Content-based music retrieval

Outline:
- Introduction
- Music similarity
- Music classification
- Transcription / separation oriented approaches
- Lyrics and paralinguistic information
- Efficient indexing techniques

Introduction

- Music information retrieval (MIR) is currently an active research area
- See proceedings of ISMIR conference and annual MIREX evaluations
- Background for this growth:
  We have huge amounts of music available and easily accessible over the Internet

[iTunes]

[Spotify]
Introduction
Plenty of music – Consequences

- Traditional ways of finding music are no longer sufficient
  - we cannot browse through all the music we would potentially like
  - record companies and radio stations are no longer critical gatekeepers in music distribution

- Relying just on popularity statistics is not effective
  - music tastes are so different that averaging opinions does not produce precise information for an individual
  - "UK Singles Chart" etc. sales statistics work badly as a guide for the consumer

Modern ways of searching music

Two complementary approaches:

1. Collaborative filtering
   - based on \((\text{users} \times \text{items})\) matrix = "music likings metadata"
   - recommend music by comparing user profiles and predicting likings for new pieces
   - measure similarity of music pieces (acoustics, usage, etc.) based on piece profiles

2. Content-based retrieval \(<\text{topic of this talk}\>
   - either based on automatic signal analysis or collaborative tagging by users

- Old ways of discovering music are still relevant too (though ineffective)
  - talking to friends, relying on experts (e.g. listening to FM radio you like)
Introduction

Audio-based MIR is needed

- Collaborative filtering (CF) does not solve it all
  - CF does not allow separating the various dimensions of music similarity, but these are all mixed in the piece profile
  - CF alone is not able to deal with items that are new or do not have many listeners
- Audio-based MIR addresses the above problems
  - enables "truly musical queries" with specific musical criteria, such as requesting pieces with certain vocal characteristics or slow tempo
  - can be employed even on media libraries that do not have any audience of listeners
  - also, enables musically interesting listening UIs that encourage music understanding

On the other hand, audio-based MIR alone cannot measure aspects like quality, usage, or culture
→ The two approaches are complementary

Introduction

Manual tagging as an alternative?

- Music annotation by human experts is costly and limits the coverage
  - not easy to integrate in music production process since music making is anarchistic
  - Pandora.com is audio-based MIR service (US only) based on expert tagging

Collaborative tagging by music service users (for example last.fm) is effective for items that are sufficiently popular
Tagging games can achieve better coverage, but (currently) less users

Music retrieval 9

SGN-24006 / A.K.

Manual tagging as an alternative?

- Music annotation by human experts is costly and limits the coverage
  - not easy to integrate in music production process since music making is anarchistic
  - Pandora.com is audio-based MIR service (US only) based on expert tagging

Collaborative tagging by music service users (for example last.fm) is effective for items that are sufficiently popular
Tagging games can achieve better coverage, but (currently) less users

Music retrieval 10

SGN-24006 / A.K.

MIR user interfaces

- Non-speech audio is difficult to describe with words
→ expressing a music query is hard
- Query mechanisms
  - query by example
  - browse by similarity (see Figure)
  - query by humming or tapping
  - tempo
  - lyrics
  - music categories (genre, mood, tags)
- Tailor UI to match user abilities
  → thumbnail extraction (chorus detection etc.)
  → also spatialisation for simultaneous presentation has been tried

Music similarity

- Music similarity estimation enables query by example and browsing by artist similarity
- Widely used acoustic features
  - Mel-frequency cepstral coefficients (MFCCs) → timbre/instrumentation
  - chroma [Bartsch-2001]: collapse spectral content into one octave and use 12 bins for the total spectral energy on each pitch class (c, c#, d,.....b) → harmonic content
  - rhythmogram (or, fluctuation patterns): cosine transform in blocks that extend in time direction → rhythm
- For a more comprehensive list, see e.g. [Peeters-2004]

Music retrieval 11

SGN-24006 / A.K.

Music retrieval 12

SGN-24006 / A.K.
Features define "similarity"

- "Similarity" as such is not well-defined
- For example: Is *Bohemian rhapsody* by Queen more similar to
  a) *Bohemian rhapsody* by London Symphony Orchestra, or
  b) *Killer Queen* by Queen?

The riddle is solved by choosing the acoustic features
- chroma → a) is more similar (composition)
- MFCCs → b) is more similar (instrumentation)

User may wish to specify the features when doing query by example

Narrowing down "a perfect piece" by using multiple examples
- enabled when huge amounts of music is available

Similarity measures between audio clips

- Similarity between two audio signals is *typically* calculated based on the statistics of extracted features
- "Bag of features" approach: collapse all temporal structure in data
- Traditional distance measures utilize means and covariances of the features. For example *Mahalanobis distance* between clips $f$ and $g$:

$$D_{\text{mah}}(f, g) = \left( \mu_f - \mu_g \right)^T \Sigma^{-1} \left( \mu_f - \mu_g \right)$$

- $\mu_f$, mean of features in clip $f$
- $\Sigma$, covariance of all features
**Similarity measures between audio clips**

- **Cross-likelihood ratio test** is a bit more sophisticated distance measure:

\[ C(A, B) = \frac{p_s(A) p_B(B)}{p_s(B) p_A(A)} \]

where

\[ p_s(A) = p(a_1, a_2, ..., a_n | \theta_s) \]

is the probability of feature vectors extracted from \( A \) given the model trained for \( B \) (GMM model, HMM model, or some other).

- Accurate and well motivated, but requires going through the feature vectors $\rightarrow$ computationally inefficient

**Music similarity: some evaluation results**

- **Results from MIREX 2007 evaluation** [Downie-ismir-2007]
  - organised by IMIRSEL Group at University of Illinois

<table>
<thead>
<tr>
<th>Rank</th>
<th>Participant</th>
<th>Avg. Fine Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pohle &amp; Schnitzer</td>
<td>0.568</td>
</tr>
<tr>
<td>2</td>
<td>Tzanetakis, G.</td>
<td>0.554</td>
</tr>
<tr>
<td>3</td>
<td>Barrington, Turnbull, Torres &amp; Lanskriek</td>
<td>0.541</td>
</tr>
<tr>
<td>4</td>
<td>Bastuck, C. (1)</td>
<td>0.539</td>
</tr>
<tr>
<td>5</td>
<td>Lidy, Rauber, Pertusa &amp; Iliea (1)</td>
<td>0.519</td>
</tr>
<tr>
<td>6</td>
<td>Mandel &amp; Ellis</td>
<td>0.512</td>
</tr>
<tr>
<td>7</td>
<td>Lidy, Rauber, Pertusa &amp; Iliea (2)</td>
<td>0.491</td>
</tr>
<tr>
<td>8</td>
<td>Bastuck, C. (2)</td>
<td>0.446</td>
</tr>
<tr>
<td>9</td>
<td>Bastuck, C. (3)</td>
<td>0.439</td>
</tr>
<tr>
<td>10</td>
<td>Bosteels &amp; Kerre (1)</td>
<td>0.412</td>
</tr>
<tr>
<td>11</td>
<td>Paradzinski &amp; Chen</td>
<td>0.377</td>
</tr>
<tr>
<td>12</td>
<td>Bosteels &amp; Kerre (2)*</td>
<td>0.178</td>
</tr>
</tbody>
</table>

**Music similarity: techniques used** *(reference material – do not memorize)*

- **Pohle-Schnitzer**
  - modification of [Pampalk-ismir-2006]
  - features: MFCCs, fluctuation patterns (0-10Hz at several frequency bands), “gravity” (slow of fast) + “bass” extracted from fluctuation patterns
  - features are averaged and normalised over the piece $\rightarrow$ one feature vector per song
  - cosine distance is employed as distance measure between the feature of two pieces

- **Tzanetakis**
  - several spectral features (incl. MFCCs)
  - two-step mean & std calculation of framewise features $\rightarrow$ one feature vector per song
  - normalization, Euclidean distance

- **Barrington-Turnbull-et al**
  - map audio tracks into a semantic feature space
  - resulting feature vector: 146-dimensional vector of posterior probabilities of certain concepts occurring, given the audio features $\rightarrow$ one feature vector per song
  - concepts included words that characterize the genre, instrumentation, vocals, emotion, rhythm, usage, etc.
  - similarity measured with KL divergence between two feature vectors

**pdf-based similarity measures**

- **Idea:** measure similarity by calculating distance between the probability density functions (pdfs) of features
  - each song is represented by its pdf of features instead of just one feature vector
  - more flexible and accurate than using the means and covariances of features
  - no need to go through the feature vectors (after the models have been trained), however computational complexity higher than when using 1 feature vector / song
  - For example Euclidean distance, Kullback-Leibler (KL) divergence etc.
Using temporal sequences for similarity

- Above methods collapse the time structure of feature sequences
- Using temporal sequences for similarity requires time-alignment
  - non-trivial: tempo differences, different numbers of sectional parts, etc.
- **Beat-synchronous feature extraction** reconciles for tempo differences
  - track the beat of each song and extract one feature vector per inter-beat-interval
- Used previously in
  - cover song detection using beat-synchronised chroma features (e.g. [Ellis-2006])
  - analysis of the sectional form (verse, chorus,...) of a piece (e.g. [Paulus-dafx-2008])

### Cover song identification: example methods (reference material – do not memorize)

- Serrà and Gómez
  - extract a sequence of tonal descriptors (harmonic pitch class profiles)
  - compute a similarity matrix between two pieces
  - use dynamic programming to align the two pieces in time and to obtain similarity
- Ellis and Cotton
  - beat-synchronized chroma features
  - cross-correlation of the feature sequences of two pieces

### Music classification

- Music can be classified according to genre, mood, etc.
- **Classical** train/test **supervised classification** scenario:

![Diagram of music classification process]

### Classification and identification tasks

- **Classify music into categories**
  - genre: rock, hip hop, jazz, classical,... (here 10)
  - mood: aggressive, passionate, humorous, cheerful,... (5)
  - artist identification (here 102 artists)
  - classical composer identification (here 11)

- **Train-test setup**

- **Results from MIREX 2007** [Downie-ismir-2007]
  - organised by IMIRSEL Group at University of Illinois
**Artist vs singer: task definition**

- Do we want to recognize the singer (person) or the artist name?

---

**Classification and identification tasks**

*(reference material – do not memorize)*

- Technically, the leading classification / identification systems are surprisingly similar
- Compare IMIRSEL vs. Mandel-Ellis vs. Tzanetakis vs. Guaus-Herrera
  - all use a single feature vector per audio clip
  - all obtain the feature by calculating statistics of feature over the clip (mean, std, covariances, ...)
  - frame-level features and the statistical measures do vary from system to system
  - all found support vector machine (SVM) classifier to be the best
- Participants did not vary their systems much between different tasks
- Convergence of the techniques does not mean we are done
  - glass ceiling... partly due to the fuzzyness of ground truth (genre, mood)

---

**Transcription / separation approaches**

- Approaches where some musically meaningful part of the signal, such as the melody line, is extracted and analyzed
Music transcription

- Transcription of
  - melody [Goto; Paiva; Ellis-Poliner; Dressler; Ryynänen]
  - bass line [Goto; Hainsworth; Ryynänen]
  - drums [Paulus; Yoshii; FitzGerald]
  - chords [Sheh-Ellis; Bello-Pickens; Harte; Lee]
  - key/mode (e.g. [Gomez-phd])
  - tempo, meter
  - instrument recognition

- Separation
  - vocals
  - drums

- MIREX 2007: polyphonic transcription task → [Downie-ismir-2007]

Query by humming

- Consists of two main steps:
  1. transcribing a hummed or sung query into a suitable higher-level representation
  2. matching that representation against a large database of known reference items

- Some QBH services are already available, for example SoundHound and Musipedia:

Query by humming of audio

- Example method [Ryynänen-icassp-2008]
  - preprocessing: extract melodies automatically from music pieces
  - transcribe the query
  - match by Euclidean distance between the two melodic contours (allow time scaling)
  - efficient indexing using locality sensitive hashing

- Demos
  - Example A
    - query retrieval results #1 #2 #3
  - Example B
    - query retrieval results #1 #2 #3
  - Example C
    - query retrieval results #1 #2 #3

Query by chord sequence similarity

- Query by example, determining similarity based on transcribed chords
  - database of 1294 music pieces
  - preprocessing: transcribe the chords (24 triads) from all pieces [Ryynänen-2008], Resample to get beat-synchronous chord sequence. Key normalisation.
  - query: let the user select 10 second segment from an arbitrary song
  - retrieve segment $i_t$ from song $i$ that has the most similar chord sequence

$$ i_t = \arg \min_{i,t} \sum_{m=1}^{T} d(x_{i,t}, y_{i,t}) $$

where distance $d(x,y)$ between chords $x$ and $y$ is the Euclidean distance in the chord space [Krumhansl-90] after key-normalisation
Rhythmic similarity

- Approaches
  - rhythmogram features + distance measure collapsing time structure [Dixon-03], [Paulus-dafx-2008]
  - framewise features + distance aligning with dynamic time warping [Paulus-02]
  - transcribe drums + similarity measure (e.g. Eigenrhythms [Ellis-04])
  - beat-synchronised features + distance measure comparing feature sequences
  - classification into rhythmic categories [Kapur-2004]

- Musipedia allows query by tapping (using the keyboard)
  www.musipedia.org

Lyrics: what is this song about?

- goal: recognize the words from a song

  Miranda

  in the morning

  takes her eggs sunny side up

Indexing techniques

- Locality sensitive hashing (LSH)
  - computationally efficient indexing technique to searching nearest neighbors
  1) in large databases and 2) in high-dimensional feature spaces [Datar-2004]
  - idea: project data points on random lines and subdivide the lines into hash buckets

Indexing-based audio analysis?

- Increased memory capacity and powerful indexing techniques allow storing examples as an alternative for training a (statistical) model

- Imagine a situation where we have indexed, say, $10^{10}$ audio signals on a server, and given a query example, could retrieve perceptually the most similar clips in an instant.

- Provided that some contextual information about the stored signals would be available too (which is realistic if the data is collected with mobile devices) this would result in a huge machine hearing system
Toolboxes

- See "Tools we use" page edited by Paul Lamere
  
  http://www.music-ir.org/evaluation/tools.html

Conclusions

- Music in large quantities functions as a resource that is potentially useful for many purposes: to pass time, to help concentrate in work, to improve physical exercise, to create suitable atmosphere in a social situation, etc.
- With the help of proper MIR tools, the large supply of music meets the even greater demand