Chord Recognition and Segmentation using EM-trained Hidden Markov Models

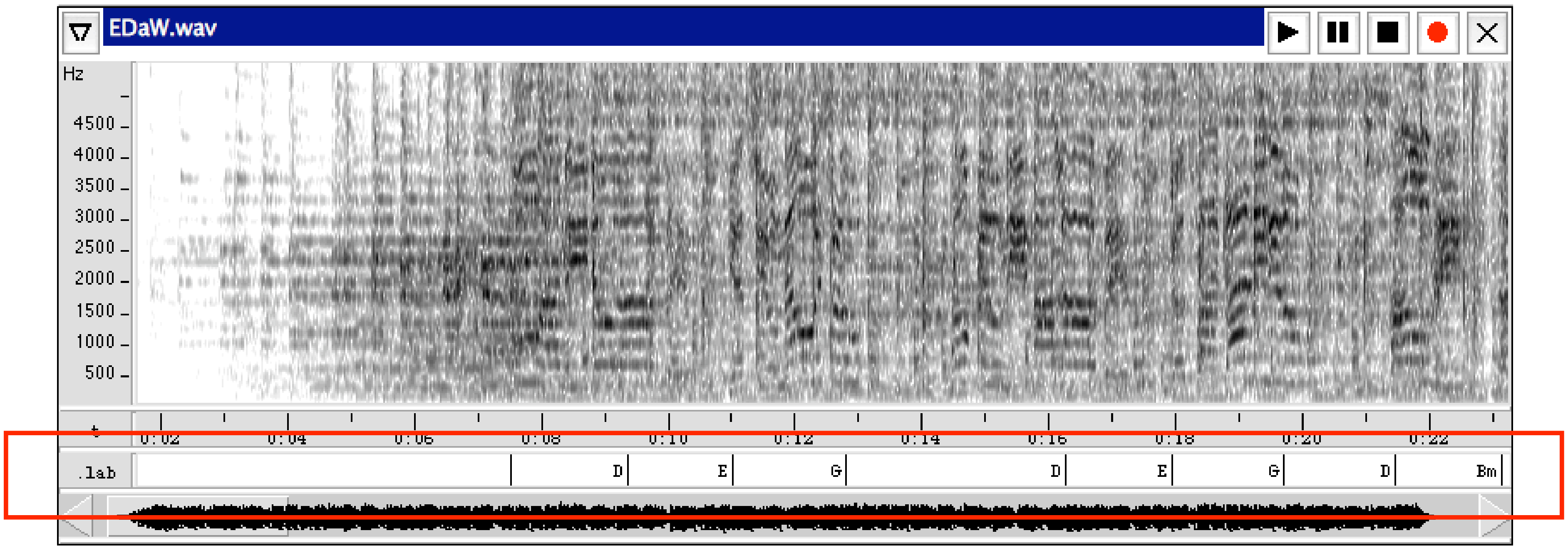
- Chord Recognition
- EM-trained HMMs
- Experiments & Results

Alex Sheh & Dan Ellis
{asheh79,dpwe}@ee.columbia.edu
LabROSA, Columbia University

Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Chord Transcription

- Basic problem:
  Recover chord sequence labels from audio



  - Easier than note transcription ?
  - More relevant to listener perception ?

Difficulties with Chord Transcription

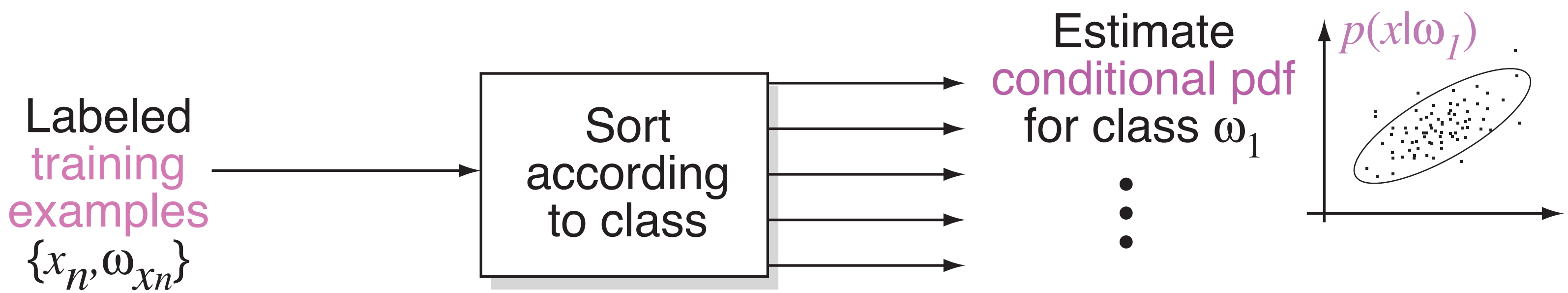
- Enharmonicity:
  Chord labels can be ambiguous
  - C# vs Db
- Many different chord classes
  - major, minor, 6th, 9th, ...
  - fold into 7 'main' classes:
    maj, min, maj7, min7, dom7, aug, dim
- Acoustic variability
  - chords are the same regardless of instrumentation

Approaches to Chord Transcription

- Note transcription, then note→chord rules
  - like labeling chords in MIDI transcripts
- Spectrum→chord rules
  - i.e. find harmonic peaks, use knowledge of likely notes in each chord
- Trained classifier
  - don't use any "expert knowledge"
  - instead, learn patterns from labeled examples

Statistical-Pattern-Recognition Chord Recognizer

- Use labeled training examples to estimate $p(x|\omega_i)$ for features x and chord class ω_i



-

- Use Bayes Rule to get posterior probabilities for each class given features:

$$p(\omega_i|x) = \frac{p(x|\omega_i) \cdot p(\omega_i)}{\sum_j p(x|\omega_j) \cdot p(\omega_j)}$$

Training Data Sources

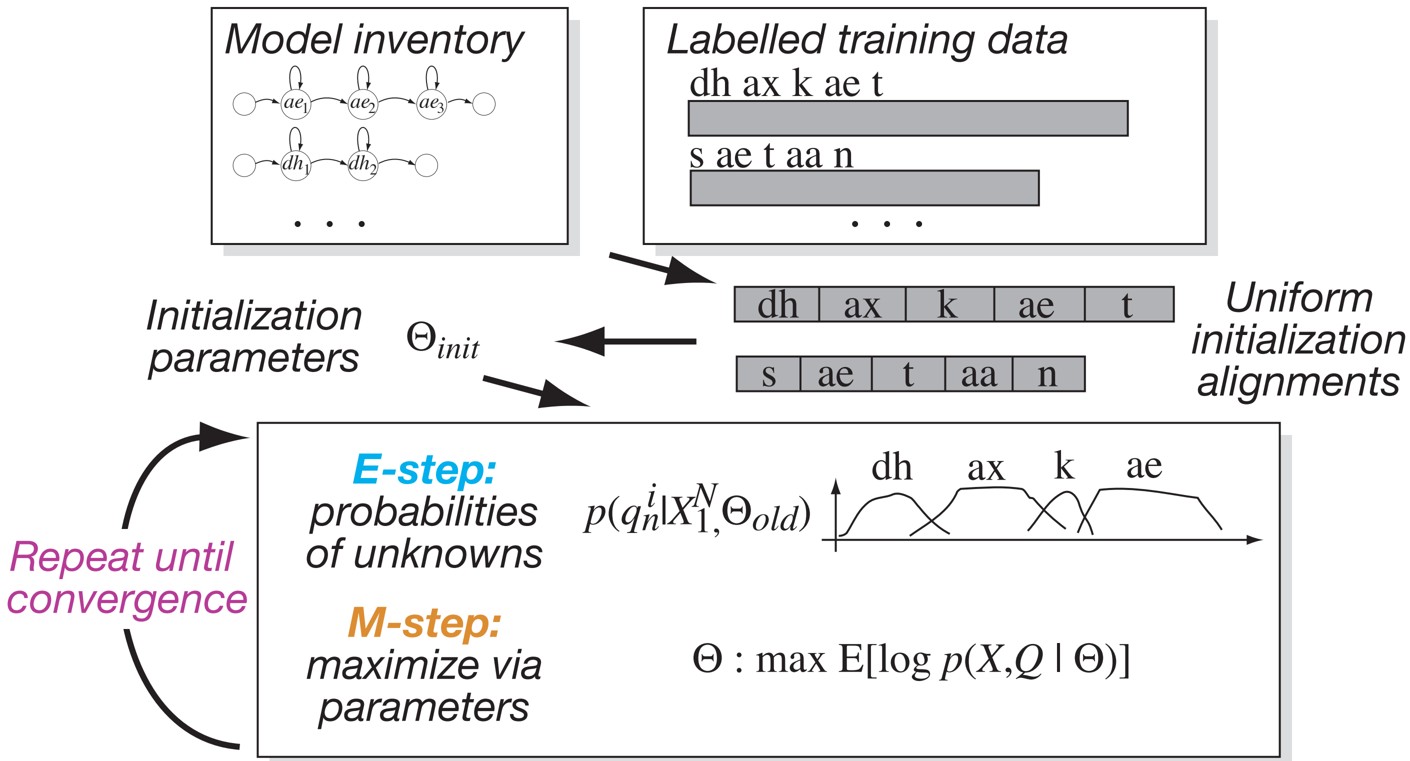
- **We need (lots of) examples of audio segments and the appropriate chord labels**
  - Not widely available!
  - Even when you can get it, there is very little
- **We could hand-mark a training set**
  - painfully time-consuming!
- **Can we generate it automatically?**
  - we could, if we already had the chord transcriptions system working...

Speech Recognition Analogy

- Chord recognition is like word recognition
  - we are trying to recover a sequence of 'exclusive' labels associated with an audio stream
  - we have a lot of potential training audio, but no time labels
- Can we do what they do in ASR?
  - .. i.e. iterative re-estimation of models and labels using Expectation-Maximization
  - Need only label sequence, not timings (i.e. words or chords in order, no times)

EM HMM Re-Estimation

- Estimate 'soft' labels using current models
- Update model parameters from new labels
- Repeat until convergence to local maximum



Chord Sequence Data Sources

- All we need are the chord sequences for our training examples
- OLGA Tab archives (http://www.olga.net/)?
  - multiple authors, unreliable quality
- Hal Leonard "Paperback Song Series"
  - many Beatles songs, consistent detail
  - manually retyped for 20 songs:
    "Beatles for Sale", "Help", "Hard Day's Night"
- Issues:
  - repeats, intros, weird bits in the middle

Experiments

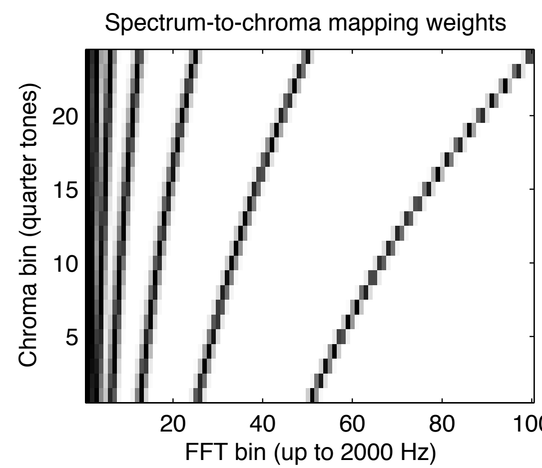
- Preliminary investigations to see if this works
  - small database to get started
  - compare different feature sets
  - different ways to evaluate results
  - can we reintroduce a little high-level knowledge?

Training/Test Conditions

- **Two training sets:**
  - Train on 18 songs, test on 2 held-out songs (fair test)
  - Train on all 20 songs, test on 2 songs from training set (cheating, rewards over-fitting, upper bound)
- **Two evaluation measures**
  - Recognition: transcribe unknown chords
  - Forced alignment: find time boundaries given correct chord sequence
  - Score frame-level accuracy against hand-markings

Features

- Can try any features with EM training...
- Use MFCCs as baseline
  - "the" feature in ASR, also useful in music IR
  - also try deltas, double deltas as in ASR
  - capture 'formants', not pitch - straw man
- "Pitch Class Profile" features (Fujishima'99)
  - collapse FFT bin energies into (24) chroma bins
  - a/k/a Chroma Spectrum (Bartsch'01) ...



Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Averaging Rotated PCP Models

- **Statistical system learns separate models for each of (7 chord types) x (21 roots)**
  - models are means, variances of feature vectors
  - only 32 actually appear in our training set
  - even so, many have few training instances
- **Expect similarity between e.g. Amaj & Bmaj**
  - same chord, just shifted in frequency
  - shift = rotation of 24-bin chroma space
- **Can align & average all transpositions of same chord after each training iteration**
  - then rotate back to starting chroma, continue...

Results: Recognition

- Models are not adequately discriminant:

Percent Frame Accuracy: Recognition

Feature	Recog train18	Recog train20
MFCC	5.9 7.7	16.7 19.6
MFCC_D	15.8 1.5	7.6 6.9
PCP	10.0 18.2	23.6 26.4
PCP_ROT	23.3 20.1	23.1 13.1

Eight Days a Week
Every Little Thing

MFCCs are poor (can overtrain)

PCPs a little better (ROT helps generalization)

(random ~3%)

Results: Alignment

- Some glimmers of hope!

Percent Frame Accuracy: Forced Alignment

Feature	Align train18	Align train20
MFCC	27.0 14.5	20.9 23.0
MFCC_D	24.1 19.9	13.1 19.7
PCP	26.3 46.2	41.0 53.7
PCP_ROT	68.8 83.3	68.3 83.8

Eight Days a Week
Every Little Thing

PCP_ROT
best accuracy
& best generalization

Chord Confusions

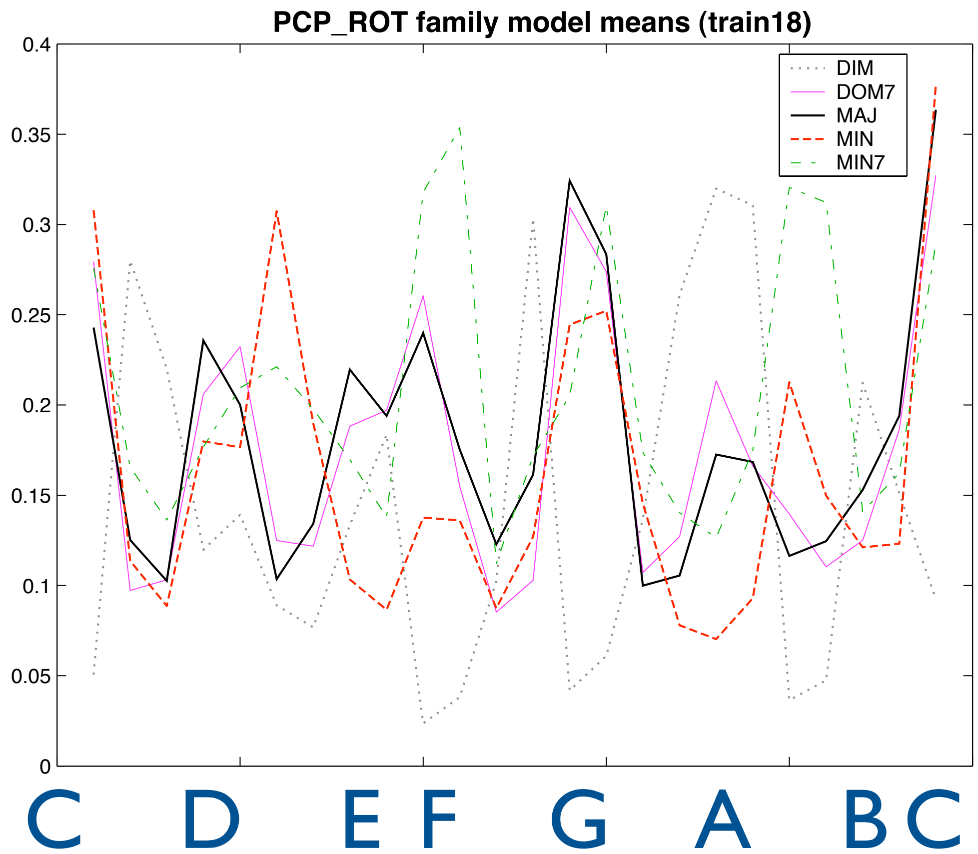
"E MAJOR" CONFUSION MATRIX
Eight Days a Week

	Maj	Min	Maj7	Min7	Dom7	Aug	Dim
E/Fb	158	115	.	9	.	.	.
E#/F	9	.	.	.	.	.	.
F#/Gb	.	.	.	11	.	.	.
G	3	.	.	.	.	.	.
G#/Ab	.	.	.	.	.	.	.
A	9	.	.	.	1	.	.
A#/Bb	.	.	.	.	.	.	.
B/Cb	8	.	.	.	,	.	.
B#/C	.	.	.	.	.	.	.
C#/Db	.	20	.	.	.	.	.
D	.	14	.	.	.	.	.
D#/Eb	.	.	.	.	.	.	.

- Major/minor confusions
- C# is relative minor (shared notes)

What did the models learn?

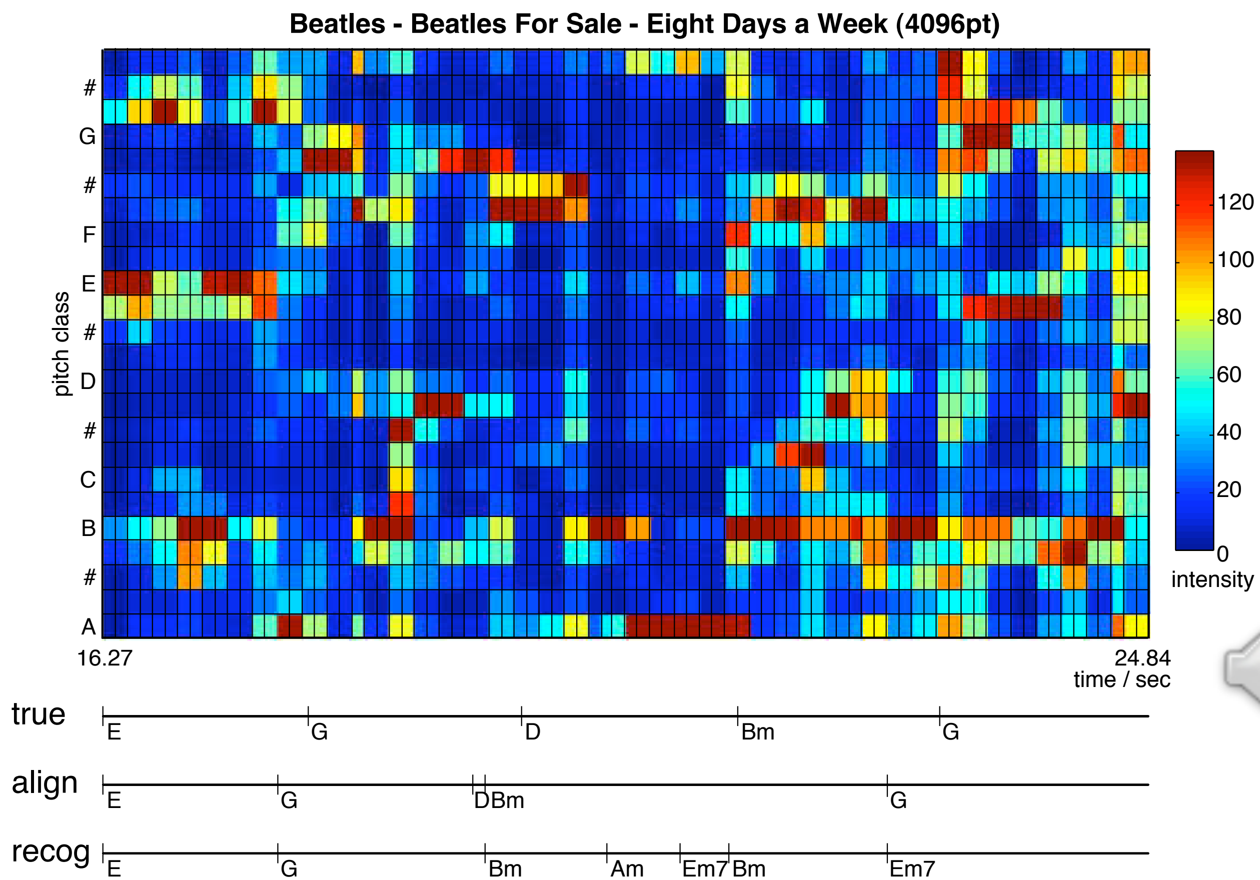
- Chord model centers (means) indicate chord ‘templates’:



(for C-root chords)

Recognition/Alignment Example

- A flavor of the features & results:



Conclusions & Future Work

- ASR-style EM is a viable approach for learning chord models
  - more training data needed
- Better features
  - capture 'global' properties of chords
  - robust to fine tuning issues?
- Better representation for training labels
  - i.e. allow for extra repeats?
- More ways to reintroduce music knowledge