection flowchart

**Beat**: the sequence of equally spaced phenomenal impulses which define a tempo for the music
在數理上，每一個八度音程(Octave)正好對應於不同的振動模式，而兩個八度音程差的音在頻率上正好差上兩倍。例如：在第0個八度的La(記為A₀)頻率為27.5 Hertz，則第1個八度的La(記為A₁)頻率即為27.5*2=55.0 Hertz。在這每一個八度的音程中，又可再將其等分為12個頻率差相近的音，這分別對應於【C Db D Eb E F Gb G Ab A Bb B】，這樣的等分法就是所謂的十二平均律(Twelve-Tone Scale)。這當中每一個音符所對應的頻率，都可以藉由數學的方程式準確的算出。
There are 12 semitones in one octave, so a tone of frequency $f_1$ is said to be a semitone above a tone with frequency $f_2$ iff

$$f_1 = 2^{1/12}f_2 = 1.05946f_2$$
將一種音色波形的大致輪廓描繪出來，就可以表示出該音色在音量變化上的特性，而這個輪廓就稱為Envelope(波封)

一個波封可以用4種參數來描述，分別是Attack(起音)、 Decay(衰減)、 Sustain(延持)、與 Release(釋音)，這四者也就是一般稱的"ADSR"。

![Envelope Diagram](image)
Envelop Extraction

- **Full Wave Rectification**
  \[ y[n] = |x[n]| \]
  To extract the temporal envelope of the signal rather than the time domain signal itself

- **Low-Pass Filtering (smoothing)**
  \[ y[n] = (1 - \alpha)x[n] + \alpha y[n-1], \quad \alpha = 0.99 \]
  To smooth the envelope

- **Downsampling**
  \[ y[n] = x[kn] \quad k=16 \]
  Reduce the computation time

- **Mean Removal**
  \[ y[n] = x[n] - E[x[n]] \]
  To make the signal centered to zero for the autocorrelation stage
Enhanced Autocorrelation

\[ y[k] = \frac{1}{N} \sum_n x[n]x[n - k] \]

- The peaks of the autocorrelation function correspond to the time lags where the signal is most similar to itself.
- The time lags correspond to beat periodicities.
Example

After enhance autocorrelation
Peak Detection and Histogram Calculation

- The first three peaks of the enhanced autocorrelation function are selected and added to a beat histogram (BH).
- The bins of BH correspond to beats-per-minute (bpm) from 40 to 200 bpm.
- For each peak, the peak amplitude is added to the histogram.
  - Peaks having high amplitude (where the signal is highly similar) are weighted more strongly
Beat Histogram

Multiple instruments of the orchestra, no strong self-similarity
Beat Histogram Features

- A0, A1: relative amplitude (divided by the sum of amplitudes) of the first and second histogram peak
- RA: ratio of the amplitude of the second peak divided by the amplitude of the first peak
- P1, P2: period of the first and second peaks in bpm
- SUM: overall sum of the histogram (indication of beat strength)
Introduction of Pitch

- **Pitch (音高):** 構成樂音的最基本要素在於音高，也就是聲音的頻率。

- 在樂理上，樂音音符可分為七個基本音，即【Do Re Me Fa Sol La Si】，以美式的符號則記為【C D E F G A B】而第八個音則稱為高八度的Do。
Pitch Content Feature

- The signal is decomposed into two frequency bands (below and above 1000 Hz)
- Envelope extraction is performed for each frequency band.
- The envelopes are summed and an enhanced autocorrelation function is computed.
- The prominent peaks correspond to the main pitches for that short segment of sound.
Beat and Pitch Detection

- The process of beat detection resembles pitch detection with larger periods.
- For beat detection, a window of 65536 samples at 22050 Hz is used.
- For pitch detection, a window of 512 samples is used.

\[ y[k] = \frac{1}{N} \sum_n x[n]x[n - k] \]

Different range of \( k \)
Pitch Histogram

- For each analysis window, the three dominant peaks are accumulated into a pitch histogram.
- The frequencies corresponding to each histogram peak are converted to musical notes.

\[ n = 12 \times \log_2 \frac{f}{440} + 69 \]

- \( f \) is the frequency in Hertz.
- \( n \) is the histogram bin (MIDI note number).


\[
\begin{align*}
69 & \quad \text{semitone} \\
70 & \quad \text{semitone}
\end{align*}
\]
Folded and Unfolded PH

- In the folded case (FPH)
  \[ c = n \mod 12 \]
  - \( c \) is the folded histogram bin
  - \( n \) is the unfolded histogram bin

- The folded version (FPH) contains information regarding the pitch classes or harmonic content of the music. The unfolded version (UPH) contains information about the pitch range of the piece.
Modified FPH

- The FPH is mapped to a circle of fifths histogram so that adjacent histogram bins are spaced a fifth apart rather than a semitone.
  
  \[ c' = (7 \times c) \mod 12 \]

- The distances between adjacent bins after mapping are better suited for expressing tonal music relations.

- Jazz or classical music tend to have a higher degree of pitch change than rock or pop music.
Pitch Histogram Features

- **FA0**: amplitude of maximum peak of the folded histogram.
- **UP0, FP0**: period of the maximum peak of the unfolded and folded histograms.
- **IPO1**: pitch interval between the two most prominent of the folded histogram (main tonal interval relation).
- **SUM**: the overall sum of the histogram.
Evaluation

- **Classification**
  - Simple Gaussian classifier
  - Gaussian mixture model
  - K-nearest neighbor classifier

- **Datasets**
  - 20 musical genres and 3 speech genres
  - 100 excerpts each with 30 sec
  - Taken from radio, CD, and mp3. The files were stored as 22050 Hz, 16-bit, mono audio files.
Experiments

• Use a single-vector to represent the whole audio file.
• The vector consists of timbral texture features \((9(\text{FFT})+10(\text{MFCC})=19-\text{dim})\), rhythmic content features \((6-\text{dim})\), and the pitch content features \((5-\text{dim})\)
• 10-fold cross validation \((90\% \text{ training and } 10\% \text{ testing each time})\)
Results

- RT GS: for real-time classification per frame using only timbral texture feature
- GS: simple Gaussian

**TABLE 1**

<table>
<thead>
<tr>
<th>Method</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>

Random, RT GS, and GMM(3)
Other Classification Results

- The STFT-based feature set is used for the music/speech classification
  - 86% accuracy
- The MFCC-based feature set is used for the speech classification
  - 74% accuracy
Detailed Performance

The matrix shows that the misclassifications of the system are similar to what a human would do.

- 26% of classical music is wrongly classified as jazz music.
- The matrix shows that the misclassifications of the system are similar to what a human would do.
- Rock music has worst accuracy because of its broad nature.
Performance on Classical and Jazz

### TABLE III
**JAZZ CONFUSION MATRIX**

<table>
<thead>
<tr>
<th></th>
<th>BBand</th>
<th>Cool</th>
<th>Fus.</th>
<th>Piano</th>
<th>4tet</th>
<th>Swing</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBand</td>
<td>42</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Cool</td>
<td>21</td>
<td>67</td>
<td>5</td>
<td>4</td>
<td>23</td>
<td>10</td>
</tr>
<tr>
<td>Fus.</td>
<td>28</td>
<td>16</td>
<td>88</td>
<td>0</td>
<td>38</td>
<td>22</td>
</tr>
<tr>
<td>Piano</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>80</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4tet</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td>Swing</td>
<td>4</td>
<td>10</td>
<td>4</td>
<td>16</td>
<td>14</td>
<td>62</td>
</tr>
</tbody>
</table>

### TABLE IV
**CLASSICAL CONFUSION MATRIX**

<table>
<thead>
<tr>
<th></th>
<th>Choir</th>
<th>Orch.</th>
<th>Piano</th>
<th>Str.4tet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choir</td>
<td>99</td>
<td>7</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Orch.</td>
<td>0</td>
<td>58</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Piano</td>
<td>0</td>
<td>9</td>
<td>86</td>
<td>4</td>
</tr>
<tr>
<td>Str.4tet</td>
<td>1</td>
<td>26</td>
<td>5</td>
<td>86</td>
</tr>
</tbody>
</table>

BBand: bigband  
Cool: cool  
Fus.: fusion  
Piano: piano  
4tet: quartet (四重奏)  
Swing: swing  

Choir: choir  
Orch.: orchestra  
Piano: piano  
Str.4tet: String Quartet (弦樂四重奏)
Importance of Texture Window Size

- 40 analysis windows was chosen
Importance of Individual Feature Sets

- Pitch histogram features and beat histogram features perform worse than the timbral-texture features (STFT, MFCC)

The rhythmic and pitch content feature sets seem to play a less important role in the classical and jazz dataset classification.

It’s possible to design genre-specific feature sets.
Human Performance for Genre Classification

- Ten genres used in previous study: blues, country, classical, dance, jazz, latin, pop, R&B, rap, and rock
- 70% correct after listening to 3 sec
- Although direct comparison of these results is not possible, it’s clear that the automatic performance is not far away from the human performance.
Conclusion

- Three feature sets are proposed: timbral texture, rhythmic content, and pitch content features
- 61% accuracy has been achieved
- Possible improvements:
  - Information from melody and singer voice
  - Expand the genre hierarchy both in width and depth
  - More exploration of pitch content features

Multimedia Content Analysis, CSIE, CCU
Introduction

- Model and detect three sound effects: laughter, applause, and cheer
- Sound effect detection must handle the following cases:
  - Model more particular sound classes
  - Recall the expected sound effects only and ignore others
- Characteristics:
  - High recall and precision
  - Extensibility: it should be easy to add or remove sound effect models for new requirements.
All audio streams are 16-bit, mono-channel, and down-sampled to 8kHz. Each frame is of 200 samples (25 ms), with 50% overlaps.

- **Features**
  - Short-time energy
  - Average ZCR
  - Sub-band energies
  - Brightness and bandwidth
  - 8 order MFCC

- These features form a 16-dimensional feature vector for a frame. To describe the variance between frames, the gradient feature of adjacent frames is also considered, and is concatenated to the original vector. Thus we have a 32-dim feature vector for each frame.
HMMs can describe the time evolution between states using the transition probability matrix.

A complete connected HMM is used for each sound effect, with the 4 continuous Gaussian mixtures modeling each state.

Training data: 100 pieces of samples segmented from audio-track. Each piece is about 3s-10s and totally about about 10 min training data for each class.

A clustering algorithm is used to determine the state numbers of HMM.

2 for applause, and 4 for cheer and laughter
Sound Effect Detection

- 1s moving window with 50% overlapping
- Each data window is further divided into 25ms frames with 50% overlapping
- Silence window is skipped

$$\text{Average}STE < \delta_E \& \text{Average}ZCR < \delta_Z$$

- Non-silence window is compared against each sound effect model to get likelihood score $$P(O|\lambda)$$ or $$\log P(O|\lambda)$$
Log-Likelihood Scores Based Decision Method

- Unlike audio classification, we can’t simply classify the sliding window into the class which has the maximum log-likelihood score.
- Each log-likelihood score is examined to see if the window data is “accepted” by the corresponding sound effect.
- Optimal decision based on Bayesian decision theory
Log-Likelihood Scores Based Decision Method

- **Cost function**

\[ C = P(C_j)P(\overline{C_j}|C_j)C_{\overline{C_j}|C_j} + P(\overline{C_j})P(C_j|\overline{C_j})C_{C_j|\overline{C_j}} \]

- \( C_{\overline{C_j}|C_j} \) is the cost of false rejection
- \( C_{C_j|\overline{C_j}} \) is the cost of false acceptance

- **To minimize the cost, use Bayesian decision rule**

\[ \frac{p(s_j|C_j)}{p(s_j|\overline{C_j})} \geq R_j \quad \text{(likelihood ratio)} \]

- \( s_j \) is the log-likelihood score under HMM of sound effect \( C_j \)
- \( p(s_j|C_j) \) and \( p(s_j|\overline{C_j}) \) are probability distributions of log-likelihood scores of the samples within and outside \( C_j \), respectively.

*Multimedia Content Analysis, CSIE, CCU*
Log-Likelihood Scores Based Decision Method

- Bayesian threshold:

\[ R_j = \frac{C_{c_j | \overline{c}_j}}{C_{\overline{c}_j | c_j}} \times \frac{P(\overline{c}_j)}{P(c_j)} \]

- The priori probabilities are estimated based on the database.

- The cost of FR is set larger than that of FA, given that a high recall ratio is more important for summarization and highlight extraction.
The distribution of samples within and outside the sound effect applause.

To approximate these distributions (asymmetric), it’s more reasonable to use negative Gamma distribution

\[ p(s_j|C_j) = \frac{1}{\beta^\alpha \Gamma(\alpha)} s^{\alpha-1} e^{-s/\beta} \]

\[ \alpha = \frac{\mu^2}{\sigma^2} \]

\[ \beta = \frac{\sigma^2}{\mu} \]

Figure 3. (a) log-likelihood scores distributions; (b) Approximate (a) with Gamma distribution
Abnormal scores are pruned first
- Score whose distance to $\mu$ are larger than $2\sigma$ are abnormal

- The windows that confirms to \( \frac{p(s_j|C_j)}{p(s_j|\overline{C}_j)} \geq R_j \) are considered to be “accepted” by a sound effect.

- If it is accepted by a sound effect, the corresponding likelihood score is considered as confidence.

- It is classified into the $i$th sound effect if

\[
i = \arg \max_j p(s_j|C_j)
\]
Overall

The confidence score of an audio segment:
(based on likelihood ratio)

\[
\frac{p(s_j|C_j)}{p(s_j|\overline{C}_j)} \geq R_j
\]
Audio attention model is constructed to describe the saliency of each sound effect.

Based on energy and confidence in sound effects

\[
\bar{E} = \frac{E_{avr}}{\text{Max } E_{avr}}
\]

\[
\bar{P}_j = \exp(s_j - \text{Max } s_j)
\]

\(E_{avr}\) and \(s_j\) denote the average energy and log-likelihood score under model \(j\) of an audio segment.

The attention model for class \(j\) is defined as

\[
M_j = \bar{E} \times \bar{P}_j
\]
Sound Effect Attention Model

Figure 4. Sound effect attention model curves
Experiments

- The testing database is about 2 hours videos, including NBC’s TV show (30 min), CCTV’s TV show (60 min), and table tennis (30 min).
- Two kinds of distribution curves – Gaussian and Gamma – are compared.
  - Gamma distribution increases the precision by 9.3%, while just affects the recall ratio by 1.8%.

<table>
<thead>
<tr>
<th>Table 1. Performance on different p.d.f. distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>p.d.f.</strong></td>
</tr>
<tr>
<td>Gaussian</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Gamma</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Experiments

- Average recall is 92.95% and average precision is 86.88%.
- Higher recall can meet the requirements for highlights extraction and summarization.
- In table tennis, reporters’ exciting voice would be detected as laughter. Moreover, sound effects are often mixed with music, speech, and other environment sounds.

![Table 2. Performance of the algorithm](image)

<table>
<thead>
<tr>
<th>Video</th>
<th>Sound Effect</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hollywood Square</td>
<td>laughter</td>
<td>0.927</td>
<td>0.879</td>
</tr>
<tr>
<td></td>
<td>applause</td>
<td>0.910</td>
<td>0.850</td>
</tr>
<tr>
<td></td>
<td>cheer</td>
<td>0.907</td>
<td>0.916</td>
</tr>
<tr>
<td>Lucky 52</td>
<td>laughter</td>
<td>0.956</td>
<td>0.813</td>
</tr>
<tr>
<td></td>
<td>applause</td>
<td>0.894</td>
<td>0.826</td>
</tr>
<tr>
<td></td>
<td>cheer</td>
<td>0.910</td>
<td>0.917</td>
</tr>
<tr>
<td>Table Tennis Championship</td>
<td>laughter</td>
<td>0.977</td>
<td>0.778</td>
</tr>
<tr>
<td></td>
<td>applause</td>
<td>0.956</td>
<td>0.945</td>
</tr>
<tr>
<td></td>
<td>cheer</td>
<td>0.957</td>
<td>0.946</td>
</tr>
</tbody>
</table>

