EE E6820: Speech & Audio Processing & Recognition

Lecture 10: Music Analysis

Michael Mandel <mim@ee.columbia.edu>

Columbia University Dept. of Electrical Engineering
http://www.ee.columbia.edu/~dpwe/e6820

April 17, 2008

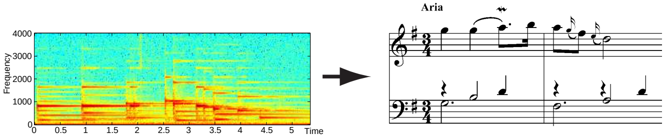
1. Music transcription
2. Score alignment and musical structure
3. Music information retrieval
4. Music browsing and recommendation

Outline

1. Music transcription
2. Score alignment and musical structure
3. Music information retrieval
4. Music browsing and recommendation

Music Transcription

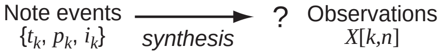
- Basic idea: recover the score



- Is it possible? Why is it hard?
  - music students do it
  - ... but they are highly trained; know the rules
- Motivations
  - for study: what was played?
  - highly compressed representation (*e.g.* MIDI)
  - the ultimate restoration system...
- Not trivial to turn a “piano roll” into a score
  - meter determination, rhythmic quantization
  - key finding, pitch spelling

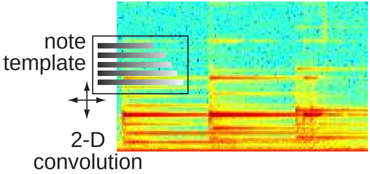
Transcription framework

- Recover discrete events to explain signal



  - analysis-by-synthesis?
- Exhaustive search?
  - would be possible given exact note waveforms

  . . . or just a 2-dimensional 'note' template?



  - but superposition is not linear in |STFT| space
- Inference depends on all detected notes
  - is this evidence 'available' or 'used'?
  - full solution is exponentially complex

Problems for transcription

- Music is practically worst case!
  - note events are often synchronized
    - $\rightarrow$ defeats common onset
  - notes have harmonic relations (2:3 etc.)
    - $\rightarrow$ collision/interference between harmonics
  - variety of instruments, techniques, ...
- Listeners are very sensitive to certain errors

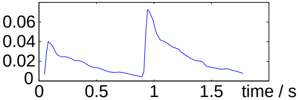
  ... and impervious to others
- Apply further constraints
  - like our 'music student'
  - maybe even the whole score!

Types of transcription

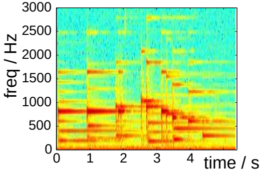
- Full polyphonic transcriptions is hard, but maybe unnecessary
- Melody transcription
  - the melody is produced to stand out
  - useful for query-by-humming, summarization, score following
- Chord transcription
  - consider the signal holistically
  - useful for finding structure
- Drum transcription
  - very different from other instruments
  - can't use harmonic models

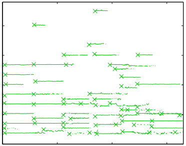
Spectrogram Modeling

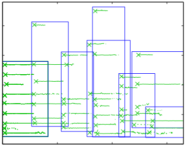
- Sinusoid model
  - as with synthesis, but signal is more complex
- Break tracks
  - need to detect new 'onset' at single frequencies



- Group by onset & common harmonicity
  - find sets of tracks that start around the same time
  - \+ stable harmonic pattern
- Pass on to constraint-based filtering. . .

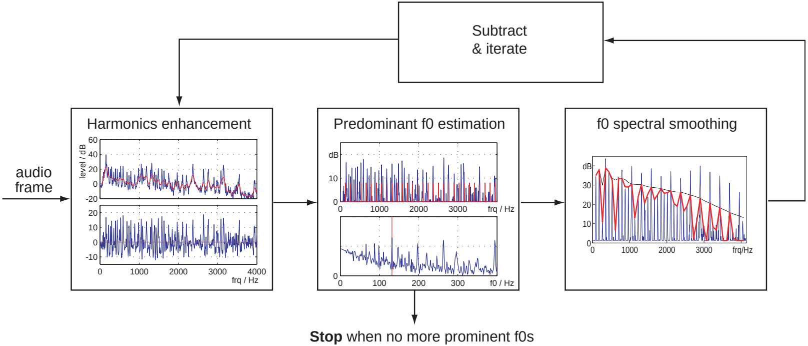






Searching for multiple pitches (Klapuri, 2005)

- At each frame:
  - estimate dominant f_0 by checking for harmonics
  - cancel it from spectrum
  - repeat until no f_0 is prominent



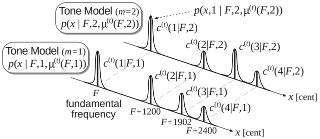
- Can use pitch predictions as features for further processing
  e.g. HMM

Probabilistic Pitch Estimates (Goto, 2001)

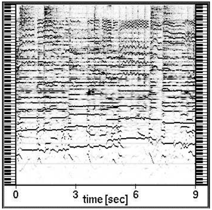
- Generative probabilistic model of spectrum as weighted combination of tone models at different fundamental frequencies:

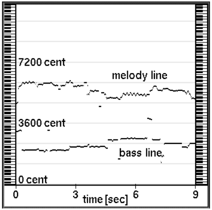
$$p(x(f)) = \int \left(\sum_m w(F, m) p(x(f) \mid F, m) \right) dF$$

- 'Knowledge' in terms of tone models + prior distributions for f_0:

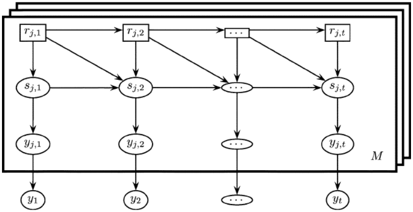


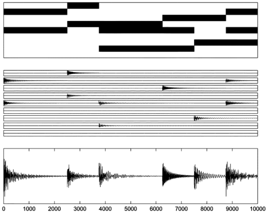
- EM (RT) results:





Generative Model Fitting (Cemgil et al., 2006)





- Multi-level graphical model
  - $r_{j,t}$: piano roll note-on indicators
  - $s_{j,t}$: oscillator state for each harmonic
  - $y_{j,t}$: time-domain realization of each note separately
  - y_t: combined time-domain waveform (actual observation)
- Incorporates knowledge of high-level musical structure and low-level acoustics
- Inference exact in some cases, approximate in others
  - special case of the generally intractable switching Kalman filter

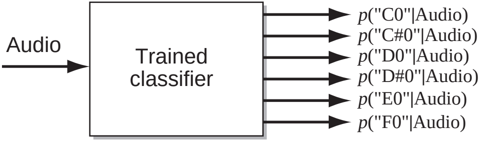
Transcription as Pattern Recognition (Poliner and Ellis)

- Existing methods use prior knowledge about the structure of pitched notes

  i.e. we *know* they have regular harmonics

- What if we didn't know that, but just had examples and features?
  - the classic pattern recognition problem

- Could use music signal as evidence for pitch class in a black-box classifier:



  - nb: more than one class at once!

- But where can we get labeled training data?

Ground Truth Data

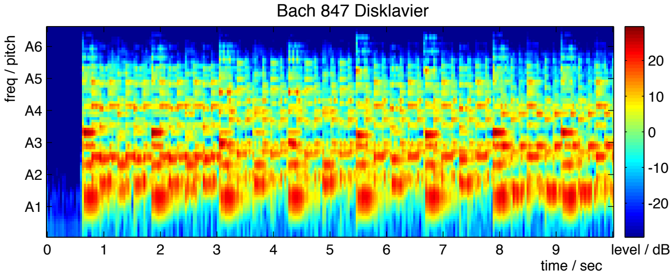
- Pattern classifiers need training data

  i.e. need {signal, note-label} sets

  i.e. MIDI transcripts of real music... already exists?
- Make sound out of midi
  - "play" midi on a software synthesizer
  - record a player piano playing the midi
- Make midi from monophonic tracks in a multi-track recording
  - for melody, just need *a capella* tracks
- Distort recordings to create more data
  - resample/detune any of the audio and repeat
  - add in reverb or noise
- Use a classifier to train a better classifier
  - alignment in the classification domain
  - run SVM & HMM to label, use to retrain SVM

Polyphonic Piano Transcription (Poliner and Ellis, 2007)

- Training data from player piano



- Independent classifiers for each note
  - plus a little HMM smoothing
- Nice results

  . . . when test data resembles training

Algorithm	Errs	False Pos	False Neg	d'
SVM	43.3%	27.9%	15.4%	3.44
Klapuri & Ryynänen	66.6%	28.1%	38.5%	2.71
Marolt	84.6%	36.5%	48.1%	2.35

Outline

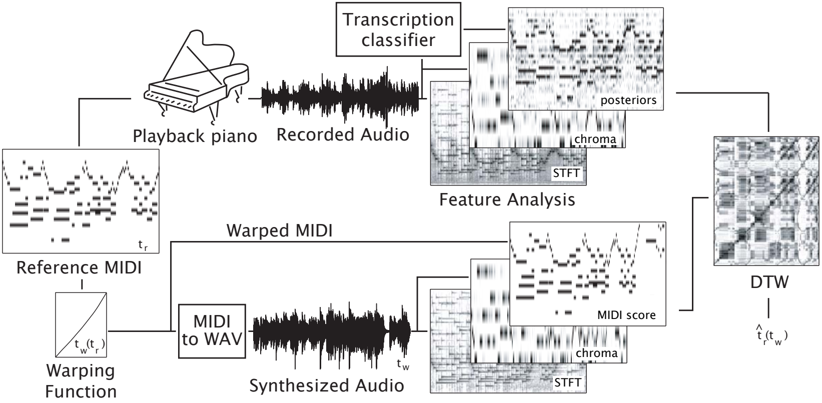
Midi-audio alignment

- Pattern classifiers need training data

  i.e. need {signal, note-label} sets

  i.e. MIDI transcripts of real music...already exists?
- Idea: force-align MIDI and original
  - can estimate time-warp relationships
  - recover accurate note events in real music!
- Also useful by itself
  - comparing performances of the same piece
  - score following, *etc.*

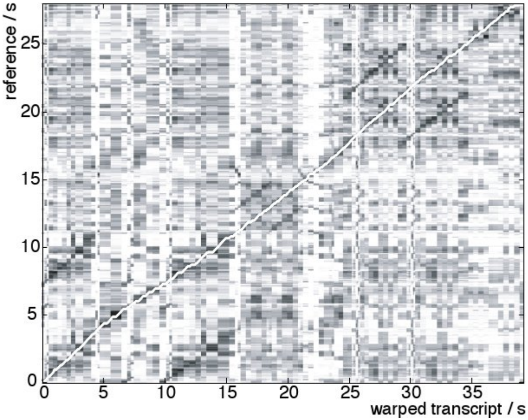
Which features to align with?



- Audio features: STFT, chromagram, classification posteriors
- Midi features: STFT of synthesized audio, derived chromagram, midi score

Alignment example

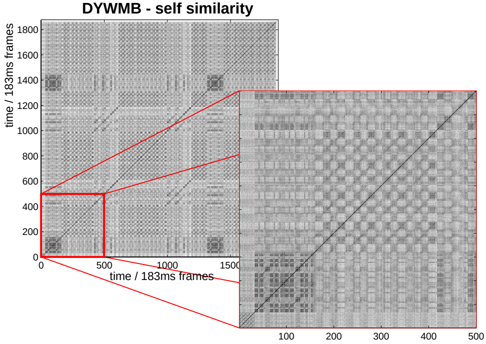
- Dynamic time warp can overcome timing variations



Segmentation and structure

- Find contiguous regions that are internally similar and different from neighbors

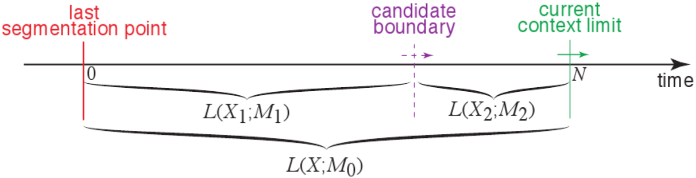
e.g. "self-similarity" matrix (Foote, 1997)



- ▸ 2D convolution of checkerboard down diagonal
- = compare fixed windows at every point

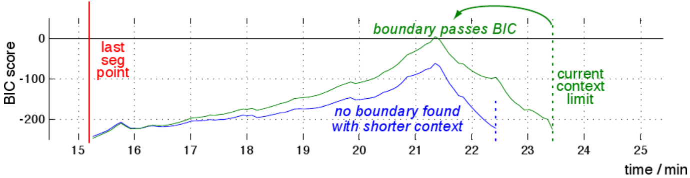
BIC segmentation

- Use evidence from whole segment, not just local window
- Do ‘significance test’ on every possible division of every possible context



$$\text{BIC}: \quad \log \frac{L(X_1; M_1) L(X_2; M_2)}{L(X; M_0)} \gtrless \frac{\lambda}{2} \log(N) \#(M)$$

- Eventually, a boundary is found:



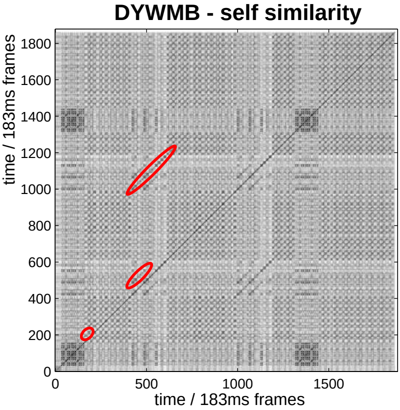
HMM segmentation

- Recall, HMM Viterbi path is joint classification and segmentation
  e.g. for singing/accompaniment segmentation
- But: HMM states need to be defined in advance
  - define a ‘generic set’? (MPEG7)
  - learn them from the piece to be segmented? (Logan and Chu, 2000; Peeters et al., 2002)
- Result is ‘anonymous’ state sequence characteristic of particular piece

```
U2-The_Joshua_Tree-01-Where_The_Streets_Have_No_Name 33677
17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17 17
17 17 17 17 17 17 17 17 17 17 17 17 17 17 1 7 17 17 17 17 17 17 17 17
17 17 17 17 17 17 17 17 17 17 11 11 11 11 11 11 11 11 11 11 11 11 11
11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11
11 1 1 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11 11
11 11 11 11 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15
15 15 15 15 15 15 15 15 15 15 15 1 5 15 15 15 15 15 15 15 15 15 15 15
15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15
15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 3 3 3 3 3 3 3 15
15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15
```

Finding Repeats

- Music frequently repeats main phrases
- Repeats give off-diagonal ridges in Similarity matrix (Bartsch and Wakefield, 2001)



- Or: clustering at phrase-level . . .

Music summarization

- What does it mean to 'summarize'?
  - compact representation of larger entity
  - maximize 'information content'
  - sufficient to recognize known item
- So summarizing music?
  - short version *e.g.* $< 10\%$ duration (< 20s for pop)
  - sufficient to identify style, artist

  e.g. chorus or 'hook'?
- Why?
  - browsing existing collection
  - discovery among unknown works
  - commerce. . .

Outline

Music Information Retrieval

- Text-based searching concepts for music?
  - "Google of music"
  - finding a specific item
  - finding something vague
  - finding something new
- Significant commercial interest
- Basic idea: Project music into a space where neighbors are "similar"
- (Competition from human labeling)

Music IR: Queries & Evaluation

- What is the form of the query?
- Query by humming
  - considerable attention, recent demonstrations
  - need/user base?
- Query by noisy example
  - "Name that tune" in a noisy bar
  - Shazam Ltd.: commercial deployment
  - database access is the hard part?
- Query by multiple examples
  - "Find me more stuff like this"
- Text queries? (Whitman and Smaragdis, 2002)
- Evaluation problems
  - requires large, shareable music corpus!
  - requires a well-defined task

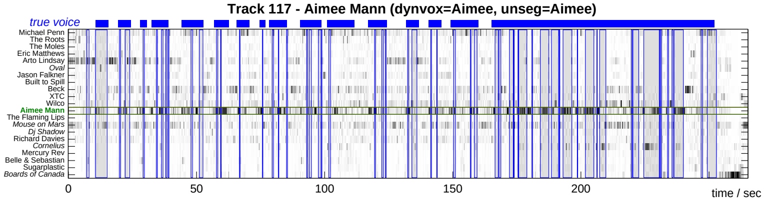
Genre Classification (Tzanetakis et al., 2001)

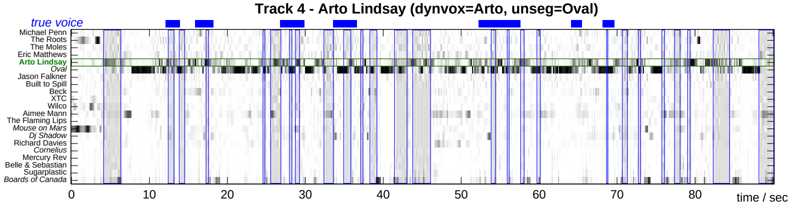
- Classifying music into genres would get you some way towards finding "more like this"
- Genre labels are problematic, but they exist
- Real-time visualization of "GenreGram":

  - 9 spectral and 8 rhythm features every 200ms
  - 15 genres trained on 50 examples each
  - single Gaussian model $\rightarrow \sim 60\%$ correct

Artist Classification (Berenzweig et al., 2002)

- Artist label as available stand-in for genre
- Train MLP to classify frames among 21 artists
- Using only "voice" segments:
  - Song-level accuracy improves 56.7% → 64.9%





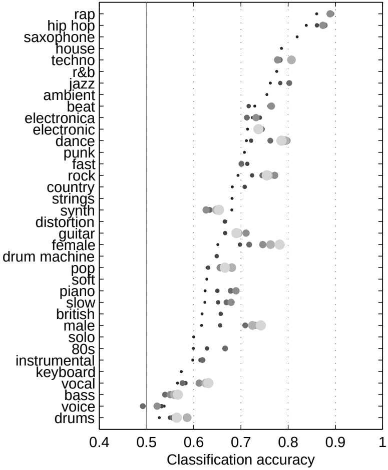
Textual descriptions

- Classifiers only know about the sound of the music not its context
- So collect training data just about the sound
  - Have humans label short, unidentified clips
  - Reward for being relevant and thorough

→ MajorMiner Music game

Tag classification

- Use tags to train classifiers
  - 'autotaggers'
- Treat each tag separately, easy to evaluate performance
- No 'true' negatives
  - approximate as absence of one tag in the presence of others



Outline

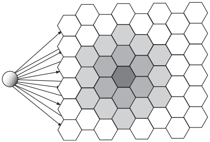
1. Music transcription
2. Score alignment and musical structure
3. Music information retrieval
4. Music browsing and recommendation

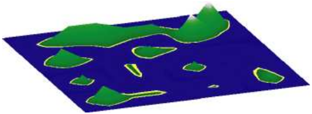
Music browsing

- Most interesting problem in music IR is finding new music
  - is there anything on myspace that I would like?
- Need a feature space where music/artists are arranged according to perceived similarity
- Particularly interested in little-known bands
  - little or no 'community data' (*e.g.* collaborative filtering)
  - audio-based measures are critical
- Also need models of personal preference
  - where in the space is stuff I like
  - relative sensitivity to different dimensions

Unsupervised Clustering (Rauber et al., 2002)

- Map music into an auditory-based space
- Build 'clusters' of nearby → similar music
    - "Self-Organizing Maps" (Kohonen)
- Look at the results:





- "Islands of music"
    - quantitative evaluation?

Artist Similarity

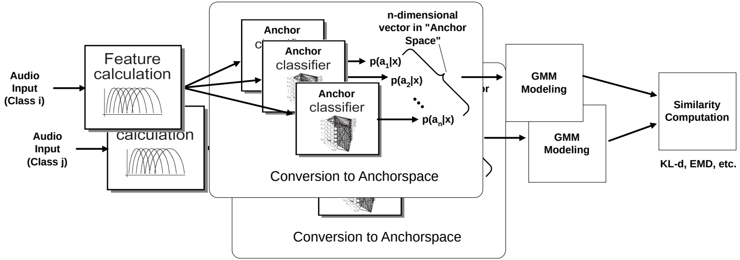
- Artist classes as a basis for overall similarity:
  Less corrupt than 'record store genres'?

- But: what is similarity
  between artists?
  - pattern recognition
    systems give a number...

- Need subjective ground truth:
  Collected via web site
  - `www.musicseer.com`
- Results: 1800 users, 22,500 judgments
  collected over 6 months

Anchor space

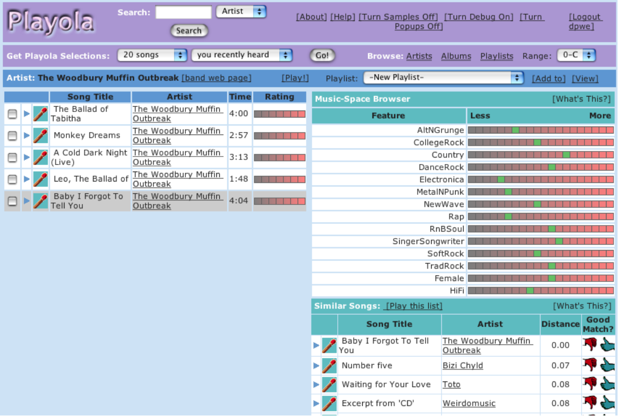
- A classifier trained for one artist (or genre) will respond partially to a similar artist
- A new artist will evoke a particular pattern of responses over a set of classifiers
- We can treat these classifier outputs as a new feature space in which to estimate similarity



- "Anchor space" reflects subjective qualities?

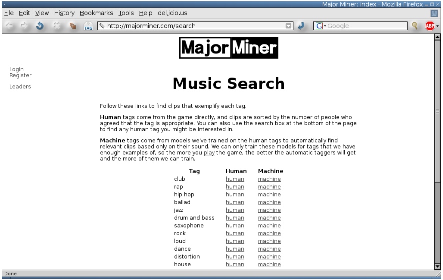
Playola interface (http://www.playola.org)

- Browser finds closest matches to single tracks or entire artists in anchor space
- Direct manipulation of anchor space axes



Tag-based search (http://majorminer.com/search)

- Two ways to search MajorMiner tag data
- Use human labels directly
  - (almost) guaranteed to be relevant
  - small dataset
- Use autotaggers trained on human labels
  - can only train classifiers for labels with enough clips
  - once trained, can label unlimited amounts of music



Music recommendation

- Similarity is only part of recommendation
  - need familiar items to build trust

  ... in the unfamiliar items (serendipity)
- Can recommend based on different amounts of history
  - none: particular query, like search
  - lots: incorporate every song you've ever listened to
- Can recommend from different music collections
  - personal music collection: "what am I in the mood for?"
  - online databases: subscription services, retail
- Appropriate music is a subset of good music?
- Transparency builds trust in recommendations
- See (Lamere and Celma, 2007)

Evaluation

- Are recommendations good or bad?
- Subjective evaluation is the ground truth

  ... but subjects don't know the bands being recommended

  - can take a long time to decide if a recommendation is good
- Measure match to similarity judgments

  e.g. musicseer data
- Evaluate on "canned" queries, use-cases
  - concrete answers: precision, recall, area under ROC curve
  - applicable to long-term recommendations?

Summary

- Music transcription
  - hard, but some progress
- Score alignment and musical structure
  - making good training data takes work, has other benefits
- Music IR
  - alternative paradigms, lots of interest
- Music recommendation
  - potential for biggest impact, difficult to evaluate

Parting thought

Data-driven machine learning techniques are valuable in each case

References

A. P. Klapuri. A perceptually motivated multiple-f0 estimation method. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 291–294, 2005.

M. Goto. A predominant-f¡sub¿0¡/sub¿ estimation method for cd recordings: Map estimation using em algorithm for adaptive tone models. In *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, volume 5, pages 3365–3368 vol.5, 2001.

A. T. Cemgil, H. J. Kappen, and D. Barber. A generative model for music transcription. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):679–694, 2006.

G. E. Poliner and D. P. W. Ellis. A discriminative model for polyphonic piano transcription. *EURASIP J. Appl. Signal Process.*, pages 154–154, January 2007.

J. Foote. A similarity measure for automatic audio classification. In *Proc. AAAI Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*, March 1997.

B. Logan and S. Chu. Music summarization using key phrases. In *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, volume 2, pages 749–752, 2000.

G. Peeters, A. La Burthe, and X. Rodet. Toward automatic music audio summary generation from signal analysis. In *Proc. Intl. Symp. Music Information Retrieval*, October 2002.

M. A. Bartsch and G. H. Wakefield. To catch a chorus: Using chroma-based representations for audio thumbnailing. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2001.

B. Whitman and P. Smaragdis. Combining musical and cultural features for intelligent style detection. In *Proc. Intl. Symp. Music Information Retrieval*, October 2002.

A. Rauber, E. Pampalk, and D. Merkl. Using psychoacoustic models and self-organizing maps to create a hierarchical structuring of music by musical styles. In *Proc. Intl. Symp. Music Information Retrieval*, October 2002.

G. Tzanetakis, G. Essl, and P. Cook. Automatic musical genre classification of audio signals. In *Proc. Intl. Symp. Music Information Retrieval*, October 2001.

A. Berenzweig, D. Ellis, and S. Lawrence. Using voice segments to improve artist classification of music. In *Proc. AES-22 Intl. Conf. on Virt., Synth., and Ent. Audio.*, June 2002.

P. Lamere and O. Celma. Music recommendation tutorial. In *Proc. Intl. Symp. Music Information Retrieval*, September 2007.