Learning Features of Music from Scratch

John Thickstun, Zaid Harchaoui, and Sham Kakade

eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

ICRA
INSTITUT CANADIEN DE RECHERCHES AVANCÉES

MusicNet

A curated collection of labeled classical music

Minutes	Labels	Recordings	Error Rate
2,048	1,299,329	330	4.0%

Ensemble	Minutes	Labels
Solo Piano	917	576,471
String Quartet	405	259,702
Accompanied Violin	148	124,886
Piano Quartet	73	60,362
Accompanied Cello	63	37,557
String Sextet	48	33,248
Piano Trio	46	28,873
Piano Quintet	25	27,545
Wind Quintet	43	24,820
Horn Piano Trio	30	18,799
Wind Octet	23	14,635
Clarinet-Cello-Piano Trio	25	13,447
Pairs Clarinet-Horn-Bassoon	24	12,218
Clarinet Quintet	26	11,184
Solo Cello	49	10,876
Accompanied Clarinet	20	10,049
Solo Violin	30	8,837
Violin and Harpsichord	16	7,469
Viola Quintet	15	4,156
Solo Flute	8	2,214

Composer	Minutes	Labels
Beethoven	1,085	736,072
Schubert	253	146,648
Brahms	192	133,109
Mozart	156	99,641
Bach	184	62,782
Dvorak	56	46,261
Cambini	43	24,820
Faure	33	22,349
Ravel	27	21,243
Haydn	15	6,404

Instrument	Minutes	Labels
Piano	1346	794,532
Violin	874	230,484
Viola	621	99,407
Cello	800	99,132
Clarinet	173	24,426
Bassoon	102	14,954
Horn	132	11,468
Oboe	66	8,696
Flute	69	8,310
Harpsichord	16	4,914
String Bass	38	3,006

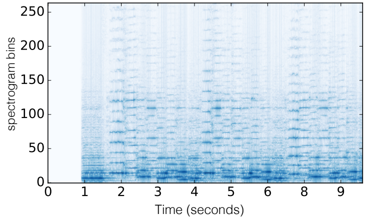
	Piano	Violin	Cello	Viola	Clarinet	Bassoon	Horn	Oboe	Flute	Bass	Harpsichord
Notes	83	51	51	51	41	36	41	28	37	43	51

A sample of labels from the MusicNet dataset:

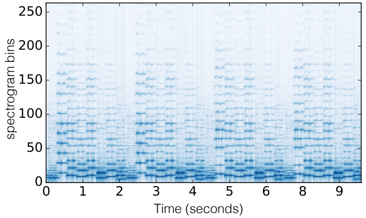
Start	End	Instrument	Note	Measure	Beat	Note Value
45.29	45.49	Violin	G5	21	3	Eighth
48.99	50.13	Cello	A#3	24	2	Dotted Half
82.91	83.12	Viola	C5	51	2.5	Eighth

Music-to-Score Alignment

Human Performance



Synthesized Performance



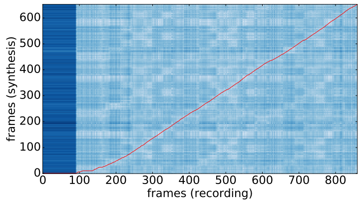
$$C(\mathbf{x},\mathbf{y}) = \|\mathbf{x} - \text{Synthesize}(\mathbf{y})\|_2$$

$$\begin{aligned} \underset{t\in\mathbb{Z}^n}{\text{minimize}} \quad & \sum_{i=1}^{n} C(\mathbf{X}_{t_i}, \mathbf{Y}_i) \\ \text{subject to} \quad & t_0 = 0, \\ & t_n = m, \\ & t_i \leq t_j \qquad \text{if } i < j. \end{aligned}$$

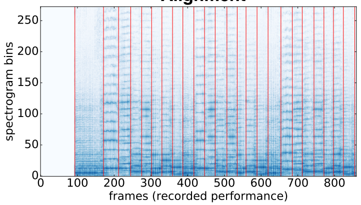
Label MusicNet by aligning performances X to scores Y:

- **synthesize a performance of Y**
- **define a cost between local frequency decompositions of**
  - **i) the human performance (top left)**
  - **ii) the synthesized performance (top right)**
- **find the minimum cost alignment (left) with DP (bottom left)**
- **map notes in the score to human performance timings via this alignment (bottom right)**

Cost Matrix



Alignment



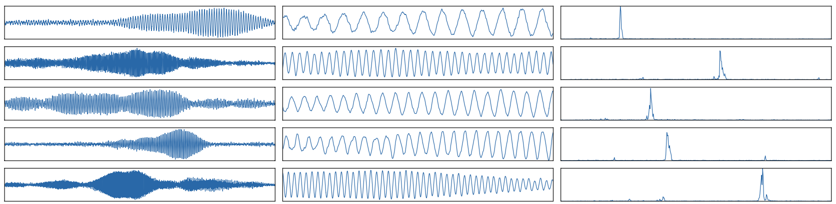
Spectrograms are approximately realizable by an MLP

$$\text{Spec}_k(\mathbf{x}) \equiv \left|\sum_{s=0}^{t-1} e^{-2\pi iks/t}x_s\right|^2 = \left(\sum_{s=0}^{t-1}\cos(2\pi ks/t)x_s\right)^2 + \left(\sum_{s=0}^{t-1}\sin(2\pi ks/t)x_s\right)^2$$

$$\approx \left|\sum_{s=0}^{t-1}\cos(2\pi ks/t)x_s\right| + \left|\sum_{s=0}^{t-1}\sin(2\pi ks/t)x_s\right|$$

Learned features of a (2-layer, ReLU) network mimic a windowed spectrogram (right). Spectrogram-inspired features are a good low-level representation of music.

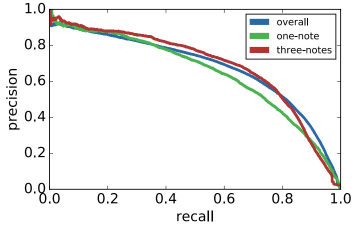
An MLP learns frequency selective filters reminiscent of spectrograms



Frame-based Transcription Results

Representation	Window Size	Precision	Recall	Average Precision
log-spectrograms	1,024	49.0%	40.5%	39.8%
spectrograms	2,048	28.9%	52.5 %	32.9%
log-spectrograms	2,048	61.9%	42.0%	48.8%
log-ReLUgrams	2,048	58.9%	47.9%	49.3%
MLP, 500 nodes	2,048	50.1%	58.0%	52.1%
MLP, 2500 nodes	2,048	53.6%	62.3%	56.2%
AvgPool, 2 stride	2,148	53.4%	62.5%	56.4%
log-spectrograms	8,192	64.2%	28.6%	52.1%
log-spectrograms	16,384	58.4%	18.1%	45.5%
MLP, 500 nodes	16,384	54.4%	64.8%	60.0%
CNN, 64 stride	16,384	60.5%	71.9%	67.8%

A Convolutional Neural Network



A CNN trained on 16,384 samples to predict notes at the center of the frame. Receptive field is 2,048 samples; stride is 8 samples. Features are pooled in groups of 16 with 50% overlap between pools.

MIREX-style results, computed by the mir_eval library

Representation	Acc	Etot	Esub	Emiss	Efa
512-point log-spectrogram	28.5%	.819	.198	.397	.224
1024-point log-spectrogram	33.4%	.715	.123	.457	.135
1024-point log-ReLUgram	35.9%	.711	.144	.377	.190
4096-point log-spectrogram	24.7%	.788	.085	.628	.074
8192-point log-spectrogram	16.1%	.866	.082	.737	.047
MLP, 500 nodes, 2048 raw samples	36.8%	.790	.206	.214	.370
MLP, 2500 nodes. 2048 samples	40.4%	.740	.177	.200	.363
AvgPool, 5 stride, 2048 samples	40.5%	.744	.176	.200	.369
MLP, 500 nodes, 16384 samples	42.0%	.735	.160	.191	.383
CNN, 64 stride, 16384 samples	48.9%	.634	.117	.164	.352

References

E. Benetos, S. Dixon, D. Giannoulis, H. Kirchoff, and A. Klapuri. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 2013.

R. J. Turetsky and D. P. W. Ellis. Ground-truth transcriptions of real music from force-aligned midi syntheses. *ISMIR*, 2003.

C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis. mir_eval: A transparent implementation of common mir metrics. *ISMIR*, 2014.

B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. *SCIPY*, 2015.

G. Hadjeres and F. Pachet. Deepbach: a steerable model for bach chorales generation. *arXiv preprint*, 2016.

S. Dieleman and B. Schrauwen. End-to-end learning for music audio. *ICASSP*, 2014.