Lecture 10: Music Analysis

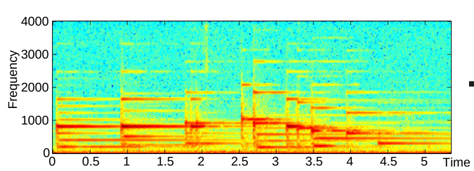
1. **Music Transcription**
2. **Music Summarization**
3. **Music Information Retrieval**
4. **Music Similarity Browsing**

Dan Ellis <dpwe@ee.columbia.edu>
http://www.ee.columbia.edu/~dpwe/e6820/

Columbia University Dept. of Electrical Engineering
Spring 2006

1 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...

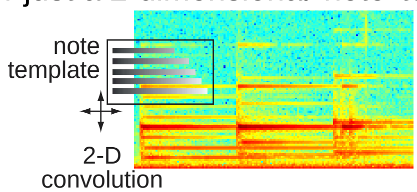
Transcription framework

- **Recover discrete events to explain signal**

  Note events $\{t_k, p_k, i_k\}$ $\xrightarrow{\textit{synthesis}}$? Observations $X[k,n]$

  - 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

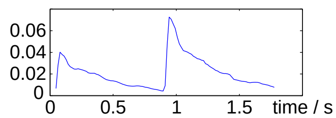
- **Inference depends on all detected notes**
  - is this evidence 'available' or 'used'?
  - full solution is exponentially complex

Problems for transcription

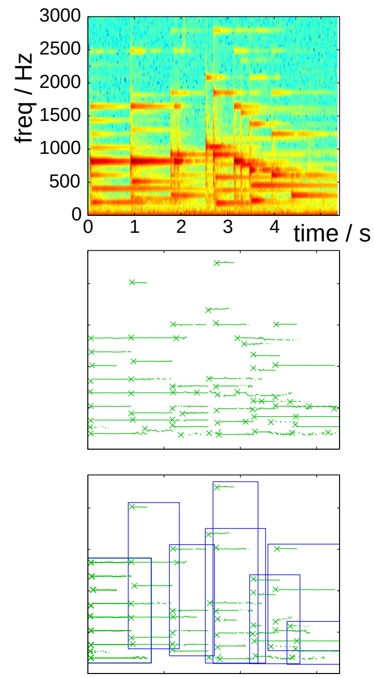
- **Music is practically worst case!**
  - note events are often synchronized
    → defeats common onset
  - notes have harmonic relations (2:3 etc.)
    → collision/interference between harmonics
  - variety of instruments, techniques, ...
- **Listeners are very sensitive to certain errors**
  - .. and impervious to others
- **Apply further constraints**
  - like our 'music student'
  - maybe even the whole score (Scheirer)!

Spectrogram Modeling

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



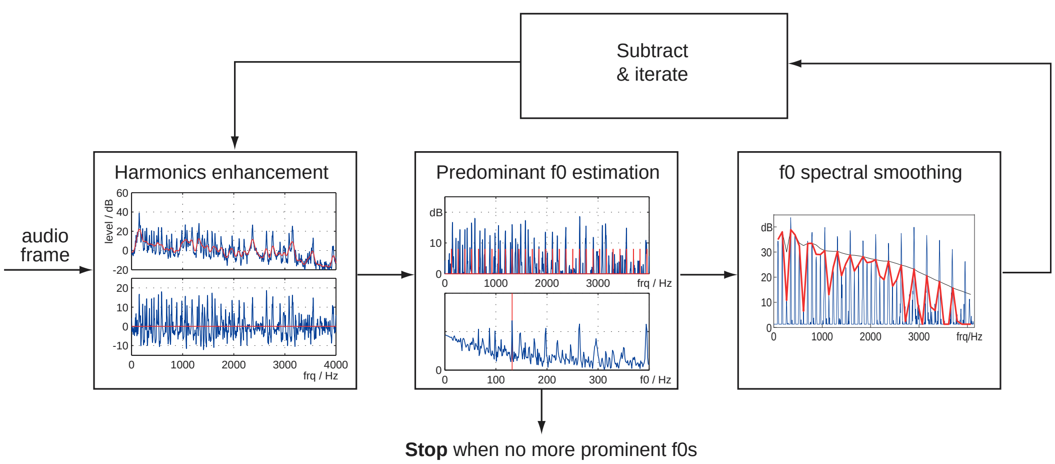
- **Group by onset & common harmonicity**
  - find sets of tracks that start around the same time
  - + stable harmonic pattern
- **Pass on to constraint-based filtering...**



Searching for multiple pitches

(Klapuri 2001)

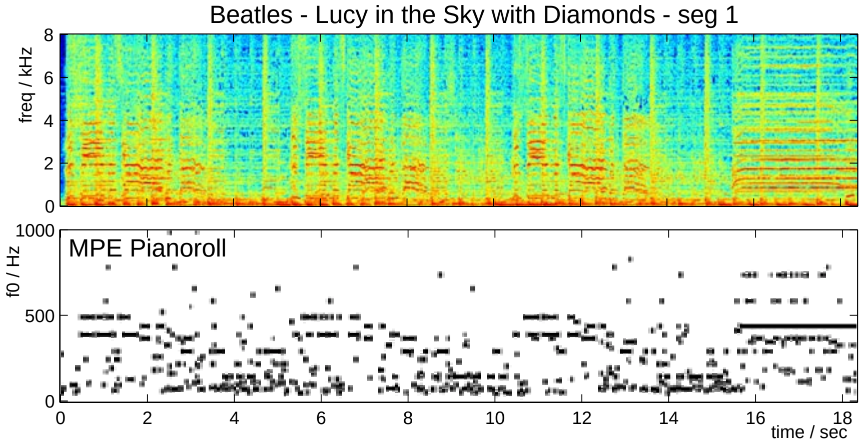
- **At each frame:**
  - estimate dominant f_0 by checking for harmonics
  - cancel it from spectrum
  - repeat until no f_0 is prominent



Multi-Pitch Extraction Results

(Rob Turetsky)

- **After continuity cleanup:**



- **Captures main notes, plus a lot else**
  - hand-tuned termination thresholds?
- **(Evaluation?)**

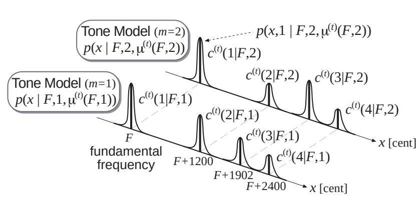
Probabilistic Pitch Estimates

(Goto 2001)

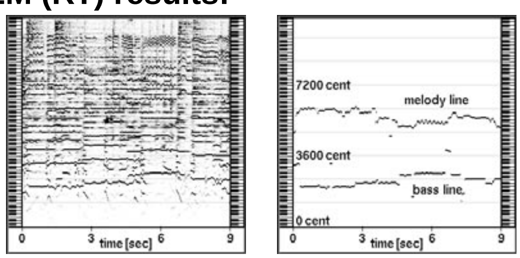
- **Generative probabilistic model of spectrum as weighted combination of tone models at different fundamental frequencies:**

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

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



- **EM (RT) results:**



Generative Model Fitting

(Walmsley et al. 1999)

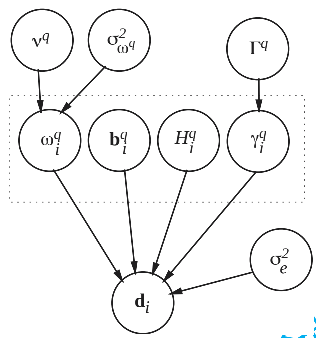
- **Generative model of harmonic complexes in the time domain:**

switch harmonic weights

samples $$\mathbf{d}_i = \sum_{q=1}^{Q} \gamma_i^q \mathbf{G}_i^q \mathbf{b}_i^q + \mathbf{e}_i$$ noise

voices harmonic bases

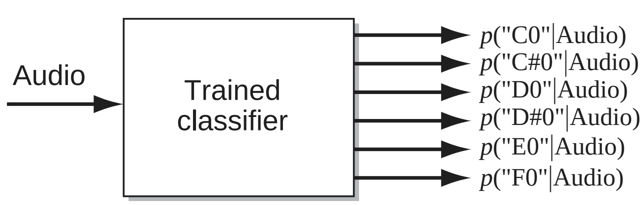
- **Too many parameters to solve by EM!**
  → Use Markov chain Monte Carlo (MCMC) to find good solution
- **Results?**



Transcription as Pattern Recognition

(Graham Poliner)

- **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:**

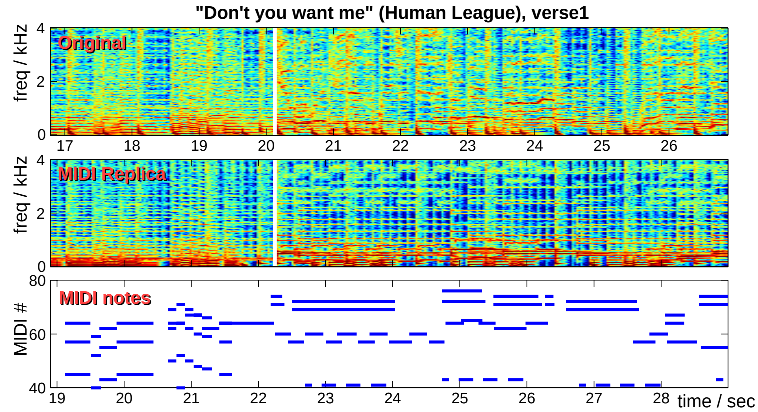


  - nb: more than one class at once!
- **But where can we get labeled training data?**

Ground Truth Data

(Turetsky & Ellis 2003)

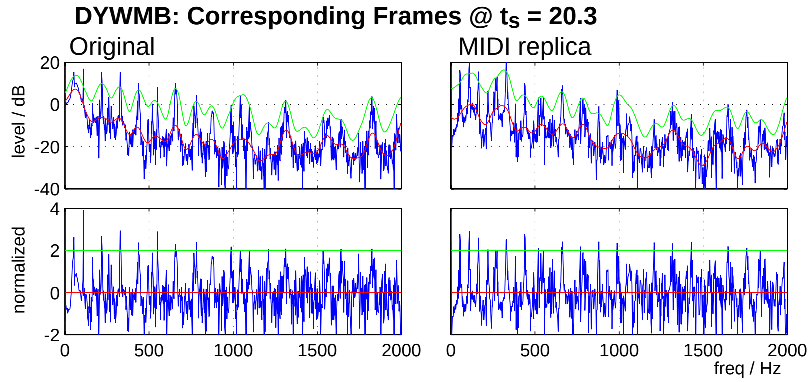
- **Pattern classifiers need training data**
  - i.e. need `{signal, note-label}` sets
  - i.e. MIDI transcripts of real music...already exist?



- **Idea: force-align MIDI and original**
  - can estimate time-warp relationships
  - recover accurate note events in real music!

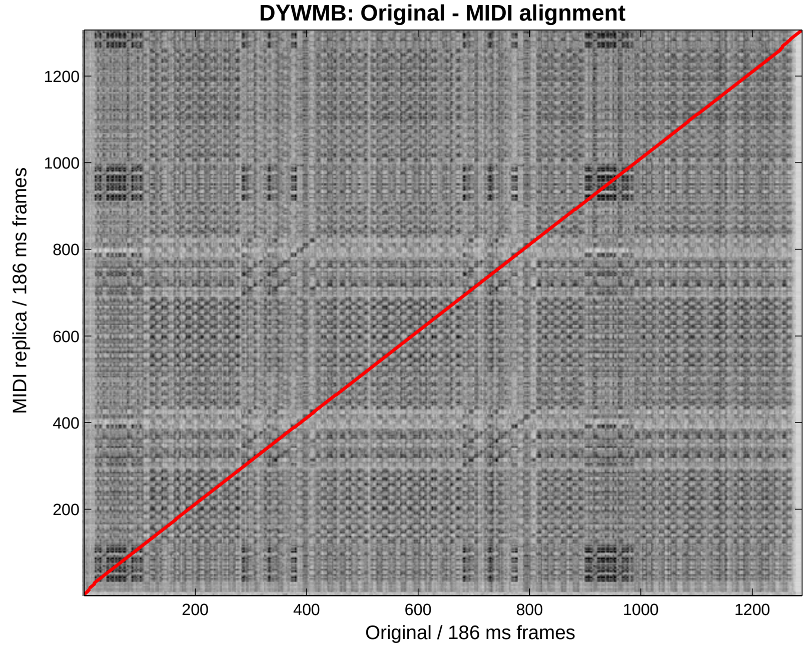
Features for MIDI alignments

- **Features that will match between MIDI replicas and original audio...**
- **Pitch is key attribute to match**
  - narrowband spectral features (but: timing...)
  - emphasize 100 Hz - 2 kHz
- **Local spectral variation, not absolute levels**
  - remove local average & normalize local range



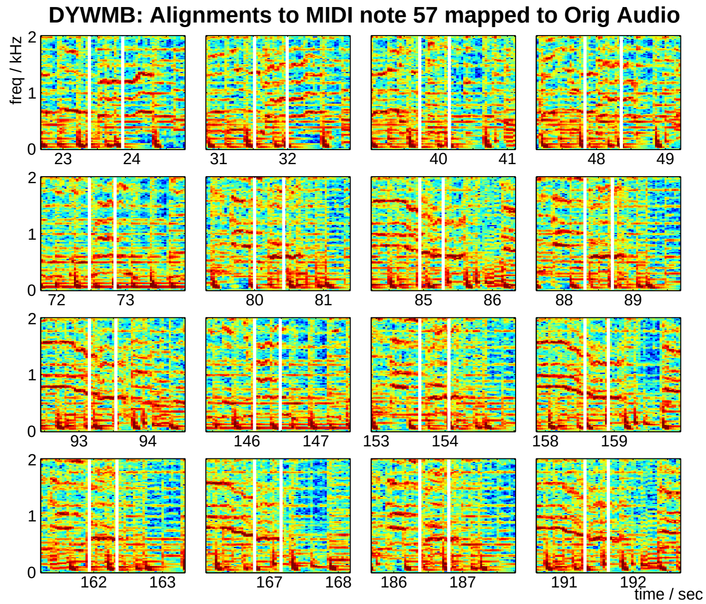
Alignment example

- **Inner-product distance on normalized spectral slices (8192 pt @ 22050 Hz):**



Extracted training data

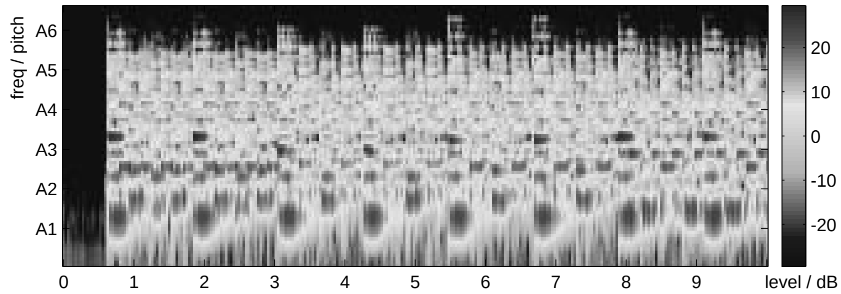
- **Want labeled** examples of notes **(in every context) to train pattern recognizer**
  - still perfecting alignment, but an example:



Polyphonic Piano Transcription

(Poliner & Ellis 2006)

- **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

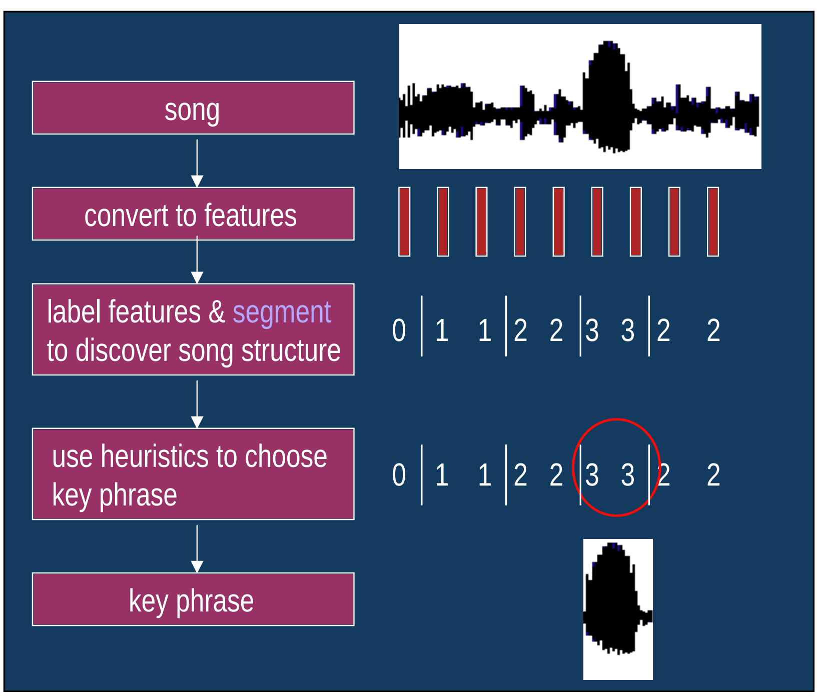
Outline

1. Music Transcription
2. **Music Summarization**
   - Segmentation
   - Identifying repetition
   - Evaluation
3. Music Information Retrieval
4. Music Similarity Browsing

2 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...

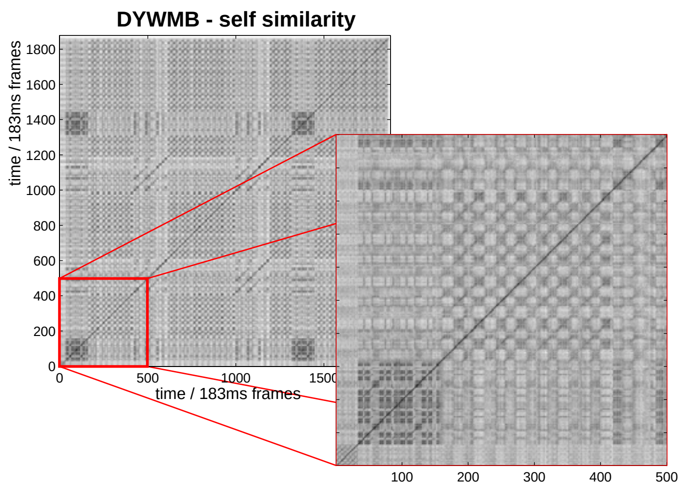
Summarization Approach



(with thanks to Beth Logan)

Segmentation

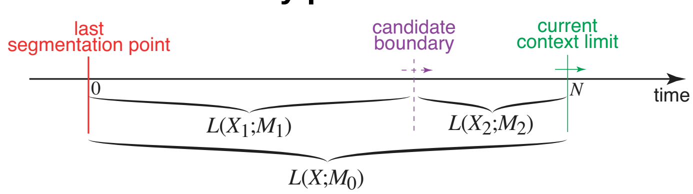
- **Find contiguous regions that are internally similar and different from neighbors**
- **E.g. "self-similarity" matrix (Foote 1997)**



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

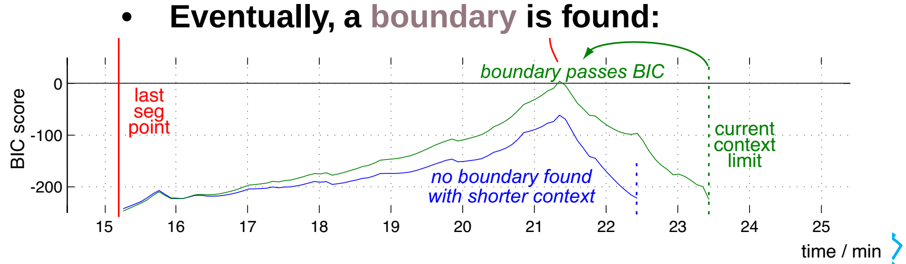
BIC segmentation

- **Want to use evidence from whole segment, not just local window**
- **Do 'significance test' on every possible division of every possible context**



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

- **Eventually, a boundary is found:**



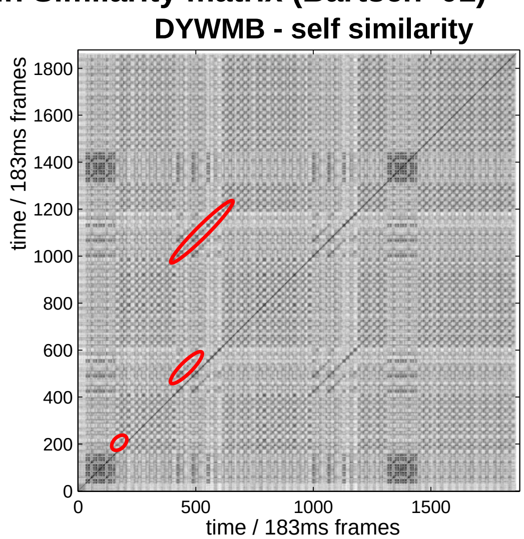
HMM segmentation

- **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? (Chu & Logan 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 1
```


Finding Repeats

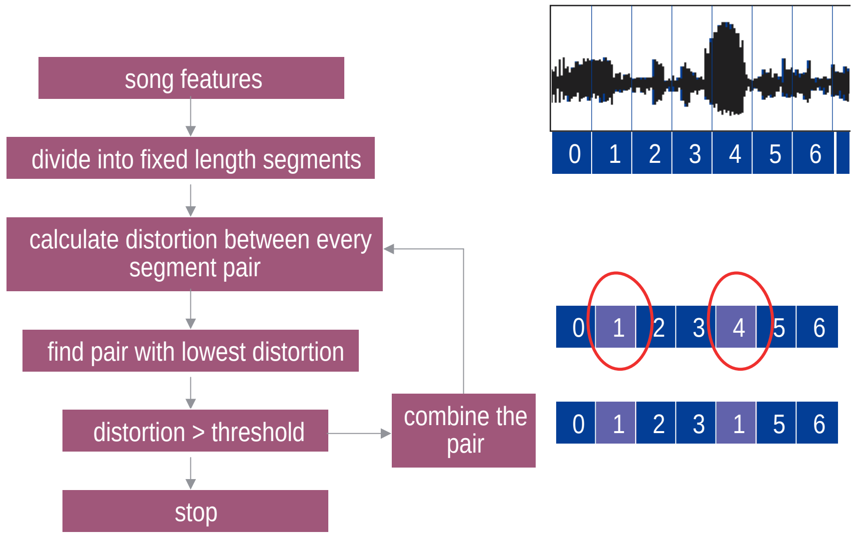
- **Music frequently repeats main phrases**
- **Repeats give off-diagonal ridges in Similarity matrix (Bartsch ’01)**



- **Or: clustering at phrase-level ...**

Clustering-based summarization
(Logan & Chu 2000)

- **Find segments in song by greedy clustering:**



- **Biggest cluster chosen as "key phrase"**
  - large contiguous block taken as example

Evaluating Summaries

- **Hard to evaluate:**
  What is the 'right answer'?
  - difficult to construct or judge a summary until you know the song...
- **Bartsch & Wakefield:**
  93 songs, 'chorus' hand-marked,
  70% frame-level precision-recall
  - aiming to find chorus/refrain
- **Chu & Logan:**
  18 Beatles #1 hits rated by 10 subjects
  as Good/Average/Poor
  - "significantly better than random"
- **Without a good metric, how to make choices to improve the algorithm?**

Outline

1. Music Transcription
2. Music Summarization
3. **Music Information Retrieval**
   - What it could mean
   - Unsupervised clustering
   - Learned classification
4. Music Similarity Browsing

3 Music Information Retrieval

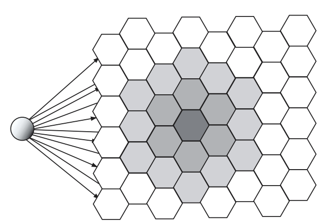
- **Text-based searching concepts for music?**
  - "musical Google"
  - finding a specific item
  - finding something vague
  - finding something new
- **Significant commercial interest**
- **Basic idea:**
  Project music into a space where neighbors are "similar"
- **(Competition from human labeling)**

Music IR: Queries & Evaluation

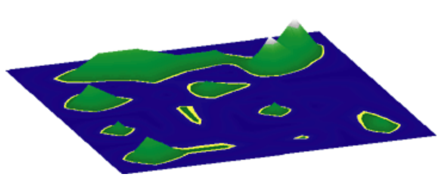
- **What is the form of the query?**
- **Query by Humming**
  - considerable attention, recent demonstrations
  - need/user base?
- **Query by noisy example**
  - "Name that tune" in a noisy bar
  - Shazam Ltd.: commercial deployment
  - database access is the hard part?
- **Query by multiple examples**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

Unsupervised Clustering
(Rauber, Pampalk, Merkl 2002)

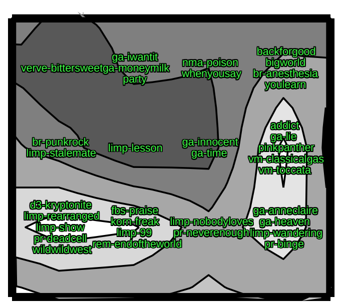
- **Map music into an auditory-based space**
- **Build 'clusters' of nearby → similar music**
  - "Self-Organizing Maps" (Kohonen)



- **Look at the results:**



"Islands of Music"

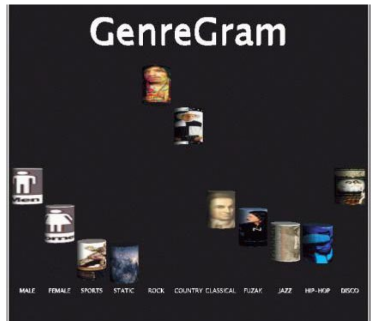


  - quantitative evaluation?

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”:**

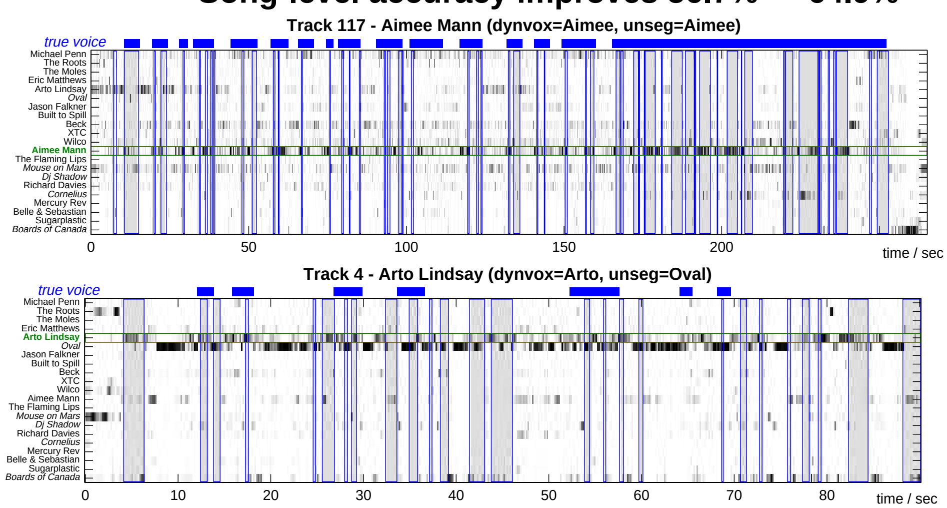


  - 9 spectral and 8 rhythm features every 200ms
  - 15 genres trained on 50 examples each, single Gaussian model → ~ 60% correct

Artist Classification

(Berenzweig et al. 2001)

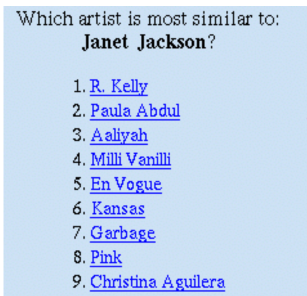
- **Artist label as available stand-in for genre**
- **Train MLP to classify frames among 21 artists**
- **Using only “voice” segments:**
  Song-level accuracy improves 56.7% → 64.9%



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

Outline

1. Music Transcription
2. Music Summarization
3. Music Information Retrieval
4. **Music Similarity Browsing**
   - Anchor space
   - Playola browser

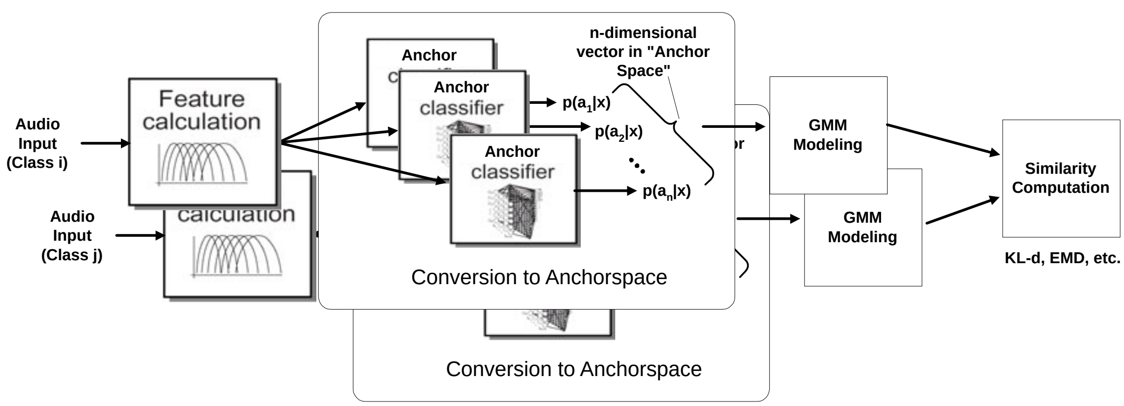
Music Similarity Browsing

4

- **Most interesting problem in music IR is finding new music**
  - is there anything on mp3.com that I would like?
- **Need a space where music/artists are arranged according to perceived similarity**
- **Particularly interested in little-known bands**
  - little or no 'community data' (e.g. collab. 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

Anchor space

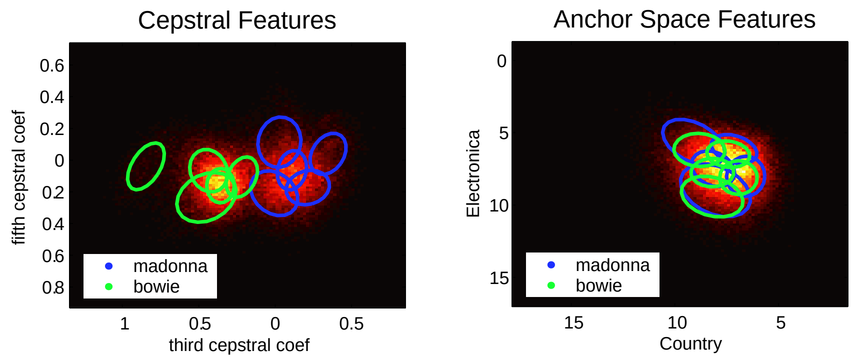
- **A classifier trained for one artist (or genre) will respond partially to a similar artist**
- **A new artist will evoke a particular pattern of responses over a set of classifiers**
- **We can treat these classifier outputs as a new feature space in which to estimate similarity**



- **"Anchor space" reflects subjective qualities?**

Anchor space visualization

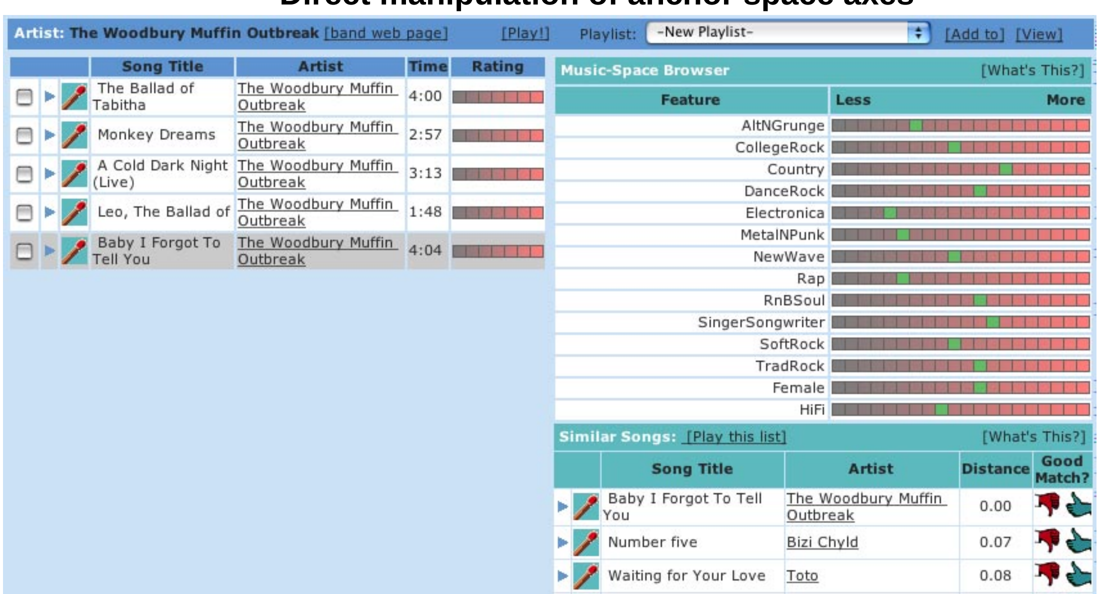
- **Comparing 2D projections of per-frame feature points in cepstral and anchor spaces:**



- each artist represented by 5GMM
- greater separation under MFCCs!
- but: relevant information?

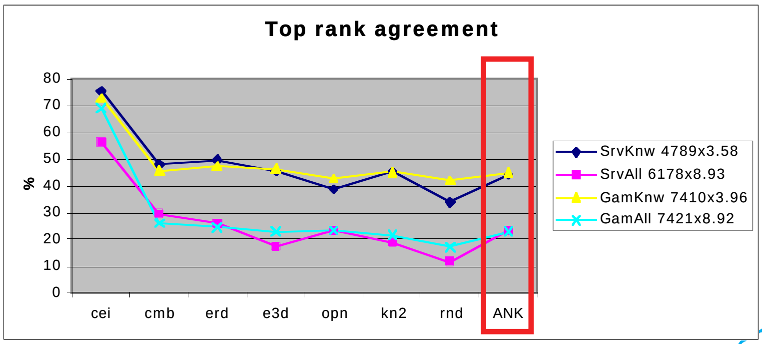
Playola interface (www.playola.org)

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



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 other similarity judgments**
  - e.g. musicseer data:



Summary

- **Music transcription:**
  Hard, but some progress
- **Music summarization:**
  New, interesting problem
- **Music IR:**
  Alternative paradigms, lots of interest

Data-driven machine learning techniques are valuable in each case

References

Mark A. Bartsch and Gregory H. Wakefield (2001) "To Catch a Chorus: Using Chroma-based Representations for Audio Thumbnailing", Proc. WASPAA, Mohonk, Oct 2001.
http://musen.engin.umich.edu/papers/ bartsch_wakefield_waspaa01_final.pdf

A. Berenzweig, D. Ellis, S. Lawrence (2002). "Using Voice Segments to Improve Artist Classification of Music ", Proc. AES-22 Intl. Conf. on Virt., Synth., and Ent. Audio. Espoo, Finland, June 2002.
http://www.ee.columbia.edu/~dpwe/pubs/aes02-aclass.pdf

A. Berenzweig, D. Ellis, S. Lawrence (2002). "Anchor Space for Classification and Similarity Measurement of Music", Proc. ICME-03, Baltimore, July 2003.
http://www.ee.columbia.edu/~dpwe/pubs/icme03-anchor.pdf

J. Foote (1997), "A similarity measure for automatic audio classification", Proc. AAAI Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora, March 1997.
http://citeseer.nj.nec.com/foote97similarity.html

Masataka Goto (2001), "A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation using EM Algorithm for Adaptive Tone Models", Proc. ICASSP 2001, Salt Lake City, May 2001.
http://staff.aist.go.jp/m.goto/PAPER/ICASSP2001goto.pdf

A. Klapuri, T. Virtanen, A. Eronen, J. Seppänen (2001), "Automatic transcription of musical recordings", Proc. CRAC workshop, Eurospeech, Denmark, Sep 2001.
http://www.cs.tut.fi/sgn/arg/klap/crac2001/crac2001.pdf

Beth Logan and Stephen Chu (2000), "Music summarization using key phrases", Proc. IEEE ICASSP, Istanbul, June 2000.
http://crl.research.compaq.com/publications/techreports/reports/2000-1.pdf

References (2)

G. Peeters, A. La Burthe, X. Rodet (2002), "Toward automatic music audio summary generation from signal analysis", Proc. ISMIR-02, Paris, October 2002.
http://ismir2002.ircam.fr/proceedings%5C02-FP03-3.pdf

Andreas Rauber, Elias Pampalk and Dieter Merkl (2002), "Using Psychoacoustic models and Self-Organizing Maps to create a hierarchical structuring of music by musical styles", Proc. ISMIR-02, Paris, October 2002.
http://ismir2002.ircam.fr/proceedings%5C02-FP02-4.pdf

G. Tzanetakis, G. Essl, P. Cook (2001), "Automatic Musical Genre Classification of Audio Signals", Proc. ISMIR-01, Bloomington, October 2001.
http://ismir2001.indiana.edu/pdf/tzanetakis.pdf

P.J. Walmsley, S.J. Godsill, and P.J.W. Rayner (1999), "Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters", Proc. WASPAA, Mohonk, Oct 1999.
http://www.ee.columbia.edu/~dpwe/papers/WalmGR99-polypitch.pdf

Brian Whitman and Paris Smaragdis (2002), "Combining Musical and Cultural Features for Intelligent Style Detection", Proc. ISMIR-02, Paris, October 2002.
http://ismir2002.ircam.fr/proceedings%5C02-FP02-1.pdf