A Deep Learning Approach to Musical Chord Recognition

Michael Kruppa

Motivation

- Automated transcription
- Live “play along”
- Use in structural segmentation, content ID, style analysis

→ Deep Learning has seen a lot of success in harder problems

Literature

- **Design and Evaluation of a Simple Chord Detection Algorithm**, Christoph Hausner, 2014
  - Template-based matching
- **Neural Networks For Musical Chords Recognition**, J. Osmalskyj et al., 2012
  - Simple network
  - Chords only, no music
- **Towards Automatic Extraction of Harmony Information from Music Signals**, Christopher Harte, 2010
  - Chord notation
- **(Textbook)**

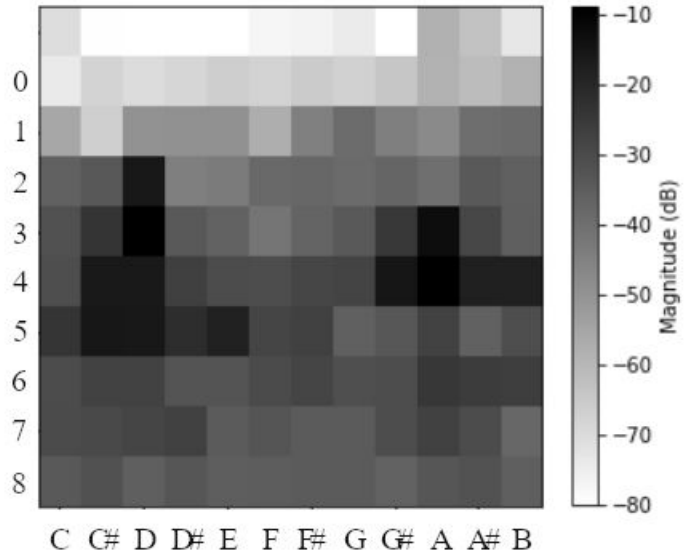
Data Source

- http://isophonics.net/content/reference-annotations-beatles
- Annotations of Beatles songs
  - Beats
  - Chords
  - Segmentations
  - Key
- Use youtube-dl to download 87 songs
  - Fair Use - free, educational, research purpose

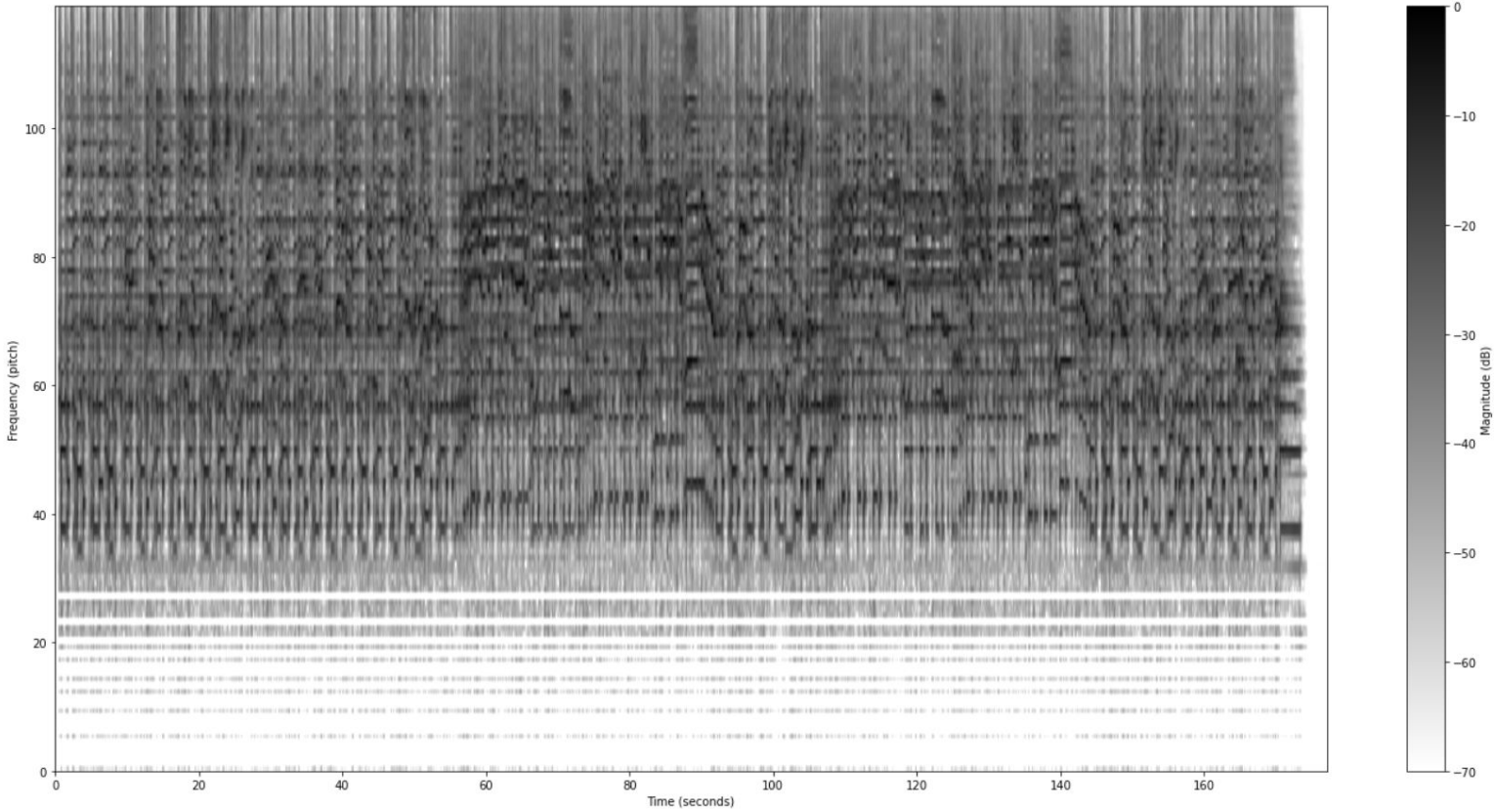
```
0.000000 0.459543 N
0.459543 0.714724 B:(6)
0.714724 0.950652 B:(1)
0.950652 1.168004 B:(6)
1.168004 3.082765 E
3.082765 4.952222 C#:min
4.952222 6.833038 F#:min9
6.833038 8.771904 B
8.771904 10.583061 E
10.583061 12.405827 G#
12.405827 14.228594 F#:min
14.228594 16.097800 B
16.097800 17.850907 E
17.850907 19.685283 G#
19.685283 21.496439 F#:min
21.496439 23.330816 B
23.330816 25.165192 E
25.165192 26.964739 C#:min
26.964739 28.764285 F#:min9
28.764285 30.552222 B
```

Feature Extraction

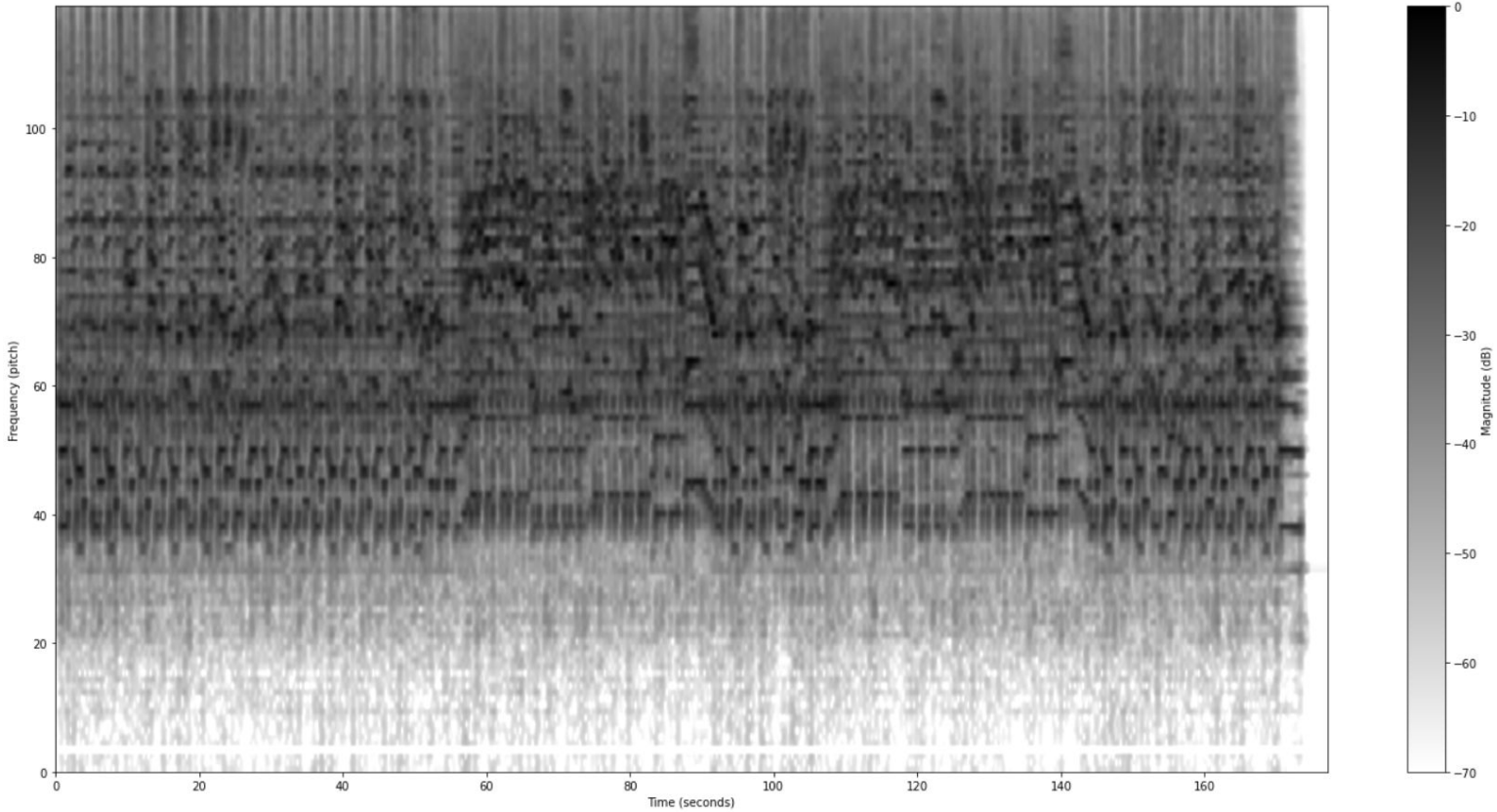
- Multiresolution STFT
  - H = 0.1s, shortWindow = 0.2s, longWindow = 0.8s
  - Try to capture both time and frequency locality
  - Long Window helps with broken chords eg. Alberti Bass
- Spectrogram Wrapping instead of Chromagram
  - Preserve octave information while suggesting cyclic structure
  - 12x10 “image” representing MIDI notes [0, 120)
- Consider only 24 major and minor root chords (and no chord)
  - Have to convert annotation to exclude 7ths, inversions, adds, sus, etc.
  - Enharmonics lost since no key information



Short Window

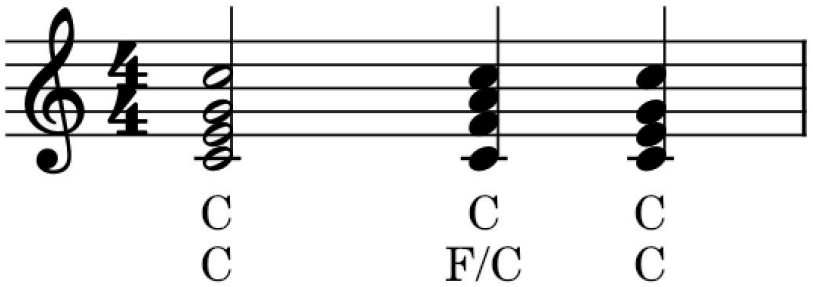


Long Window


Frequency (pitch)
0
20
40
60
80
100
Time (seconds)
0
20
40
60
80
100
120
140
160
Magnitude (dB)
0
−10
−20
−30
−40
−50
−60
−70

Practical Problems

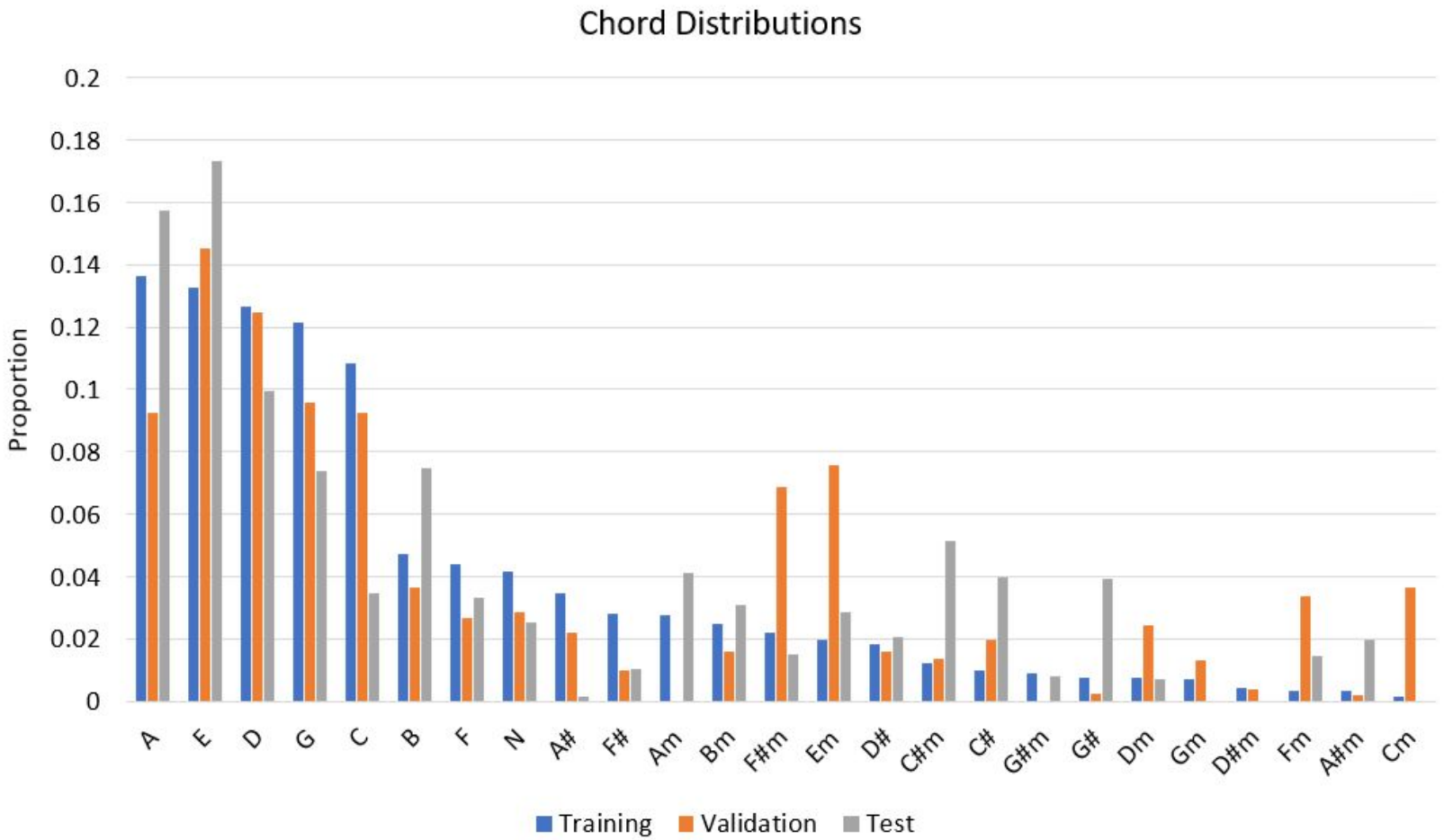
- Specificity to training data - chords or styles
- Detuning eg. Baroque A
- Passing Tones
- Arpeggiated Chords
- Only considering 24 chords might be harder - 7ths can help resolve
- Ground truth hard to define
    - Annotators mistake
    - Chord change vs passing/neighbor tones



Data Preprocessing

- Markov Chain Idea - consider previous feature as well
- 4 Spectrogram Images (Channels) per labeled chord
  - Current frame short window
  - Current frame long window
  - Previous frame short window
  - Previous frame long window
- Split into Train/Validation/Test sets
  - First by frame - problematic!
  - Then by song (71/8/8)

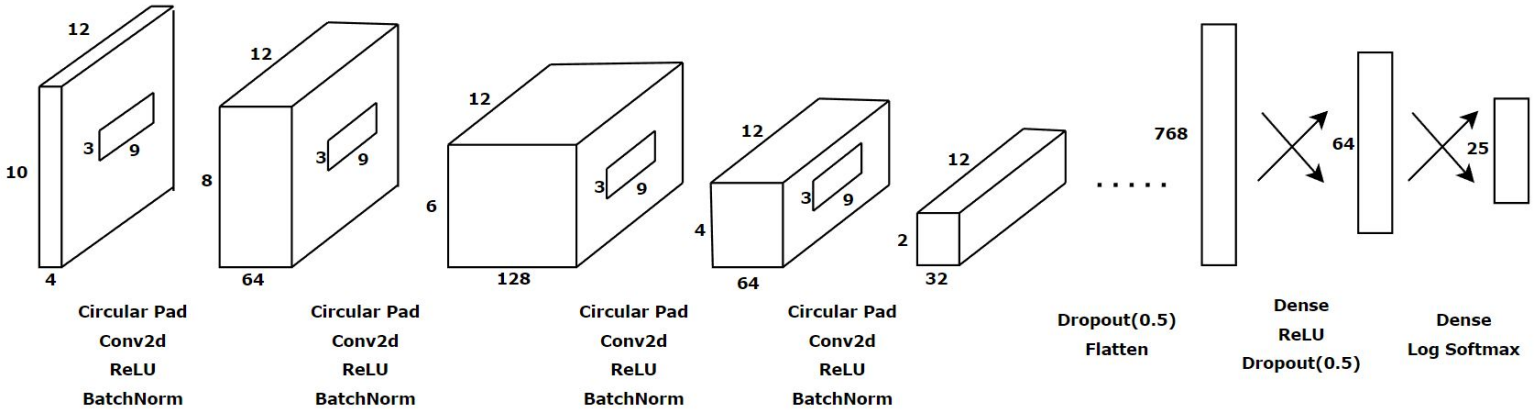
- 108123 training labels
- 10999 validation labels
- 11444 test labels


Chord Distributions
Proportion
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
A
E
D
G
C
B
F
N
A#
F#
Am
Bm
F#m
Em
D#
C#m
C#
G#m
G#
Dm
Gm
D#m
Fm
A#m
Cm
Training
Validation
Test

.	.	.	.	.	.	.	.	.	.	.	.	.			
.	A#4	B4	C4	C#4	D4	D#4	E4	.	.	.	A#4	B4	C4	C#4	D4
.	A#5	B5	C5	C#5	D5	D#5	E5	.	.	.	A#5	B5	C5	C#5	D5
.	A#6	B6	C6	.	.	.	.	.	.	.	A#6	B6	C6	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

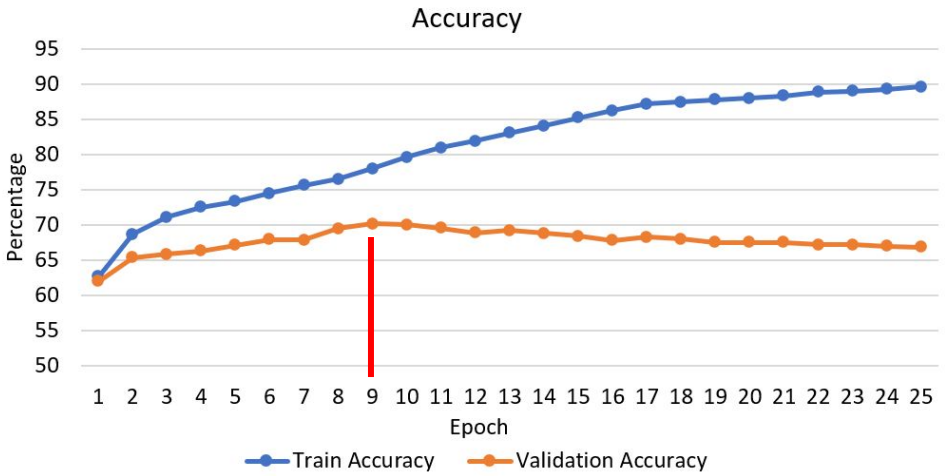
Network Architecture

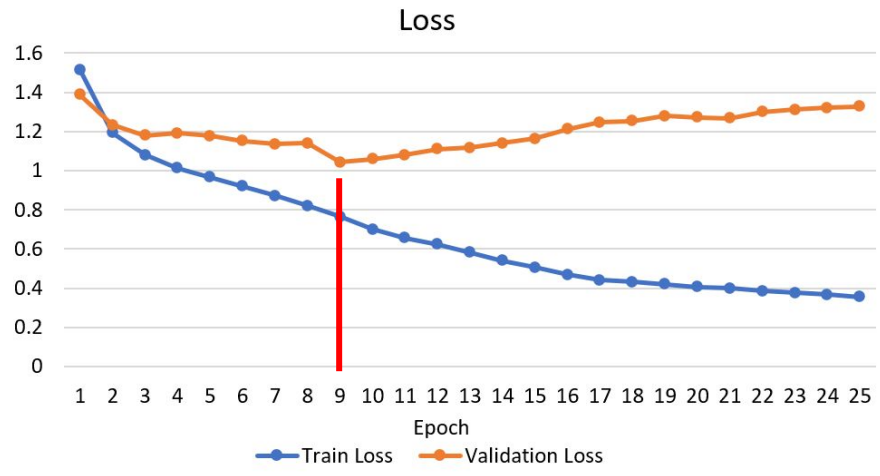
- Classification Problem
- Convolutions to pick up structures in image (harmonic series), but need to be careful
  - Edge policy - Cyclical prepadding along last axis
  - Translation variance - Pad after every convolution, no pooling
- Batch Size = 128, LR = 0.0005, Adam Optimizer, Weight Decay = 0.001
- 556281 learnable parameters



Training Results

- 70% accuracy on validation set (early stopping)

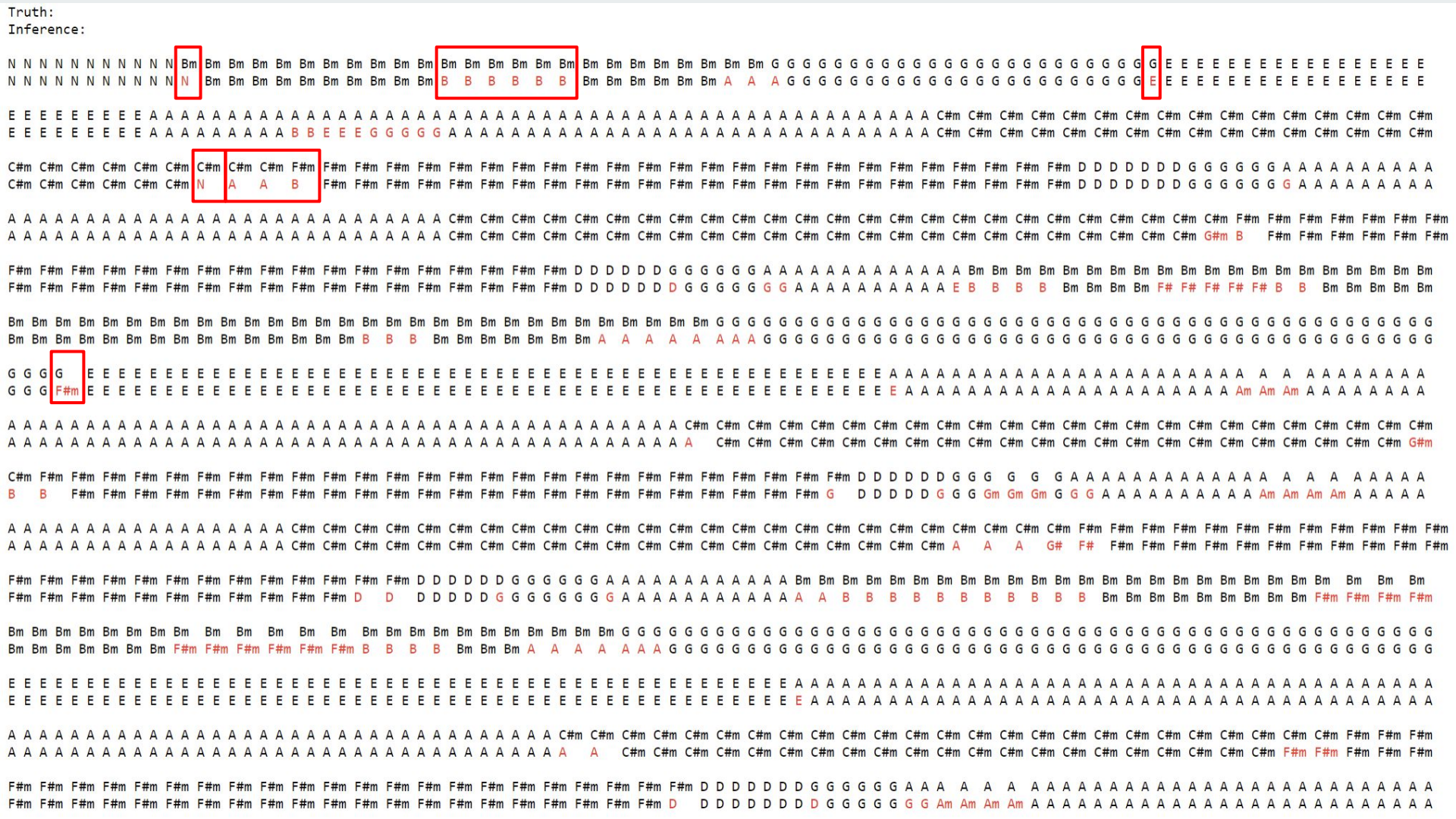




Confusion Matrix on Validation Song (Help!)

Truth \ Inference	N	C	C#	D	D#	E	F	F#	G	G#	A	A#	B	Cm	C#m	Dm	D#m	Em	Fm	F#m	Gm	G#m	Am	A#m	Bm
N	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C#	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	36	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D#	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	177	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F#	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	2	0	1	0	1	195	0	8	0	0	0	0	0	0	0	0	3	3	0	0	0	0
G#	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	2	0	0	1	0	7	0	0	14	0	386	0	2	0	0	0	0	0	0	1	0	0	13	0	0
A#	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C#m	1	0	0	0	0	0	0	0	0	1	12	0	1	0	132	0	0	0	0	2	0	2	0	0	0
Dm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D#m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Em	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F#m	0	0	0	3	0	0	0	3	1	0	0	0	3	0	0	0	0	0	0	166	0	0	0	0	1
Gm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G#m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Am	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A#m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bm	1	0	0	0	0	1	0	6	0	0	19	0	45	0	0	0	0	0	0	10	0	0	0	0	94

Average F: 0.8993589834970108


Truth:
Inference:

Error Types

- Boundaries off by a couple frames
- Major-minor confusion
- Chords that share notes
- Occasional completely wrong chords

→ Use a mode blur as post processing - edge preserving and should remove occasional wrong chords

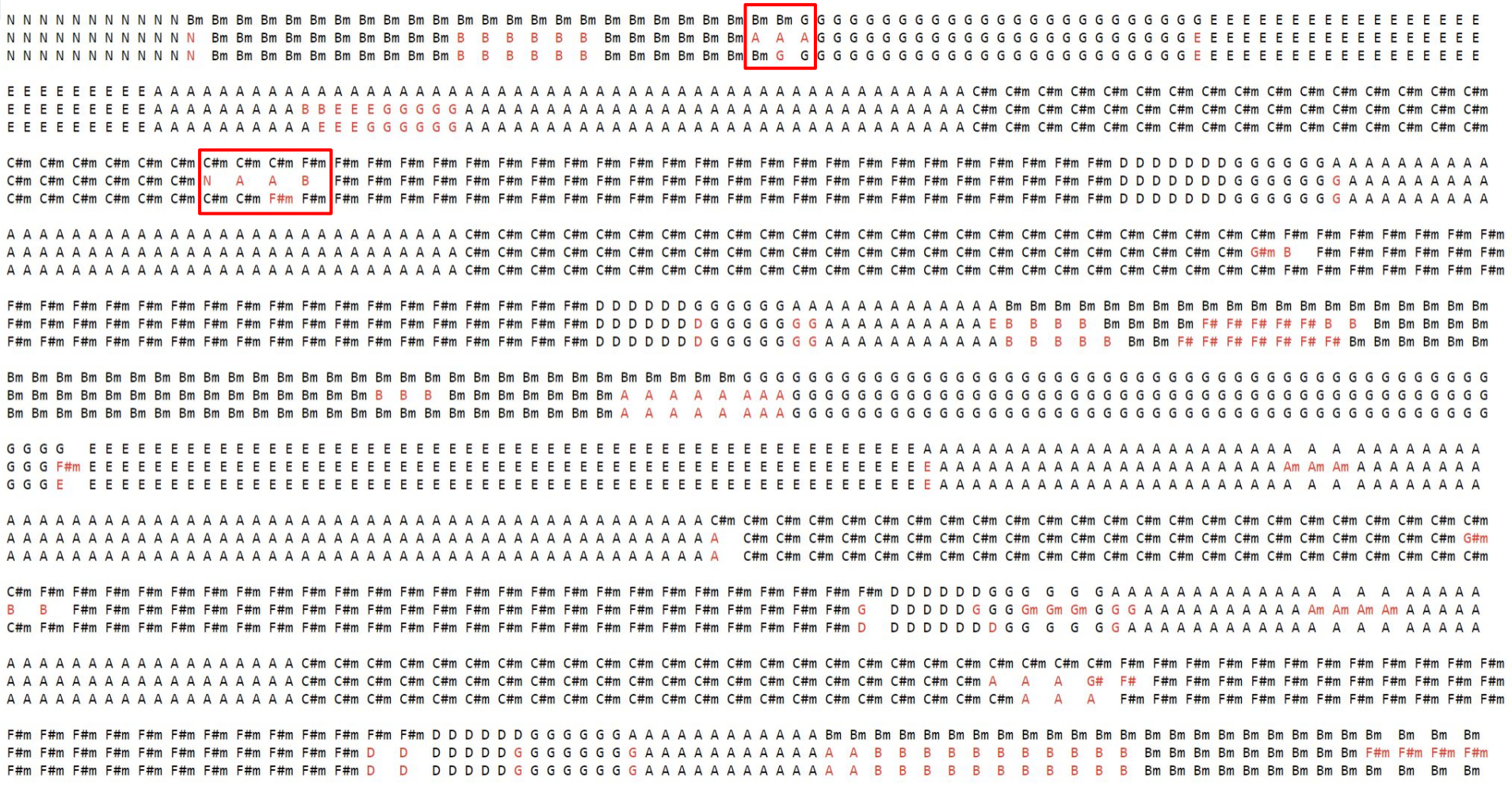
Confusion Matrix on Validation Song (Help!) with Mode Blur

Inference

Truth	N	C	C#	D	D#	E	F	F#	G	G#	A	A#	B	Cm	C#m	Dm	D#m	Em	Fm	F#m	Gm	G#m	Am	A#m	Bm
N	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C#	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	37	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D#	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	178	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F#	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	3	0	3	0	0	200	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G#	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	2	0	0	0	0	6	0	2	14	0	399	0	0	0	0	0	0	0	0	1	0	0	2	0	0
A#	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C#m	0	0	0	0	0	0	0	0	0	0	8	0	0	0	139	0	0	0	0	4	0	0	0	0	0
Dm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D#m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Em	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F#m	0	0	0	4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	172	0	0	0	0	0
Gm	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G#m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Am	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A#m	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bm	1	0	0	0	0	0	0	7	1	0	22	0	44	0	0	0	0	0	0	6	0	0	0	0	95

Average F: 0.9140323375236533

Truth:
Inference:
Blurred Inference:



Limitations and Future Work

- Errors in chord recognition, though some types are acceptable
- Only detects 24 chords, could extend to more types
- Slow, can’t be realtime
- Data augmentation on the training data to cyclically shift to every key - 12 times more data
- Enharmonics lost - need key recognition
- Overfitting
- Markov Idea not as useful as I thought - circular reasoning
    - 69% accuracy without it

Live Demo

- Inference “in the wild”
- Plays audio and shows inferred chords
- Sync issue, so actually renders a video

→ Name a piece! (a familiar piece, not a Beatles song)