Music Understanding

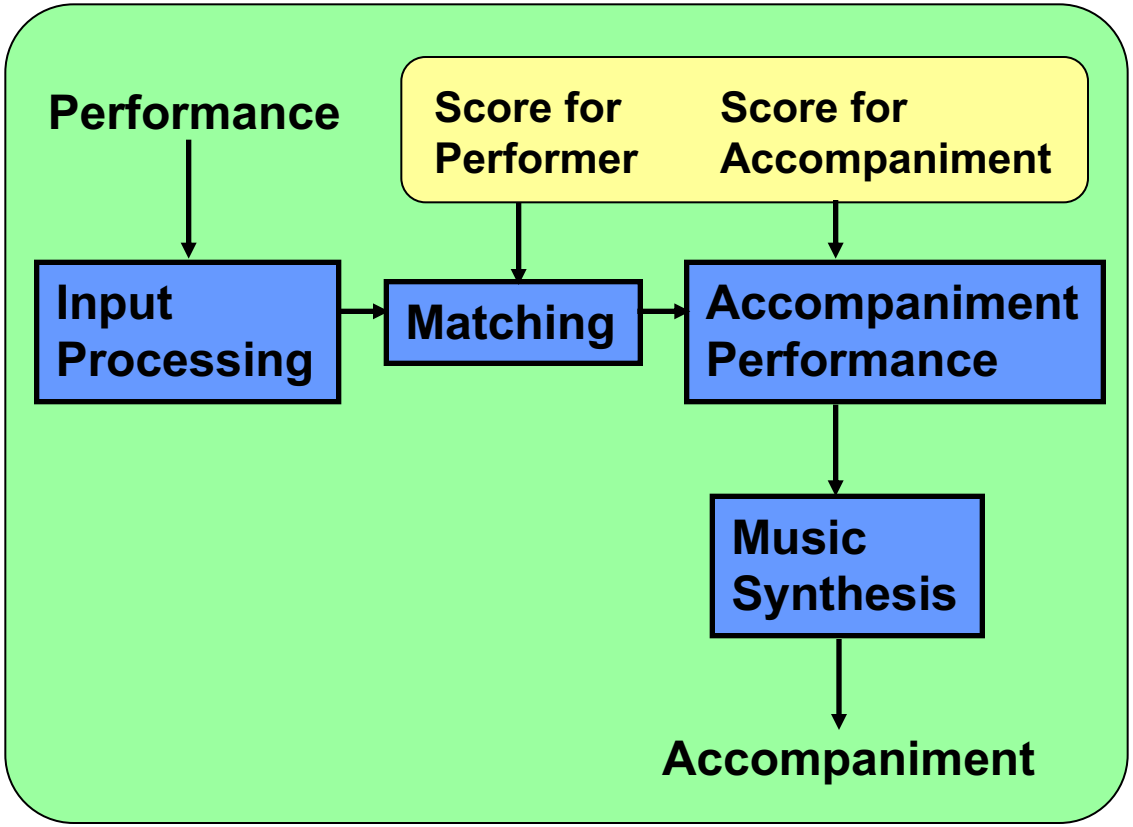
Roger B. Dannenberg
School of Computer Science

Music Understanding

- *Music Understanding*: Recognition of Pattern and Structure in Music
- Surface structure:
  - Pitch – Loudness
  - Harmony – Notes
- Deep structure:
  - Phrase relationships
  - Score following
  - Emotion
  - Expressive performance

Accompaniment Video

Computer Accompaniment

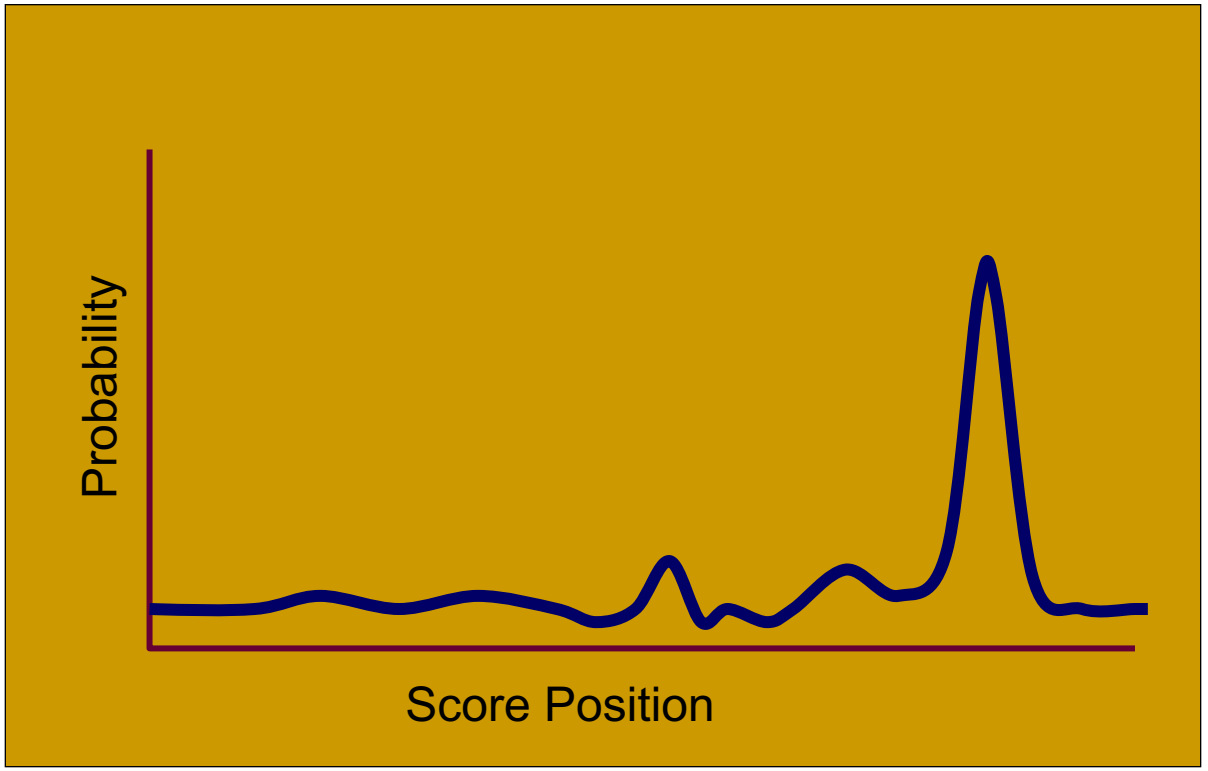


Vocal Accompaniment

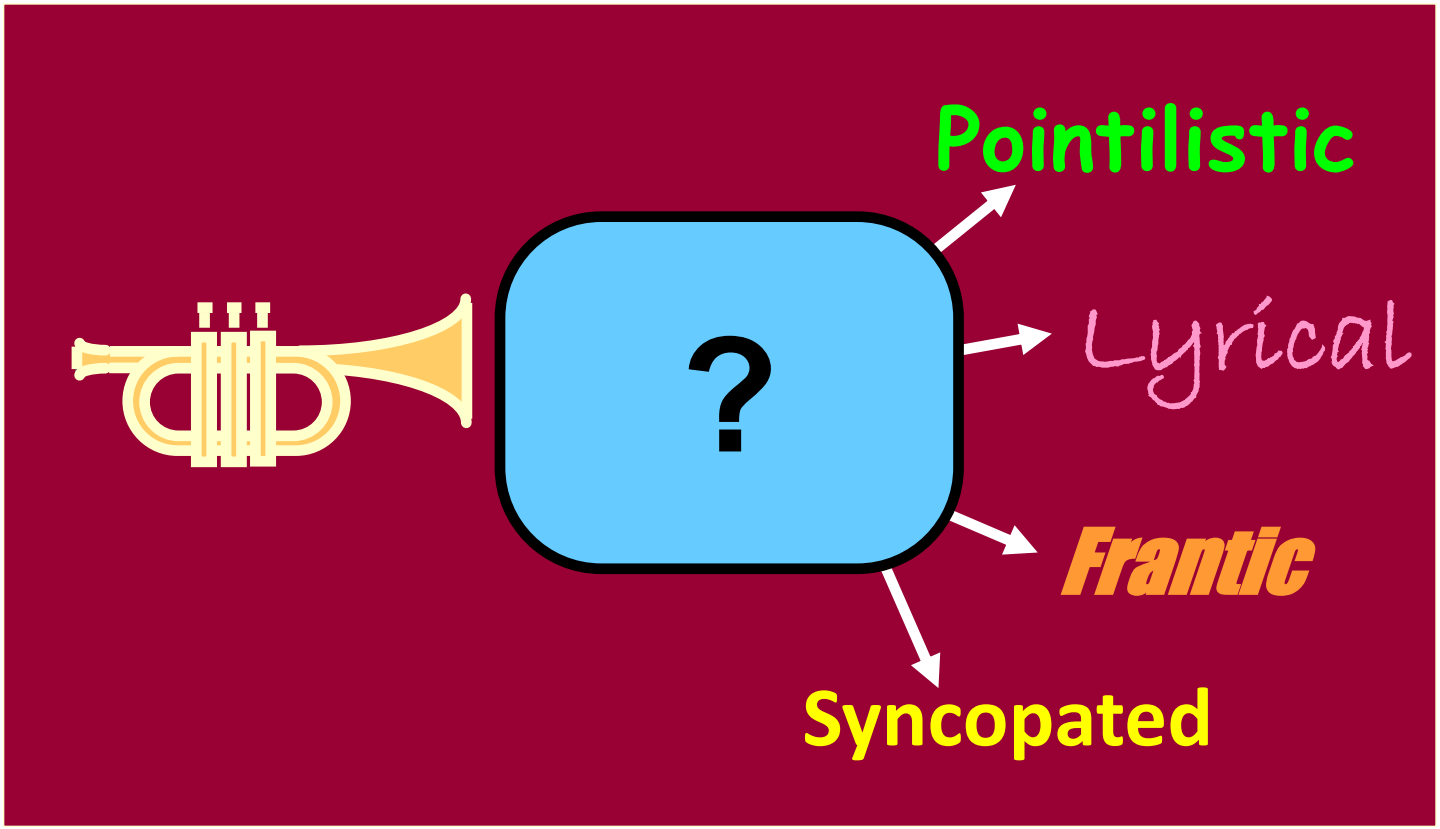
- Lorin Grubb's Ph.D. (CMU CSD)
- Machine learning used to:
  - Learns what kinds of tempo variation are likely
  - Characterize sensors
    - When is a notated G sensed as a G#?
- Machine learning necessary for good performance

Vocal Accompaniment

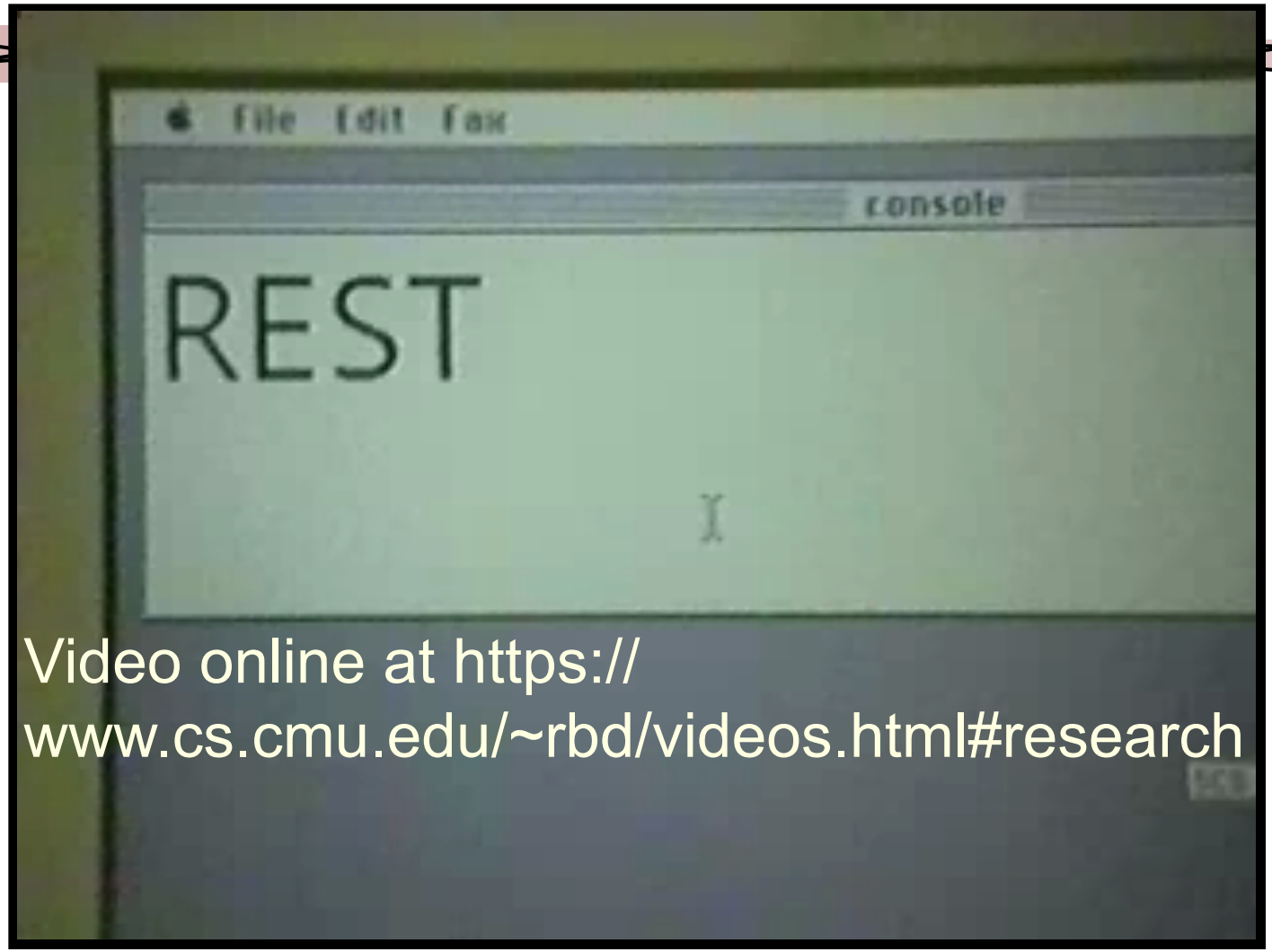
How It Works



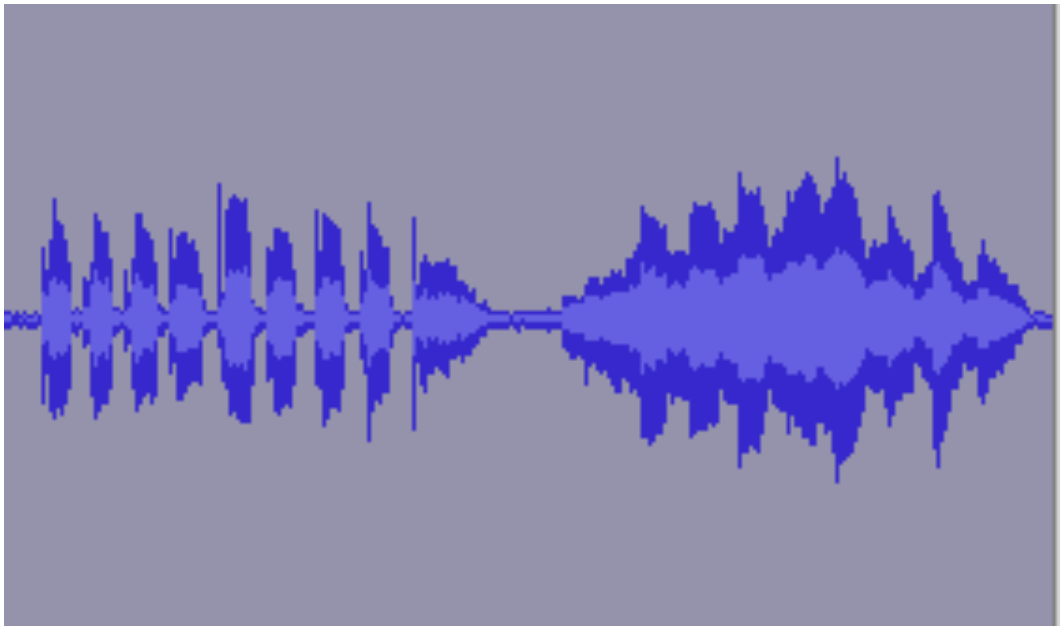
Listening to Jazz Styles



Jazz Style Recognition



Onset Detection

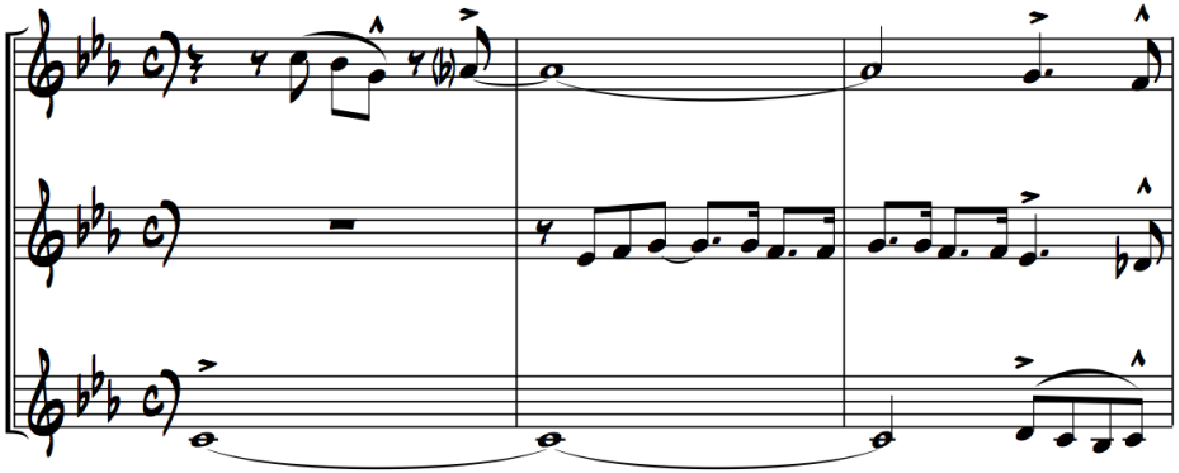


Why?

- Beat Detection
- Tempo Detection
- Computer Accompaniment
- Music Transcription
  - Query-By-Humming
- Automatic Intelligent Audio Editor

Intelligent Audio Editor

- This excerpt is included in the audio examples:



- Before: After:

Some Approaches

- Features and Thresholds
  - High Frequency
  - Phase Change
- Neural Networks
- Hierarchical Models
- HMM

A Bootstrap Method for Training an Accurate Audio Segmenter

Ning Hu and

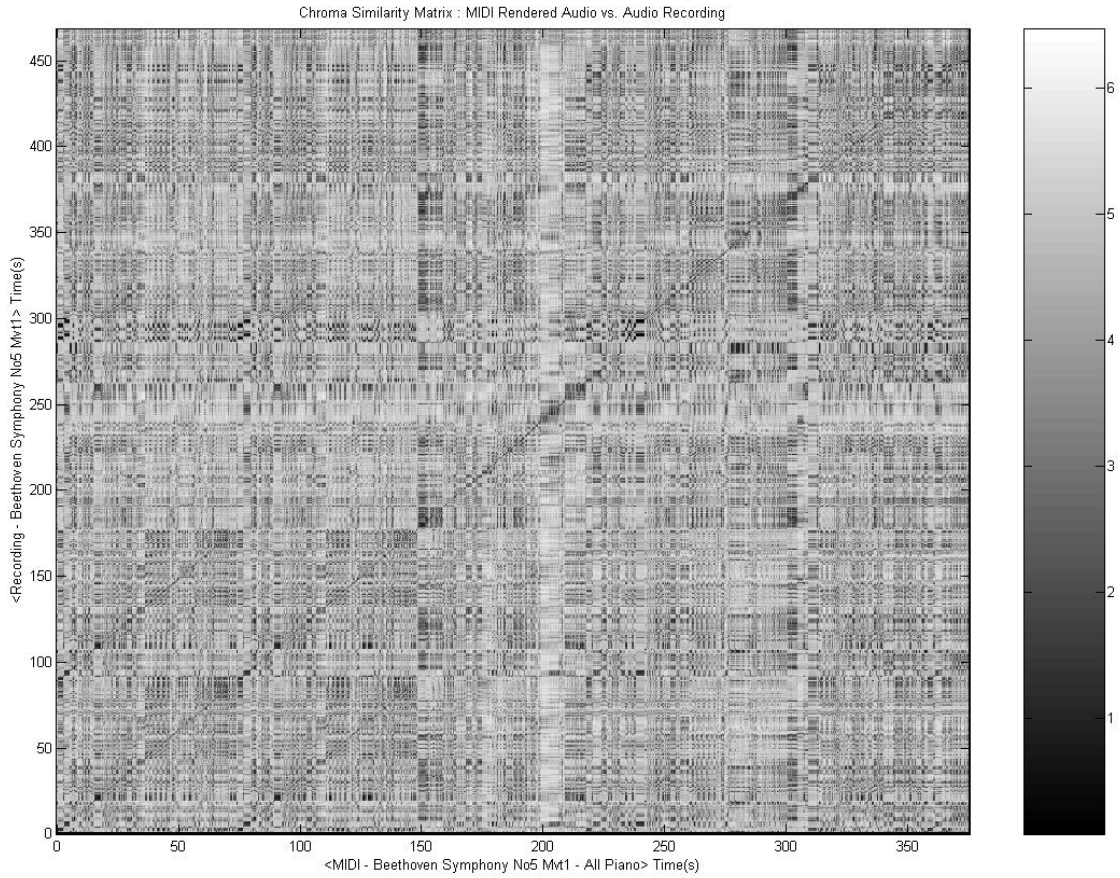
Roger B. Dannenberg

Carnegie Mellon University

Introduction

- Audio segmentation is one of the major topics in MIR research:
  - HMM approach (Raphael, 1999)
  - Neural Network approach (Marolt, et al., 2002)
  - Support Vector Machine (Lu, et al. 2001)
  - Hierarchical Model (Kapanci and Pfeffer, 2004)
- In many cases, collecting training data is time-consuming and expensive.

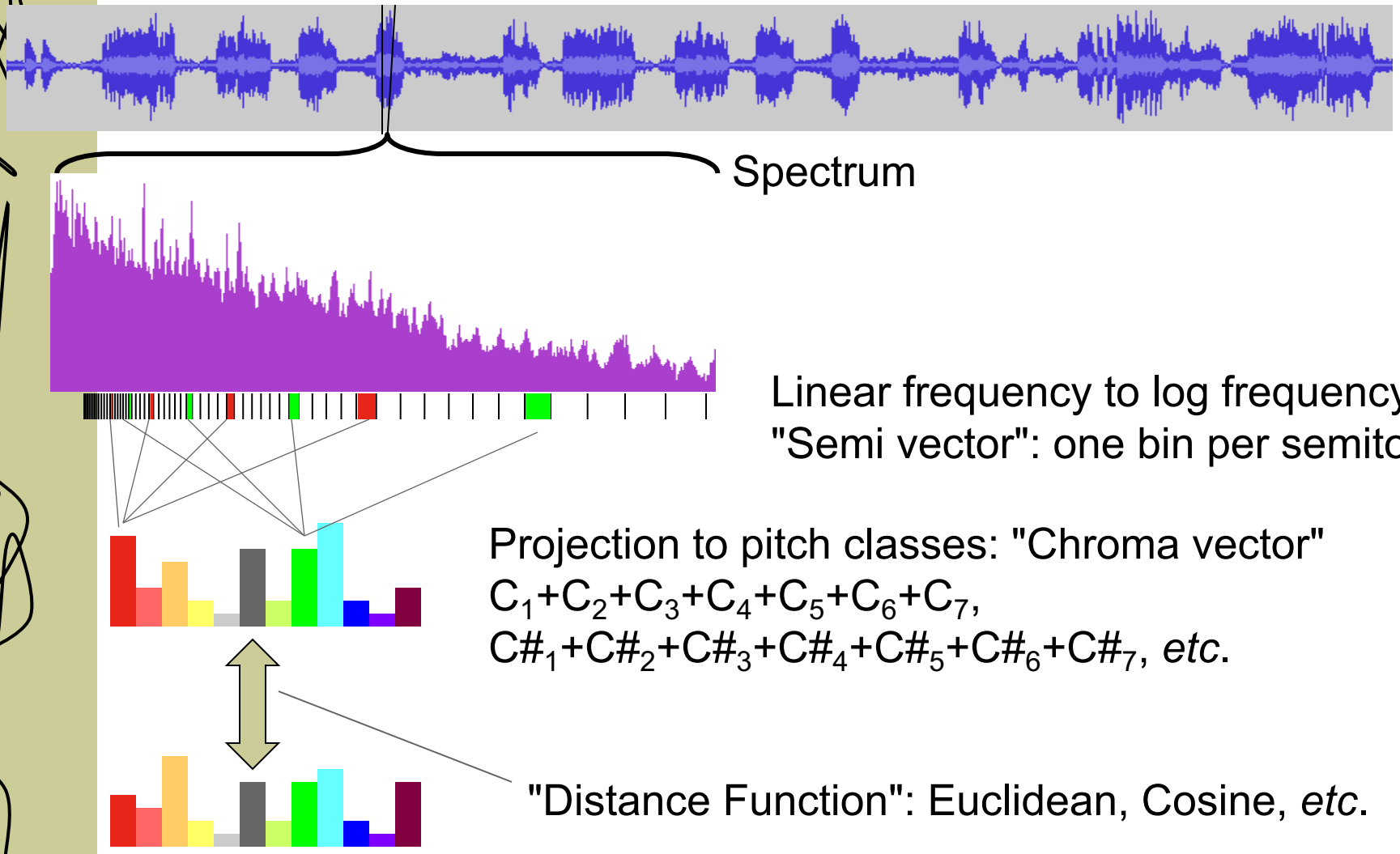
Detour - Audio Alignment



Audio Alignment Concepts

- "Score"
  - Midi File, Note List, not necessarily "real" notation
- Similarity Matrix
- Chroma Vectors
- Distance/Similarity Function
- Research on accurate alignment

Chromagram Representation



Linear frequency to log frequency:
"Semi vector": one bin per semitone

Projection to pitch classes: "Chroma vector"
$C_1+C_2+C_3+C_4+C_5+C_6+C_7$,
$C\#_1+C\#_2+C\#_3+C\#_4+C\#_5+C\#_6+C\#_7$, *etc*.

"Distance Function": Euclidean, Cosine, *etc*.

Segmentation and Alignment

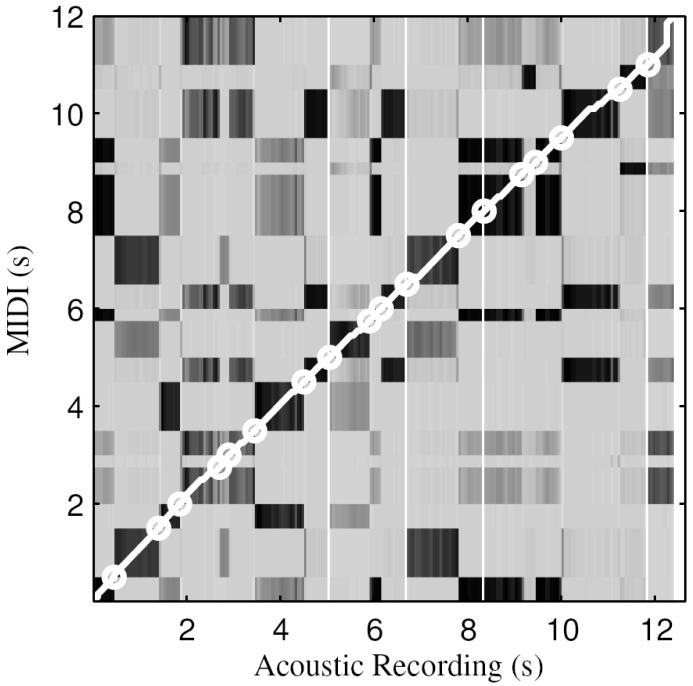
- Segmentation, audio alignment, and score-following are related
  - Rely on acoustic features
  - Precise alignment to symbolic score provides segmentation data
- We use alignment data to train a segmenter
  - Alignment avoids gross errors in segmentation
  - Segmenter learns fine-grain features that improve precision beyond initial alignment
  - $\rightarrow$ high quality segmentation and alignment

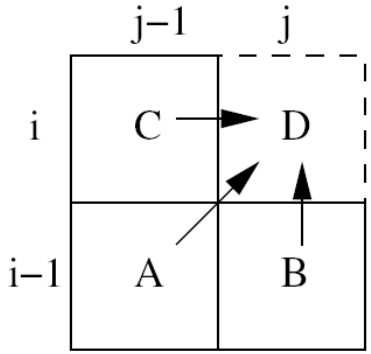
Motivation

- We need very accurate segmentation to extract trumpet envelopes (attacks ~30ms)
  - (for research on capturing synthesis models) 
- Alignment is based on chroma (100 – 250ms)
- Orio & Schwarz (2001) also use DTW and short-term features (5.8 ms windows), but alignment (an $O(N^2)$ algorithm) is slow.
  - Our system performs alignment 25x faster.
- Our small non-DTW analysis windows can use different features.

Audio-to-(MIDI)-Score Alignment

- Chromagram features from Audio
- Synthetic chromagram features for MIDI



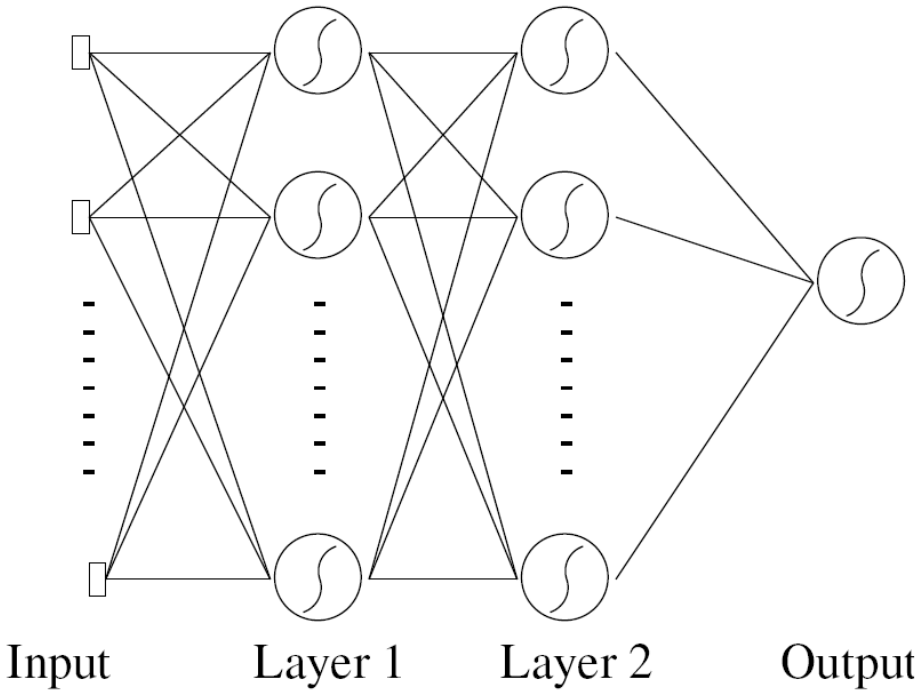


$$D = M_{i,j} = min(\begin{bmatrix} A \\ B \\ C \end{bmatrix} + \begin{bmatrix} \sqrt{2} \\ 1 \\ 1 \end{bmatrix} \times dist(i,j))$$

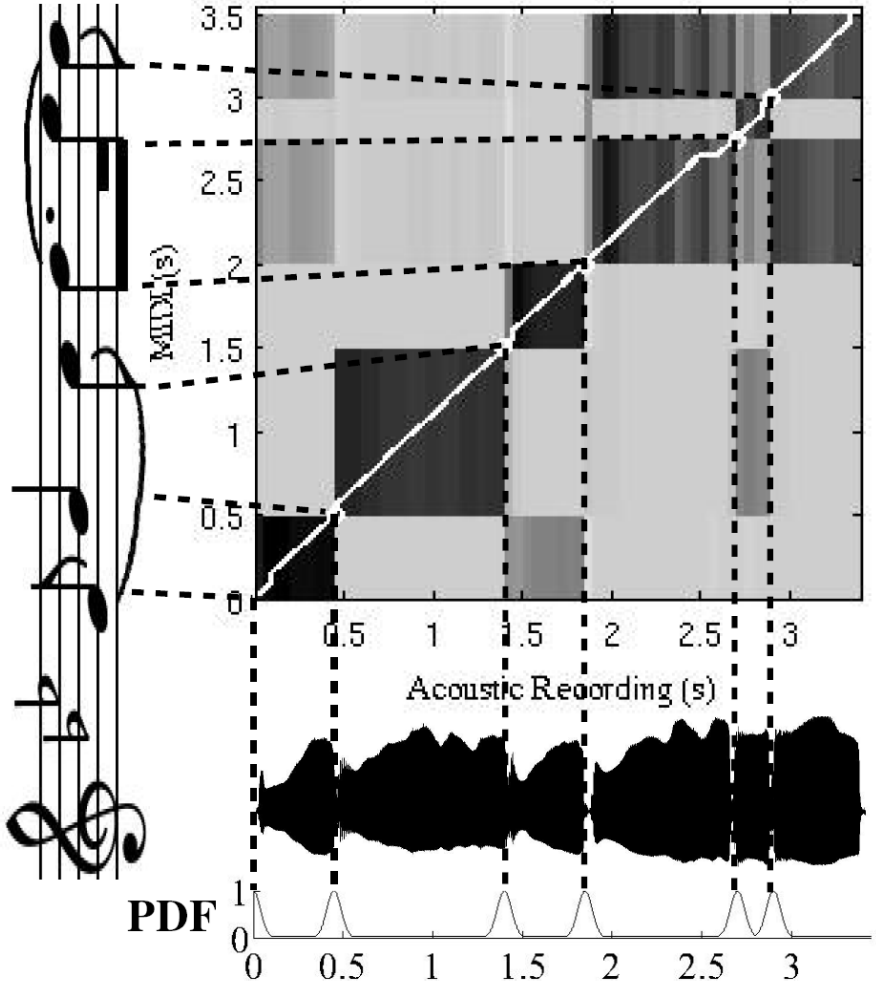
Acoustic Features for Segmentation – 5.8 ms window

- Log energy (dB)
- F0 with SNDAN's (Beauchamp) MQ analysis
- Relative strengths of first 3 harmonics:
  - $Amplitude_i / Amplitude_{overall}$
- Relative frequency deviations, first 3 harmonics:
  - $(f_i - i \times F0) / f_i$
- Zero-crossing rate
- Derivatives of all of the above

Neural Network



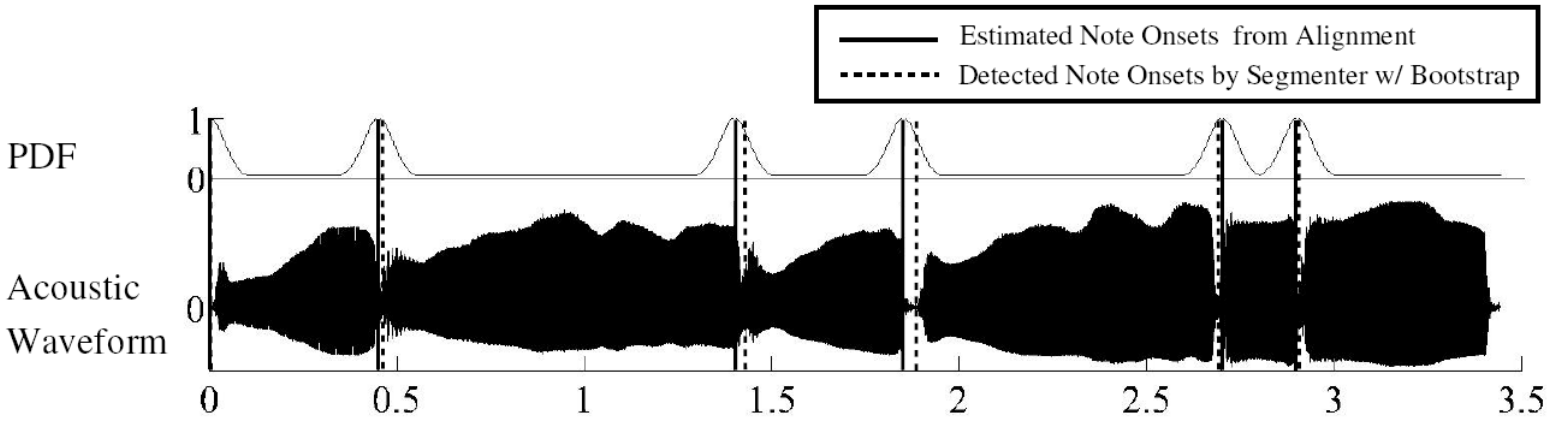
Segment boundary PDF



- Gaussians
- On alignment boundaries
- Width based on alignment window size
- P=0.04 between boundaries

Bootstrap learning process

- Multiply neural net output by PDF
- For each neighborhood around a segment boundary, find the peak → "adjusted onset"
- Retrain the neural network:
  - adjusted onsets are 1, other points are 0



Results

SYNTHETIC

Model	Miss Rate	Spurious Rate	Av. Error	STD
Baseline Segmenter	8.8%	10.3%	21 ms	29 ms
Segmenter w/ Bootstrap	0.0%	0.3%	10 ms	14 ms

REAL

Model	Miss Rate	Spurious Rate	Av. Error	STD
Baseline Segmenter	15.0%	25.0%	35 ms	48 ms
Segmenter w/ Bootstrap	2.0%	4.0%	8 ms	12 ms

Sound Examples

- Input

- Output – segmenter was trained on similar data using the bootstrap method. This input was segmented without using any score information.

Conclusions

- Supervised learning often wins over hand-crafted systems
- Segmentation training data is expensive, so supervised training is difficult
- Alignment provides strong hints, but not accurate enough for training
- Bootstrapping allows segmenter to generate its own training data
- Dramatic improvements in accuracy, even when tested without alignment “hints”

Summary

- Computer Accompaniment
- Offline Score Alignment
- Onset Detection