# Ground Truth Transcriptions of Real Music from Force-Aligned MIDI Syntheses

Robert J. Turetsky and Daniel P. W. Ellis
{rob, dpwe}@ee.columbia.edu
Columbia University, LabROSA
Dept. of Electrical Engineering

# MIDIAlign: Talk Organization

- The Polyphonic Transcription Problem
- Proposed Solution: MIDI Alignment
- Method and Experiments
- Time to evaluate the situation
- Immediate and Future Directions

# The Polyphonic Transcription Problem

- Simple Melodies: Easy to transcribe


toms_diner.mp3

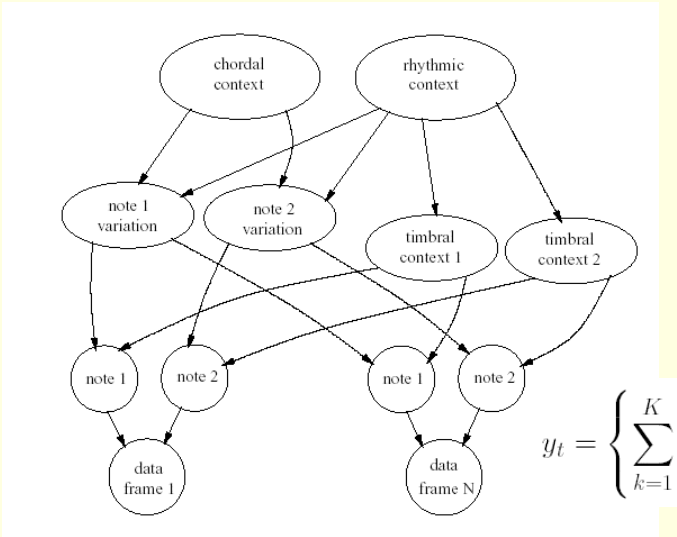
- Complicated Melodies:
  - Layers of instrumentation
  - Vocals and special effects
  - Harmonic interference
  - The rhythm section

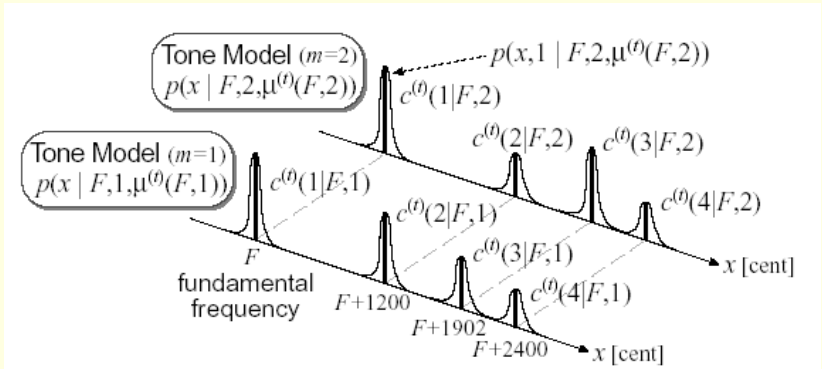

School.mp3

- Only “trained ears” with exposure to large amounts of music and music theory can transcribe rich real-world polyphonic music

# The Polyphonic Transcription Problem

- Feature-based methods for simple (e.g. monophonic) melodies
- For real-world transcription need statistical pattern recognition framework



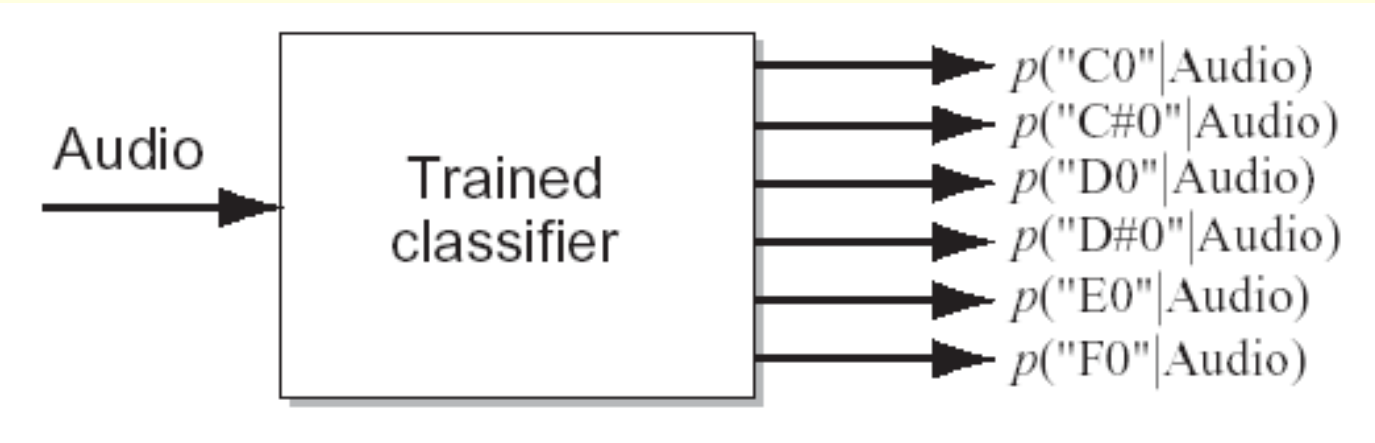




$$y_t = \left\{ \sum_{k=1}^{K} \sum_{m=1}^{M_k} \sum_{i=1}^{I} a_{k,m,i}\phi_{i,t} \cos\left[(m+\delta_{k,m})\omega_{0,k}t\right] + b_{k,m,i}\phi_{i,t} \sin\left[(m+\delta_{k,m})\omega_{0,k}t\right] \right\} + v_t$$

Walmsley et al 1999, Goto 2001

# The "Black-Box" Approach

- We could use the signal as evidence for pitch class in a black-box classifier




- To train and evaluate these methods, we need large corpus of labeled note examples!
- N.E.R.: Note Error Rate for evaluation

# Labeling Music: Prior Work

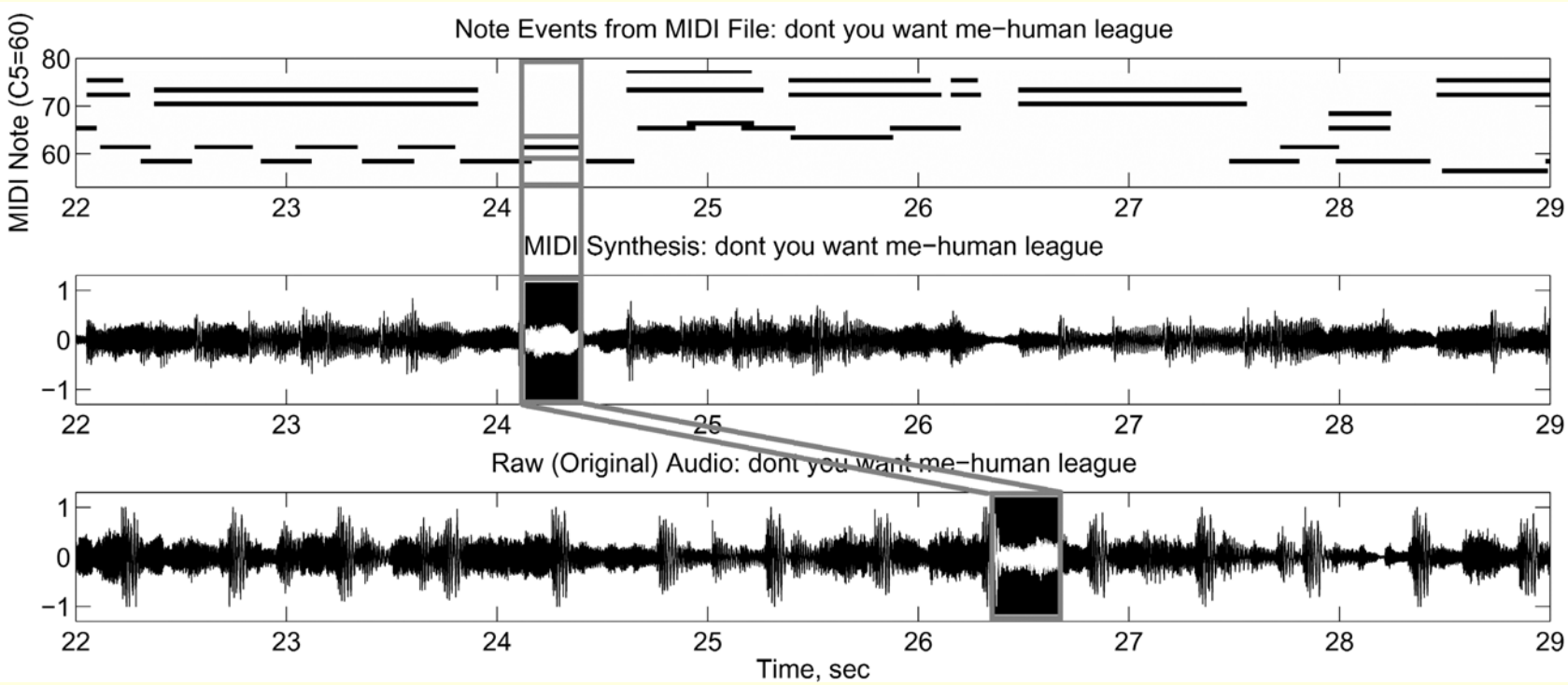
- Problem: Need labels for training and evaluation, but most music transcriptions not available!
- Solution 1: Sheet music
  - Only gives “flavor” of popular music
- Solution 2: Synthesized MIDI files
  - 128 General MIDI instruments cannot characterize the full “sonic range” of modern studios
- Solution 3: Record MIDI tracks directly from studio
  - Tracks extremely difficult to come by
  - Popular music often produced with samples

# The Trouble with MIDI Transcriptions

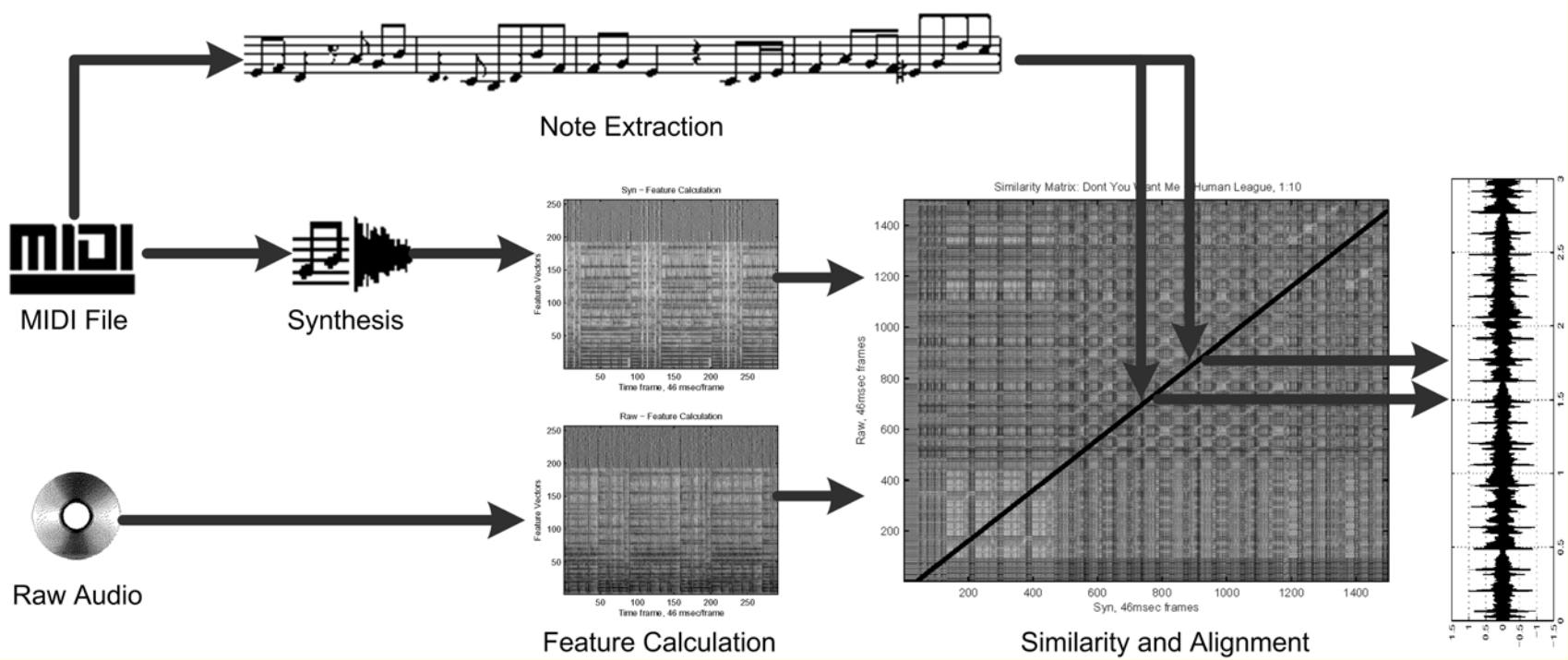
- MIDI transcriptions are freely available, but:
- MIDI files have transcriptions of music but do not accurately represent music signal
  - Can only (poorly) approximate vocals, timbres
- Transcriptions are an interpretation of the original work by an amateur:
  - Does not respect precise timing and duration of original song
  - May not be accurate transcription
    - Transposition, phrase sequence, different versions, possible melody characterization

# Our Solution: MIDI Alignment

- Goal: To map each note event from MIDI transcription (syn) into window of original song (raw)



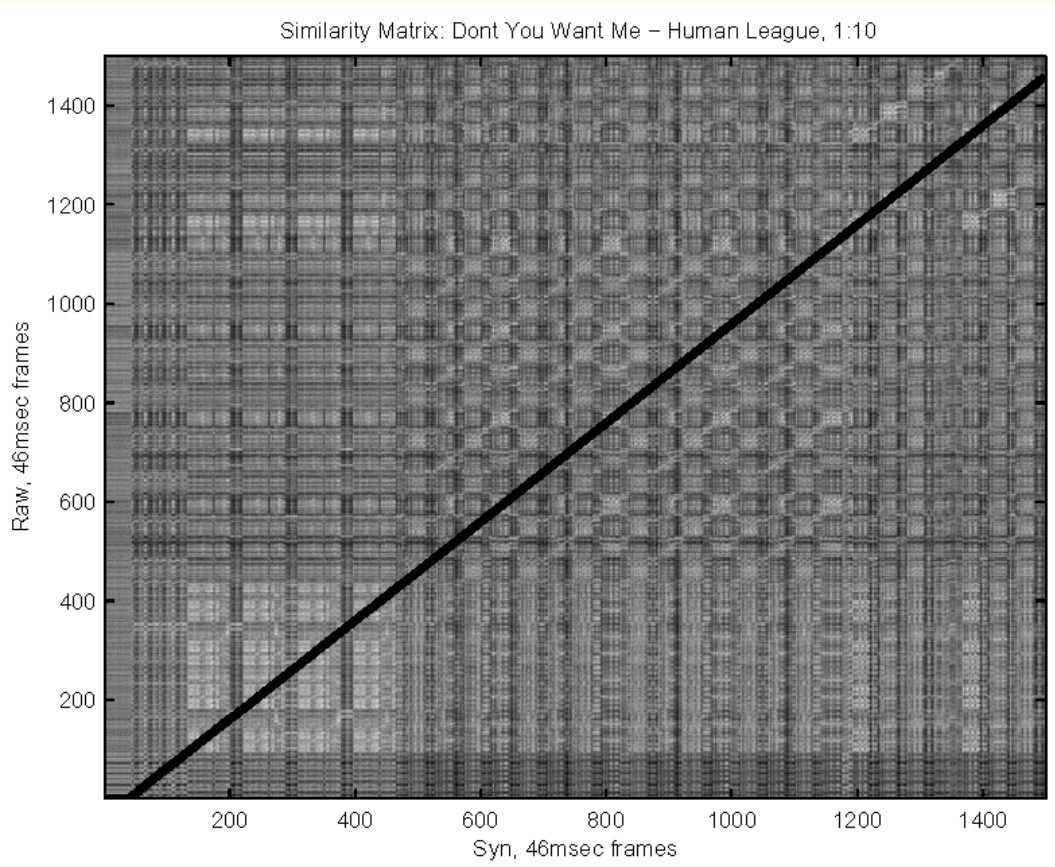
# The MIDIAlign Method



# MIDIAlign: Feature Selection

- Want to match pitches present but not timbres ➔ spectral features, not MFCC
- Base Feature: FFT bins up to 2.8 kHz
- Added Features:
  - spec_power {‘dB’, 0.3, 0.5, 0.7, 1, 2}
  - diff_points {0, [1 16], [1 64], [224 256]}
    Highlight note onset/offset times
  - freq_diff {0, 1}
    Focus on local harmonic structure
  - noise_supp {0, 1}
    Attempt to remove noise from drums and fx

# Alignment: The Similarity Matrix

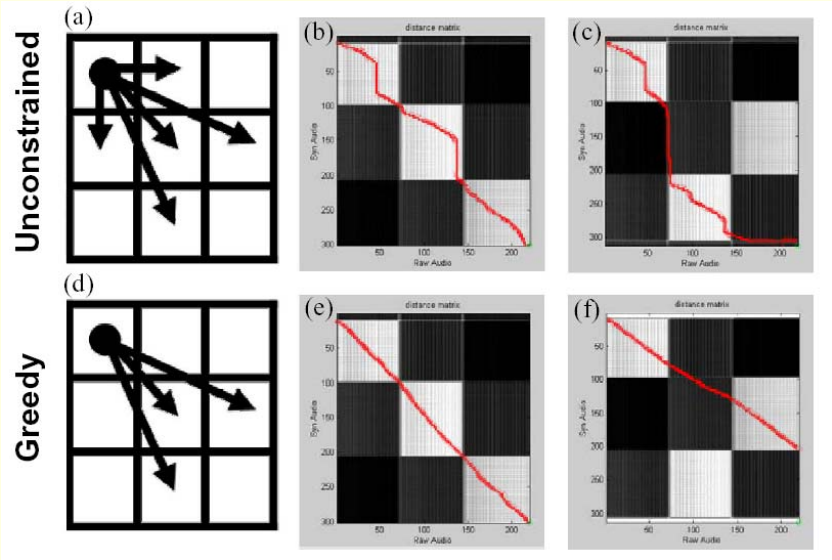



$$SM(i,j) = \frac{spec_{raw}(i)^T spec_{syn}(j)}{|spec_{raw}(i)||spec_{syn}(j)|},$$
$$\text{for } 0 \leq i < N, \quad 0 \leq j < M$$

- Each point in Similarity Matrix (SM) is cos distance between frames in raw/syn
- By searching for "best path" in SM we can find mapping from syn to raw
- Based on Foote, 1999

# DP: Finding the Best Path

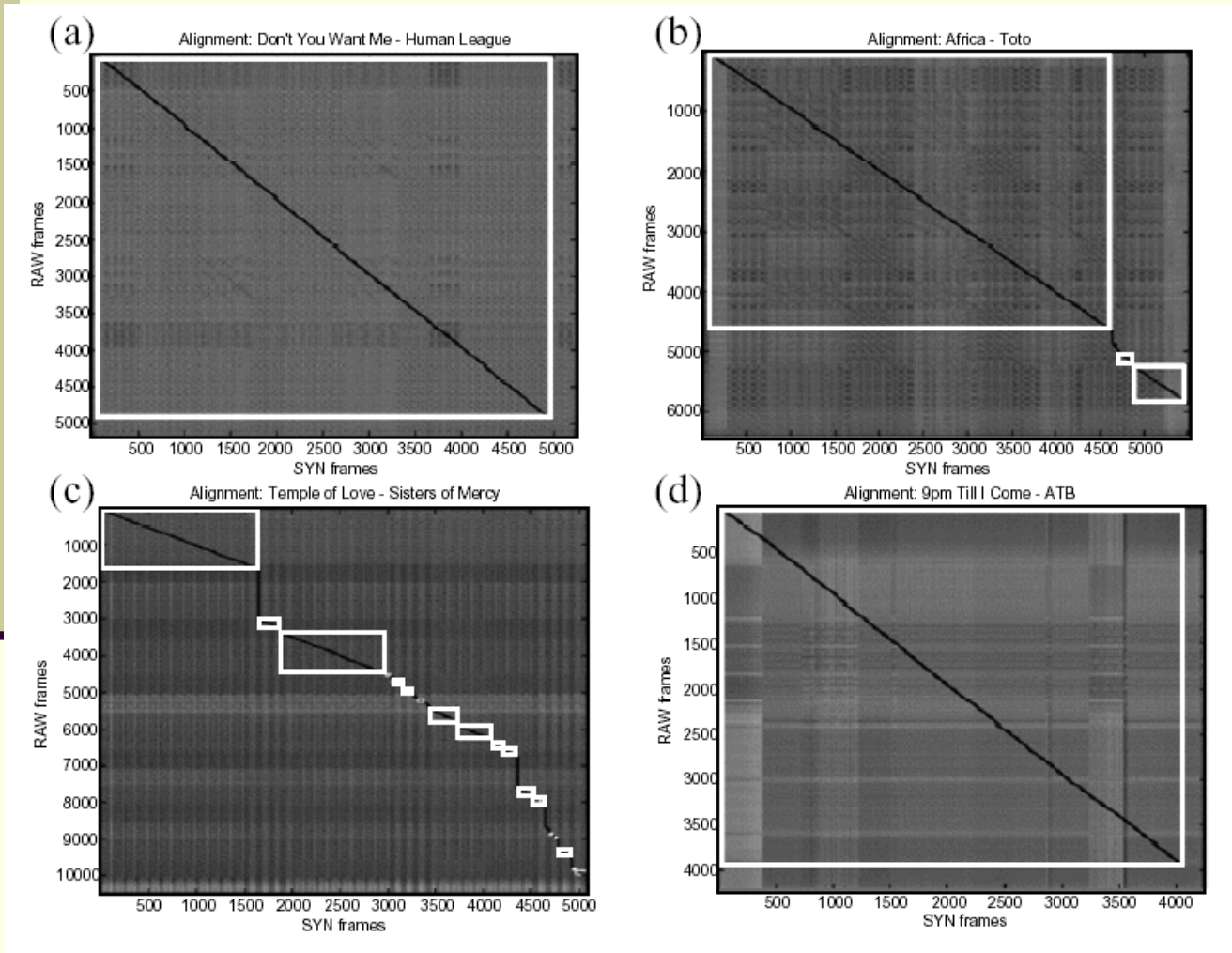
- Best warping function is the “best path” through the similarity matrix
- SM can be treated like dynamic programming (DP) distance matrix to search for best path
- Two DP flavors: unconstrained vs. greedy




Unconstrained DP can navigate around missed notes, but gives a rough alignment

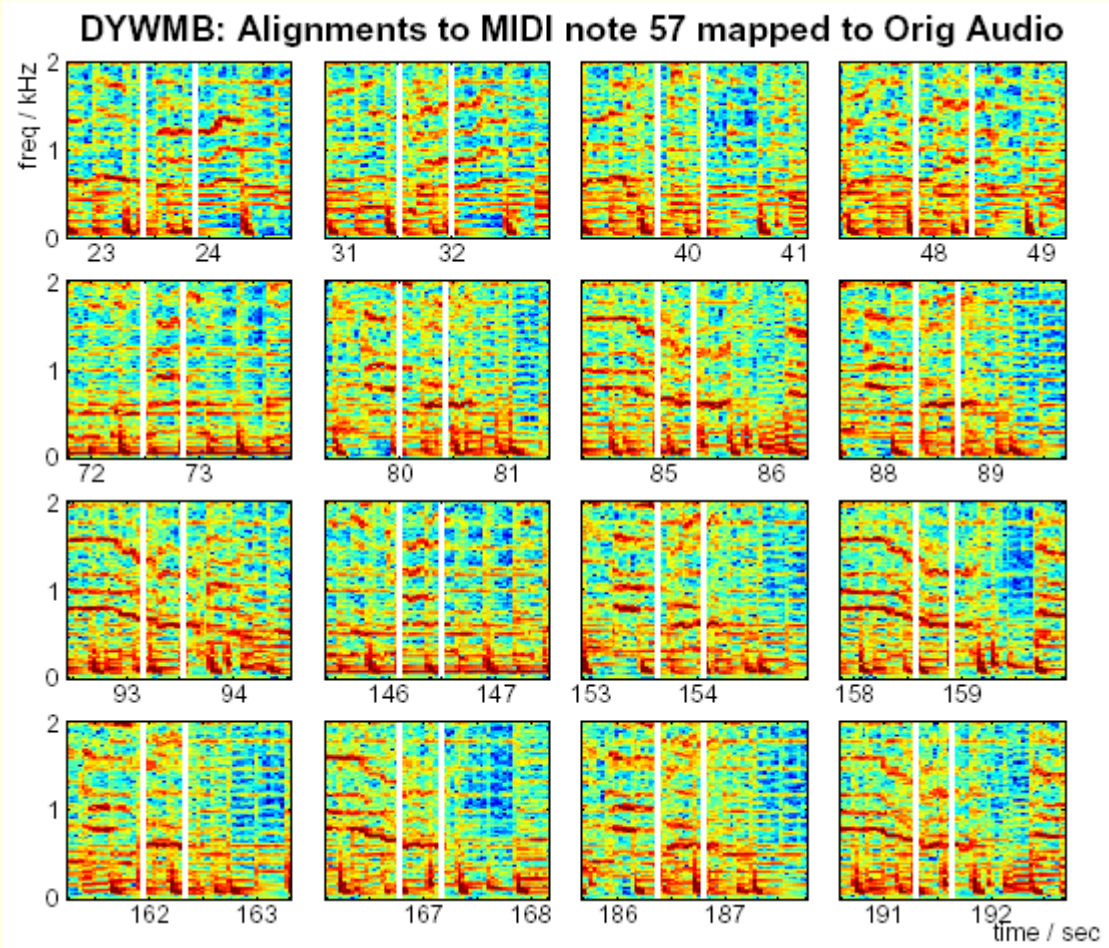
Greedy DP has smooth structure when notes are transcribed correctly, but is catastrophic with errors!

# Two Stage Alignment: Getting the Best of Both Worlds




- Unconstrained DP to find "accurate" regions that align well
- Use greedy DP to smooth out rough alignment estimates for final mapping
- Median filter can find "straight" segments

# MIDIAlign Examples




Dywmb_ex.mp3

Temple.mp3

Dont_speak_remidi.mp3

Dywmb-mn57.wav

More examples: http://www.ee.columbia.edu/~rob/midialign

# MIDIAlign Experiments

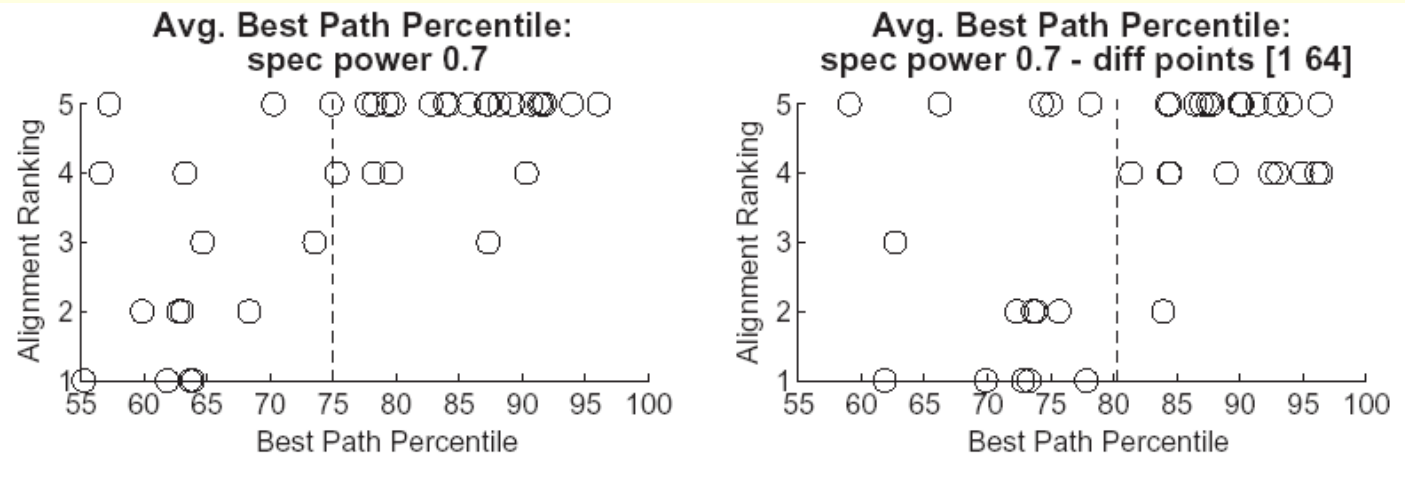
- Set of 30 songs from different genres
- MIDI files downloaded from user sites indexed on http://www.musicrobot.com
- MIDI synthesized using WAVmaker III
- Notes extracted using Perl::MIDI
- Audio files taken from local music library or filesharing services
- All analysis done at 22.05 kHz
- Alignments for all feature values evaluated aurally (1..5) to derive best featureset.

# MIDIAlign: Results

| Featureset | Score 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| spec_power = 0.5 | 6 | 2 | 7 | 6 | 15 |
| spec_power = 0.7 | 6 | 3 | 4 | 10 | 14 |
| spec_power = 1.0 | 6 | 4 | 13 | 5 | 9 |
| spec_power = 0.7, diff_points = [1 16] | 5 | 6 | 4 | 11 | 13 |
| spec_power = 0.7, diff_points = [1 64] | 6 | 7 | 0 | 4 | 21 |
| spec_power = 0.7, freq_diff = 1 | 7 | 5 | 6 | 11 | 4 |
| spec_power = 0.7, noise_supp = 1 | 28 | 1 | 0 | 1 | 0 |

# Evaluating the Alignments

- In order to develop large corpus, we need many alignments
- Many MIDI files available on www are of marginal quality from dubious sources
- Automatic evaluation of alignments on MIDI files spidered from web is imperative
- Most discriminatory Evaluation Metric: average “best path” distance percentile amongst entire SM



# Future Directions

- Improve alignment features: e.g. overtones
- Account for decay after “note off” MIDI event
- Evaluate corpus generated from web spider
  - 35 hours of music with 1.5 million note events
- “Proof of concept” classifier to perform pitch extraction based on training data only
- Improve classifiers with music models similar to Walmley, et al 1999 based on statistical analysis of MIDI files
- Evaluate the effect of studio effects (e.g. reverb) on classifiers built with aligned labels