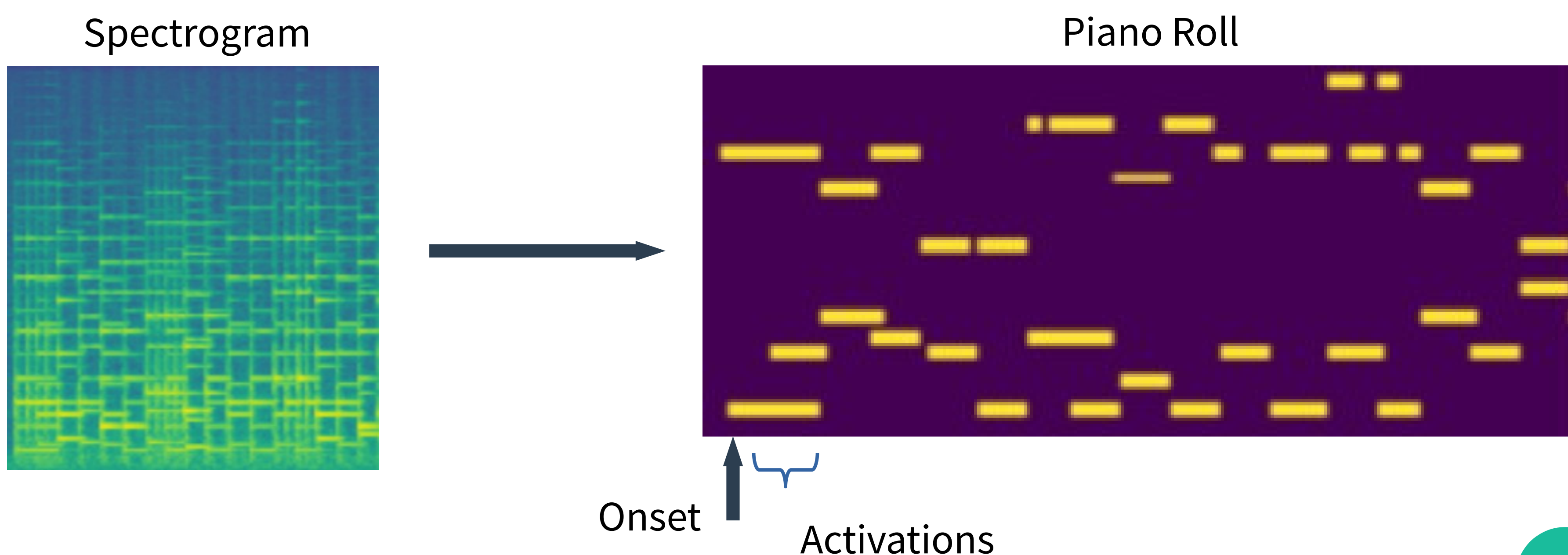
Piano Transcription Using Deep Neural Networks

Nicholas Esterer
December 20th, 2018

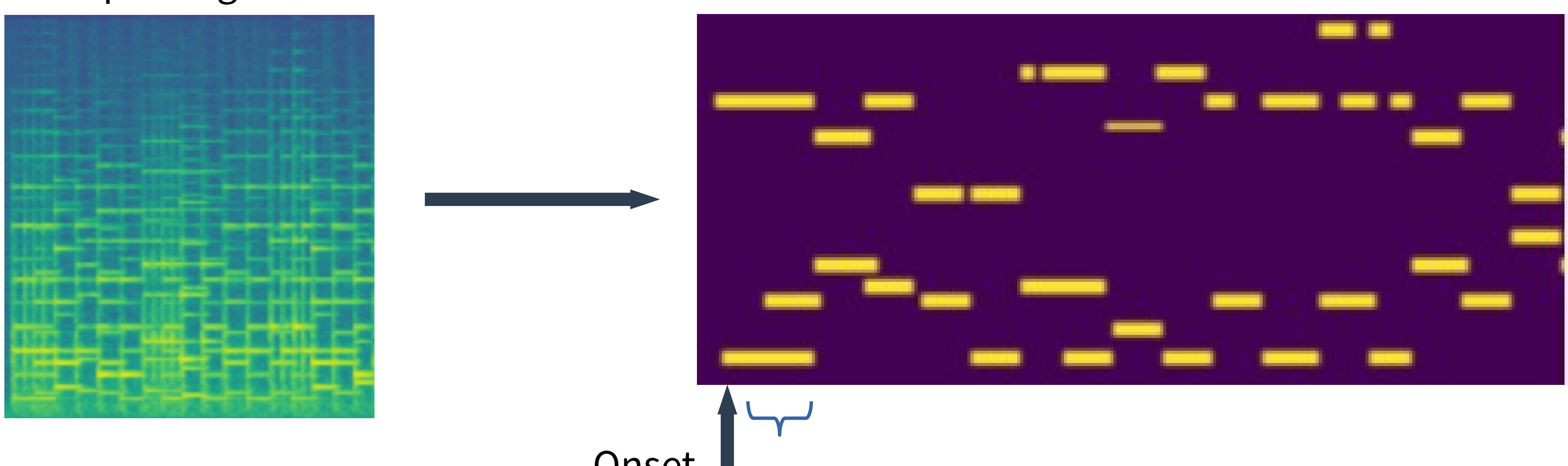
Goal

- **Transcribe note activations and onsets from recordings of piano performances.**



Goal

- Transcribe note activations and onsets from recordings of piano performances.
- Only interested in activations (not velocity or sustain information)



Problem

- Example recording from dataset

Problem

- Example recording from dataset

Problem

- Example of a real recording

Problem

- Example of a real recording

Goals

- **We would like to transcribe recordings like the latter**
  - We work with recordings of Glenn Gould
- **But we only have examples like the former to work with**
  - Real recordings noisy, more variation in speed, dynamic, expression.

Goals

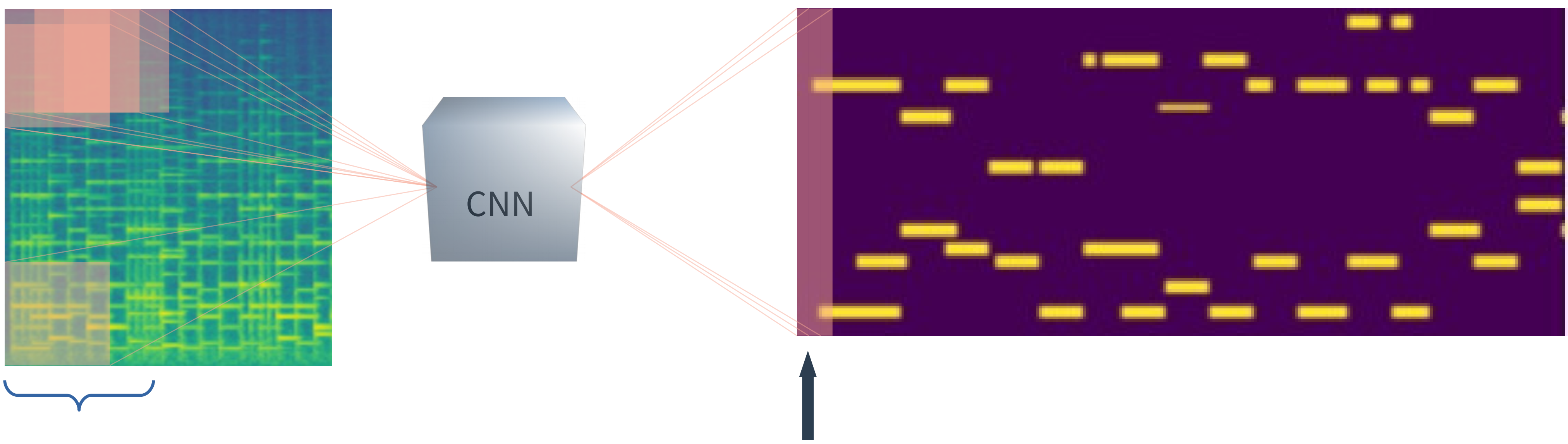
- **We would like to transcribe recordings like the latter**
  - We work with recordings of Glenn Gould
- **But we only have examples like the former to work with**
  - Real recordings noisy, more variation in speed, dynamic, expression.
- **Transcription should be performed “online”**

Outline

- **Technique**
  - Convolutional neural network, recurrent neural network
- **Initial results**
  - Performance on test set and real recordings
- **Directions for improvement**
  - Adding noise and regularization
  - Dataset augmentation
  - Alternative piano roll representations

Methods

- Convolutional neural network (Kelz 2016)



“Slice” of spectrogram

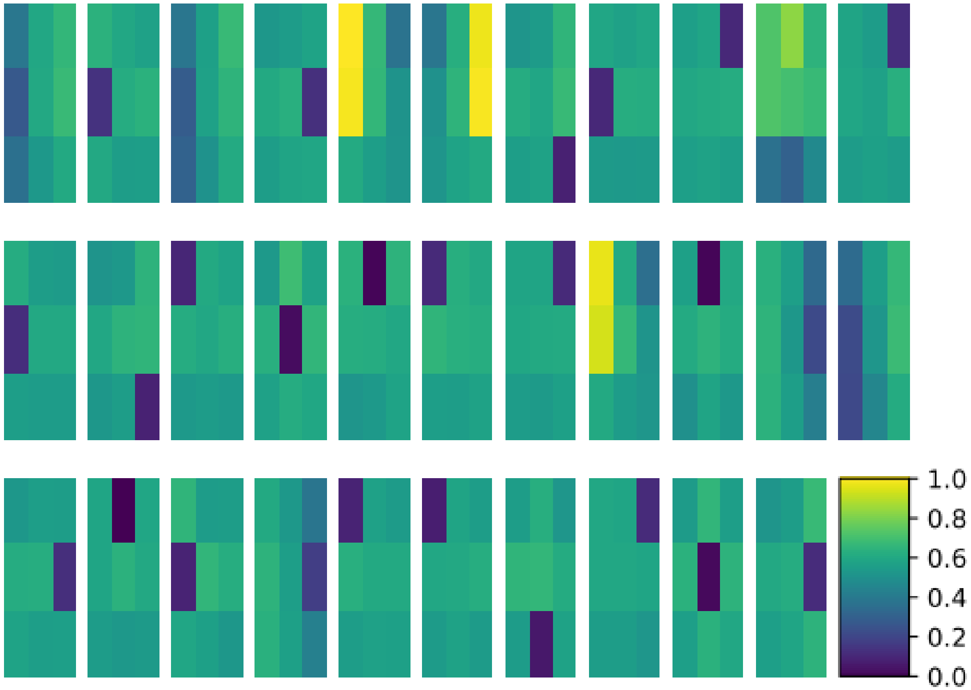
“Slice” of piano roll

Methods

- Example kernels from the frame CNN.

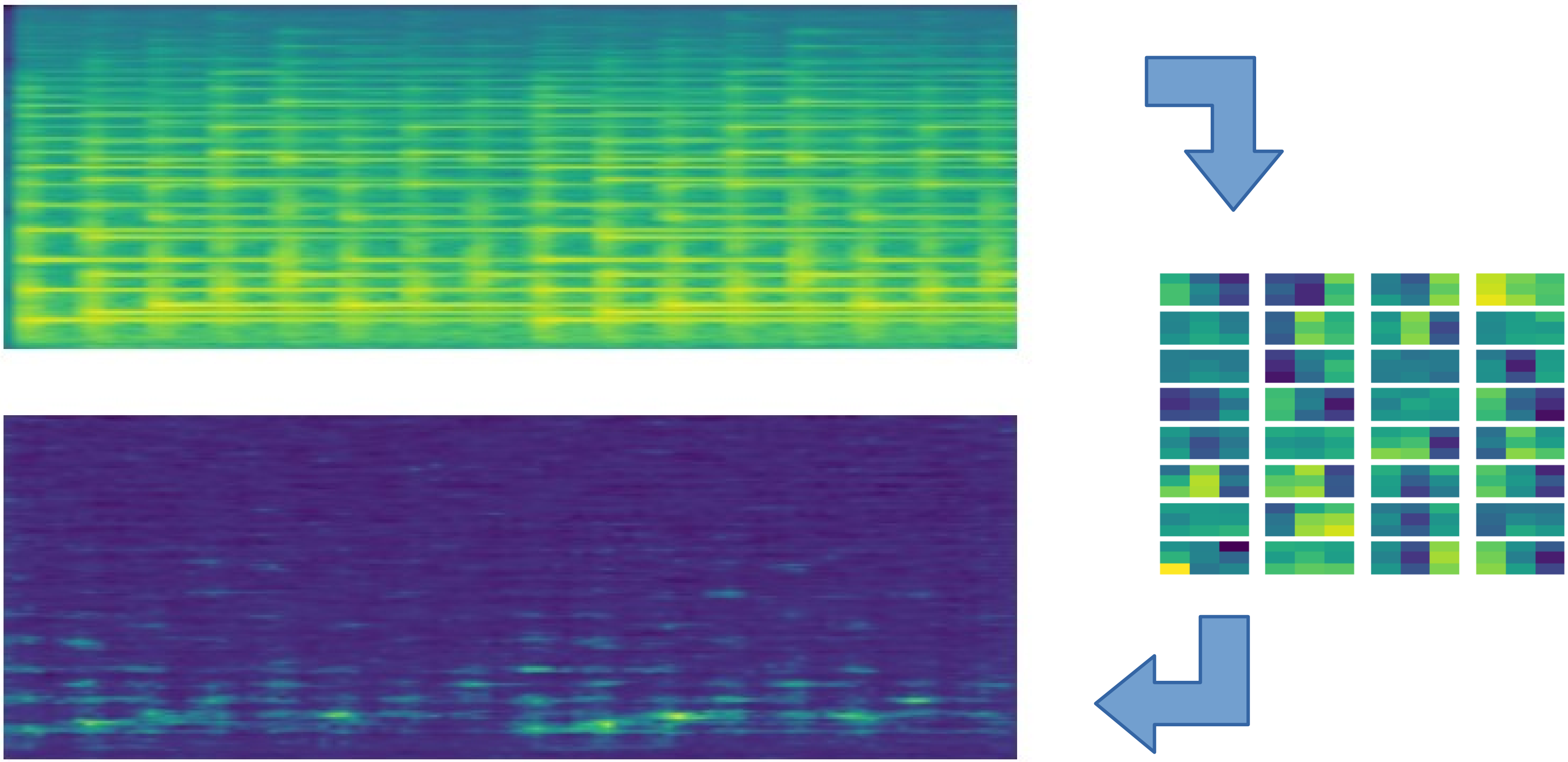
Methods

- Example kernels from the frame CNN.



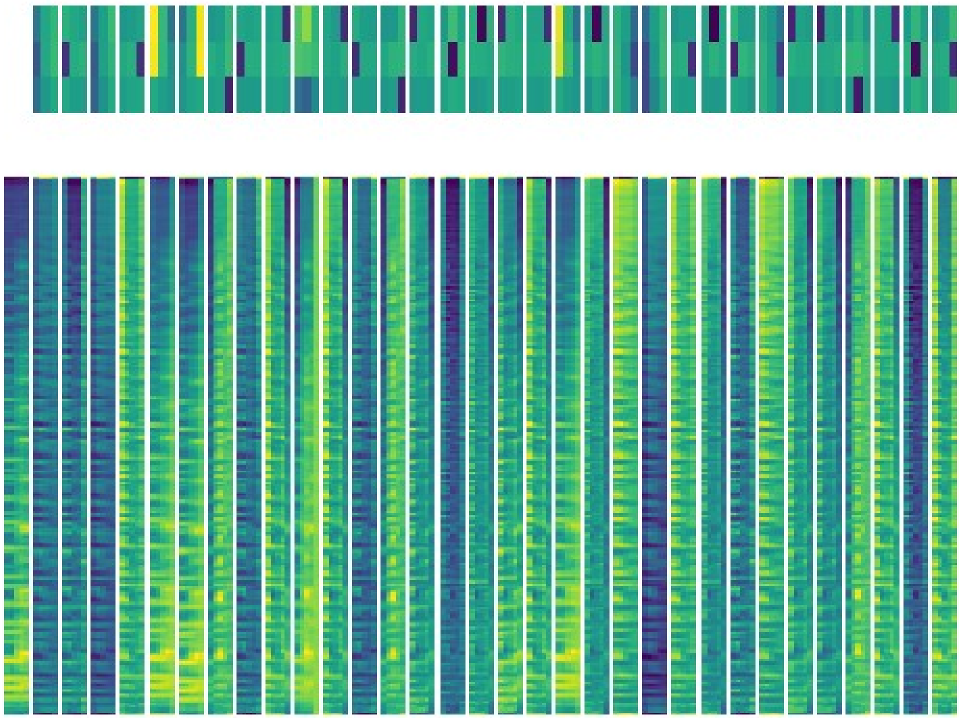
Methods

- Action of the kernels on some input

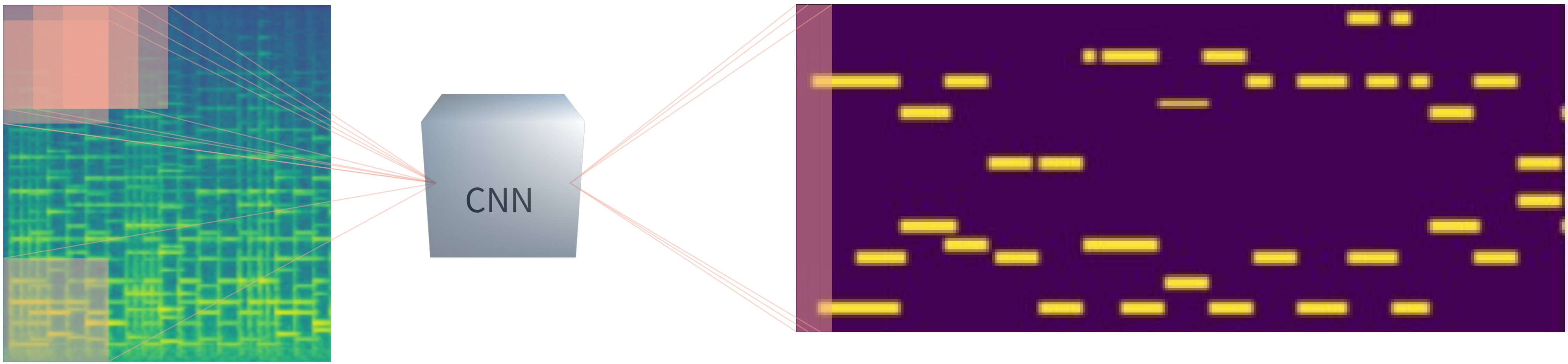


Methods

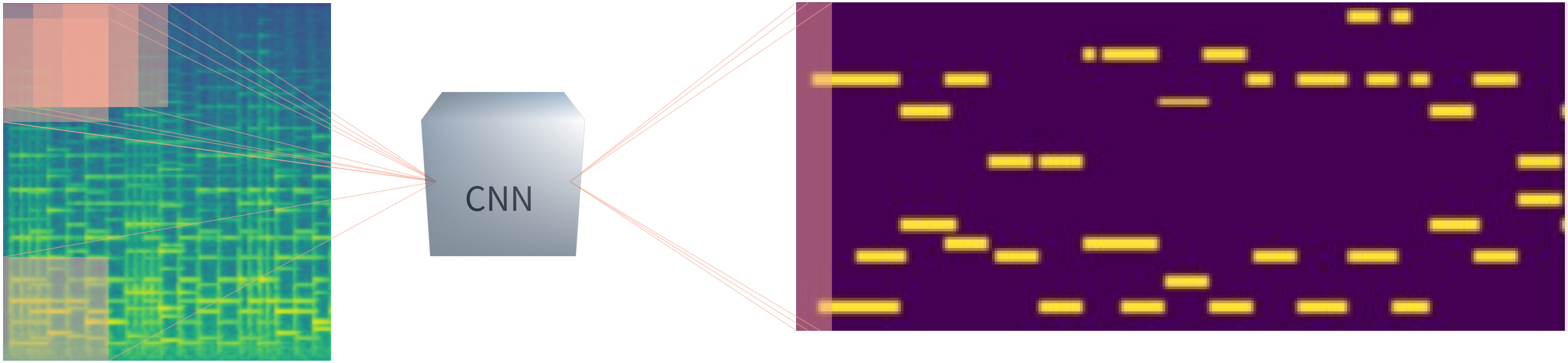
- Kernels and the resulting activations from the frame CNN



Methods



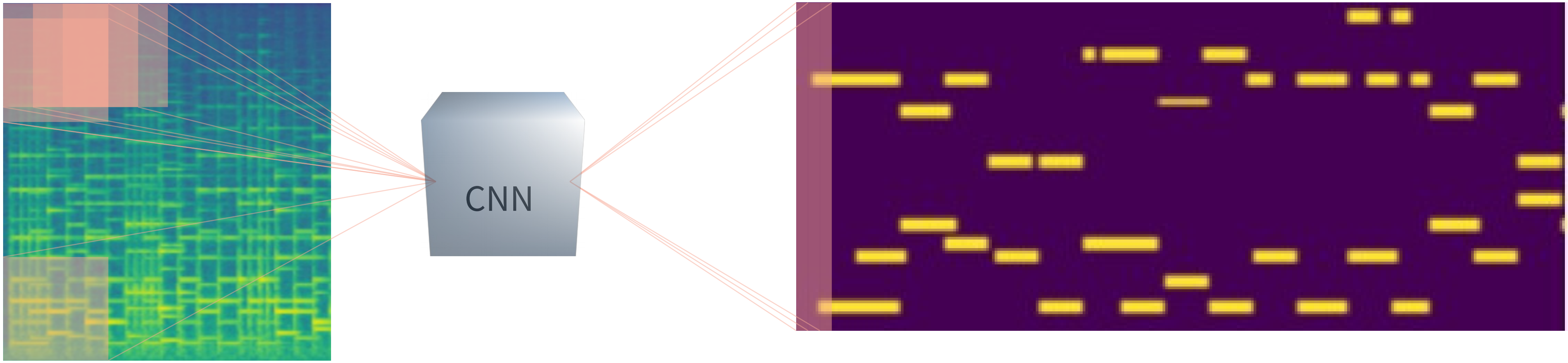
Methods



	Activation	Onset	Note with offset
Kelz et al 2016	.7160	.5094	.2314

Methods

- Music is contextual



	Activation	Onset	Note with offset
Kelz et al 2016	.7160	.5094	.2314

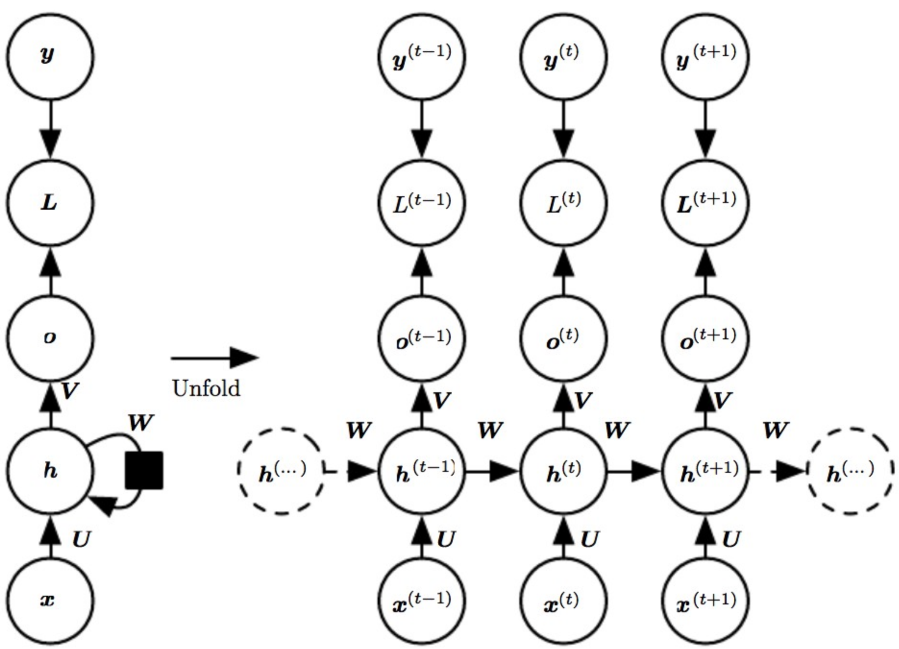
Methods

- This motivates the use of recurrent networks

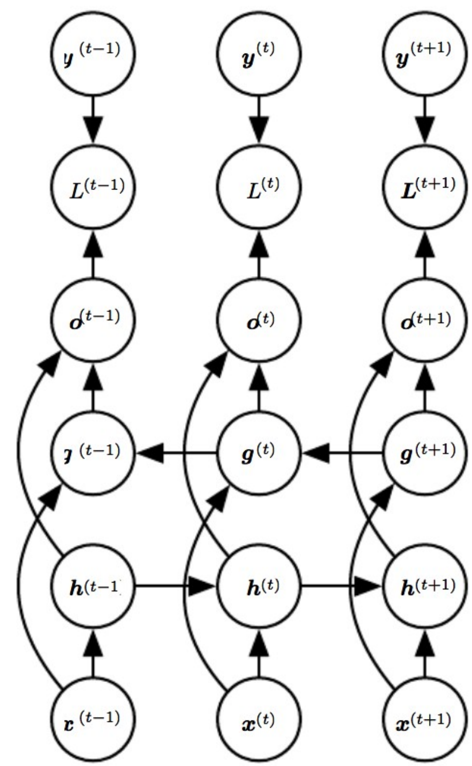
Methods

- This motivates the use of recurrent networks

Unidirectional recurrent network



Bidirectional recurrent network

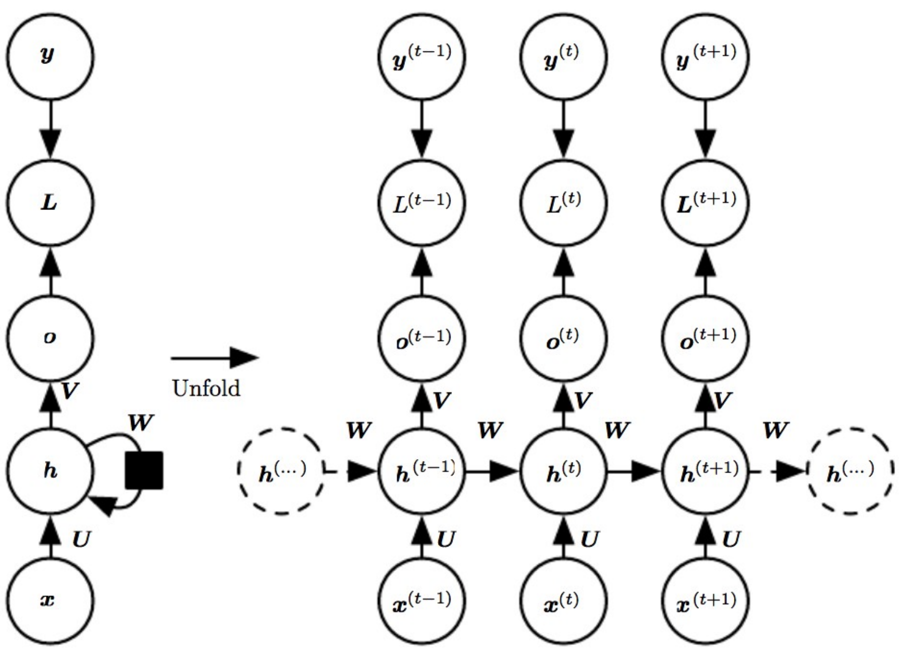


Goodfellow et al (2016)

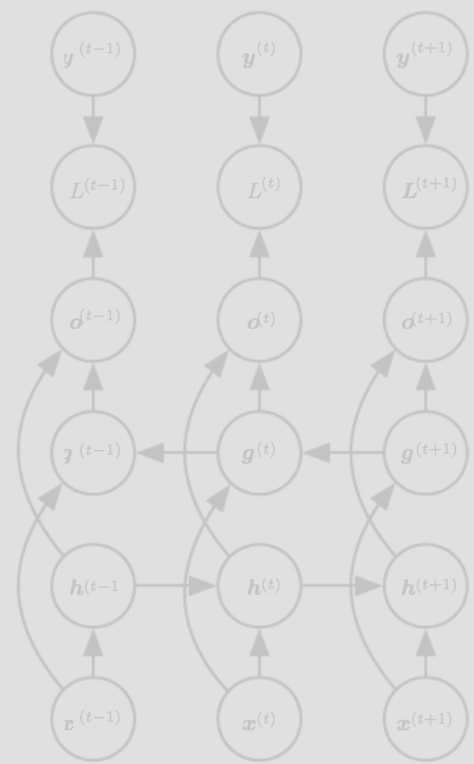
Methods

- This motivates the use of recurrent networks

Unidirectional recurrent network



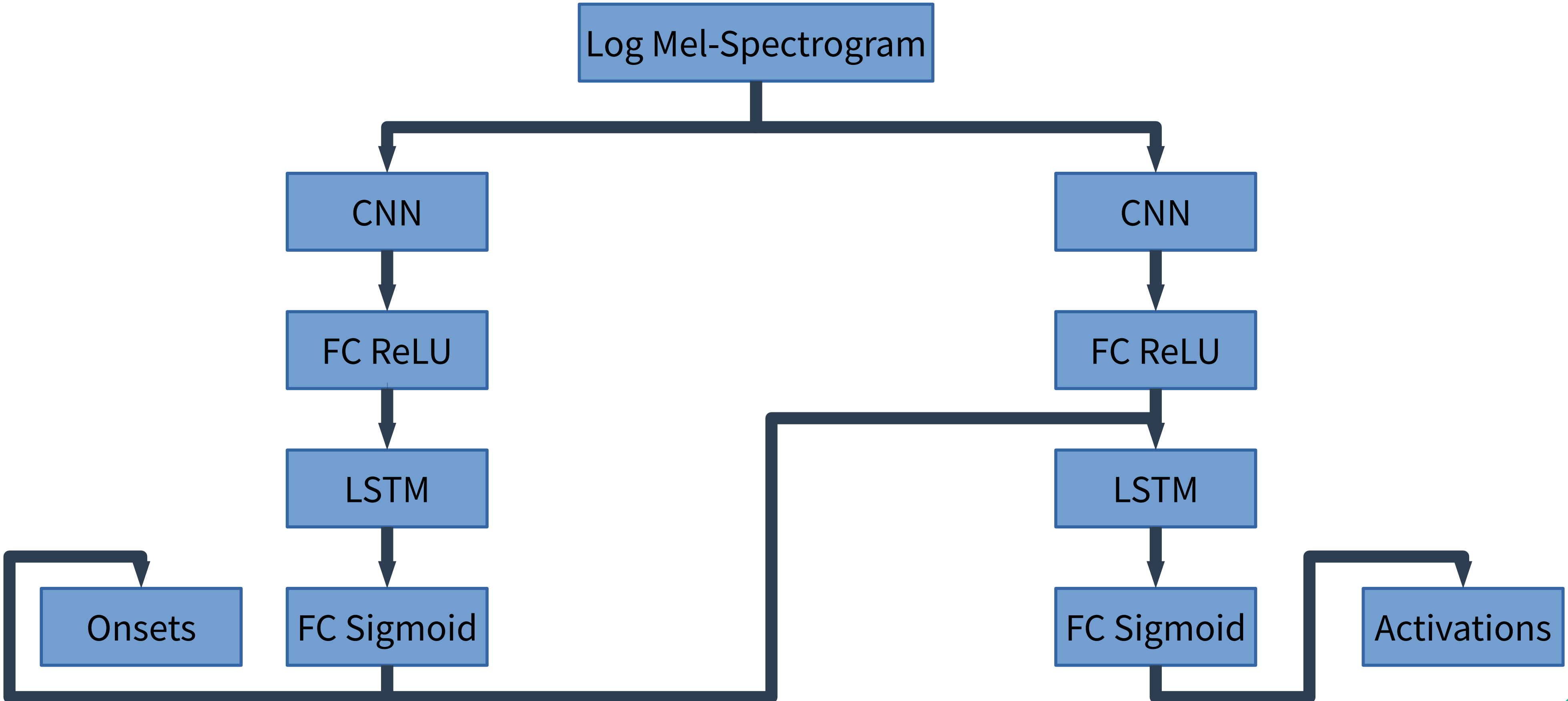
Bidirectional recurrent network



Goodfellow et al (2016)

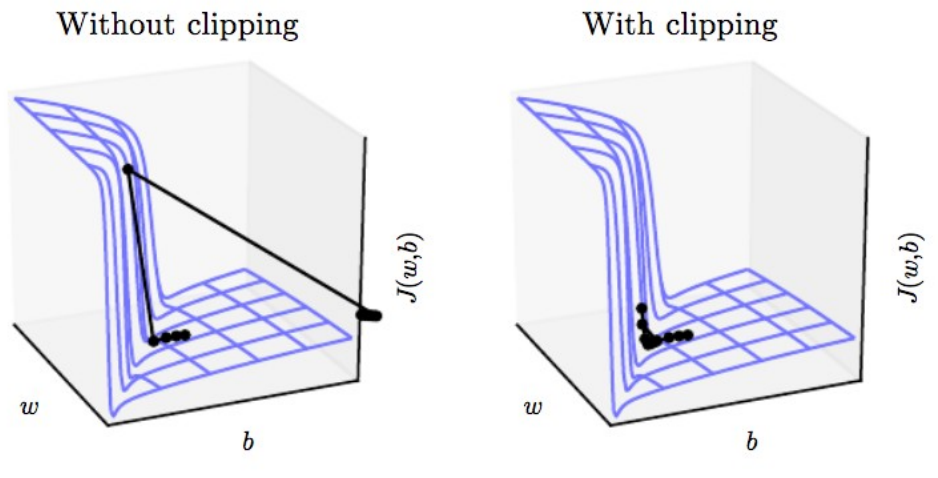
Network architecture

- Based on (Hawthorne et al 2018)
- Our LSTM only looks into the past (online)

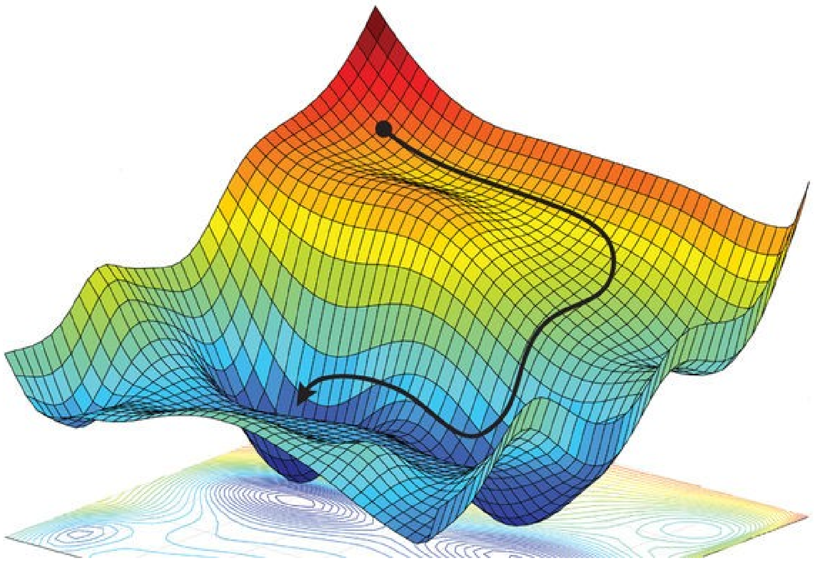


Training Strategy

- **Training using Adam optimizer**
  - Adaptive gradient descent algorithm
  - Learning rate scaling
  - Gradient clipping



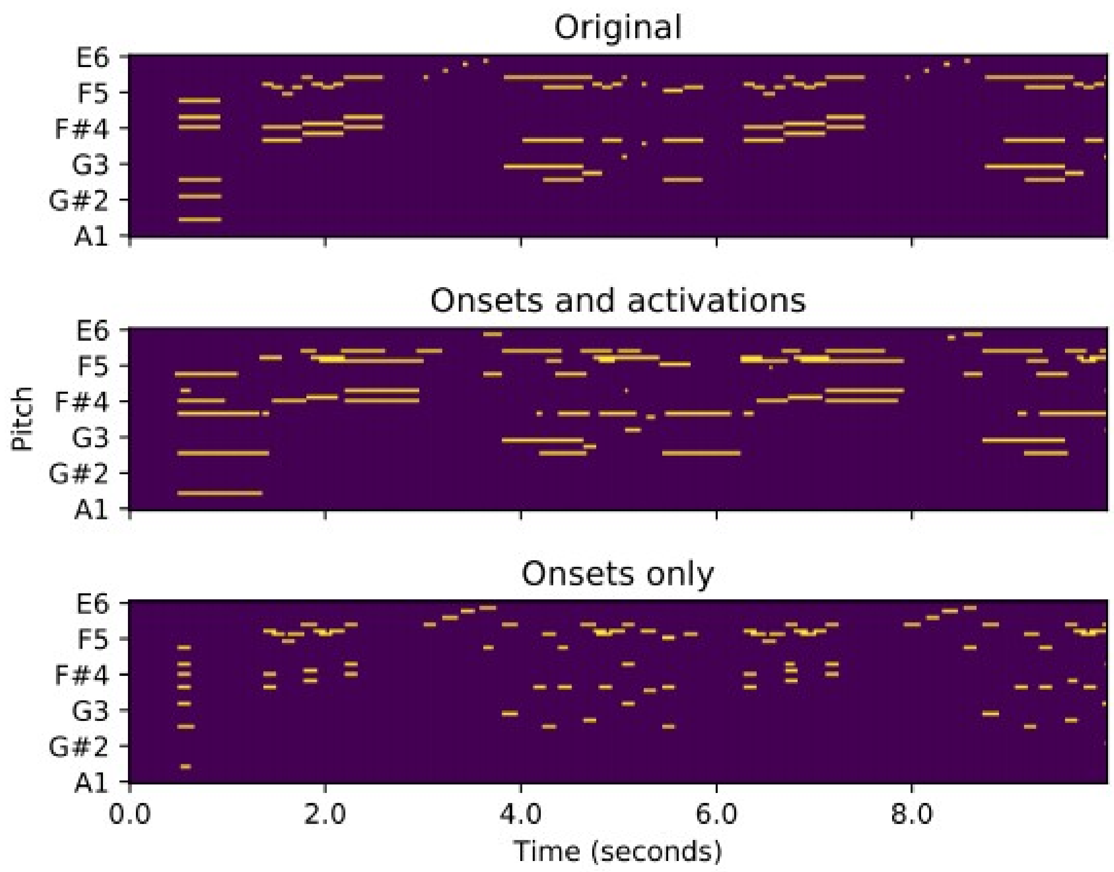
Goodfellow et al (2016)



https://www.sciencemag.org/

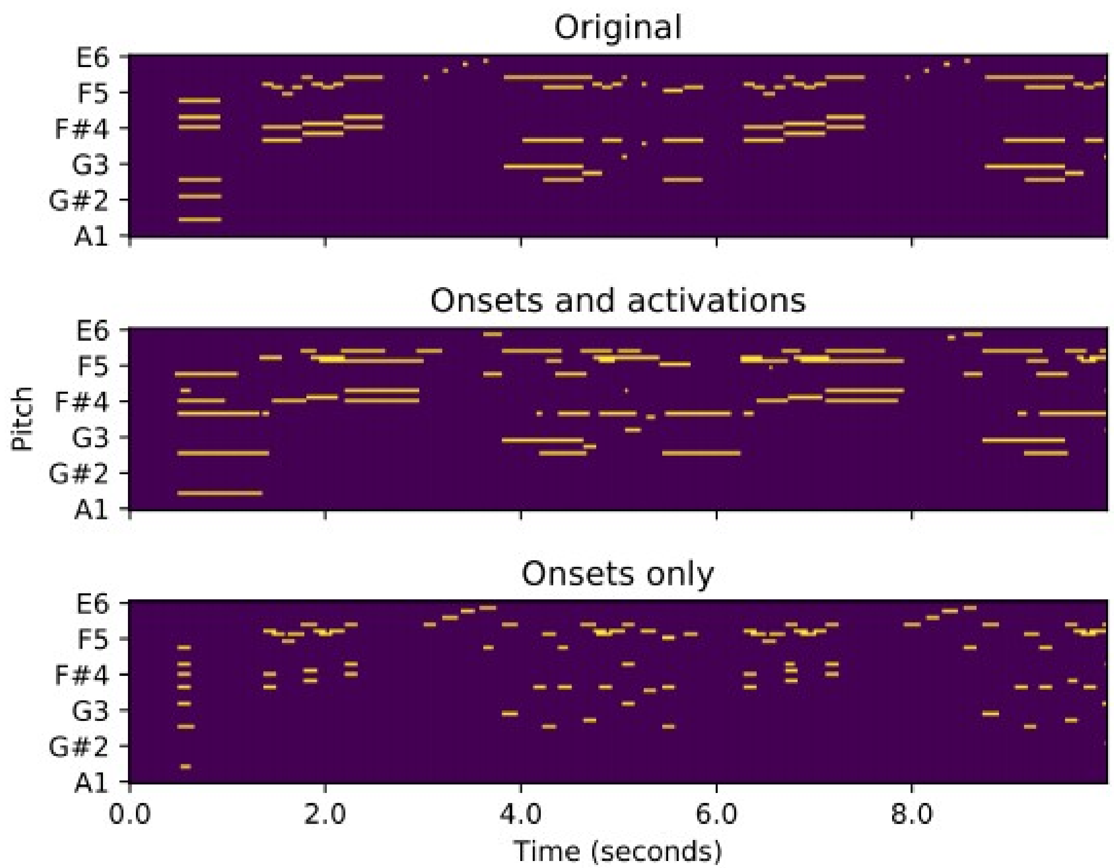
Results

- Transcription of Mozart Sonata from test set



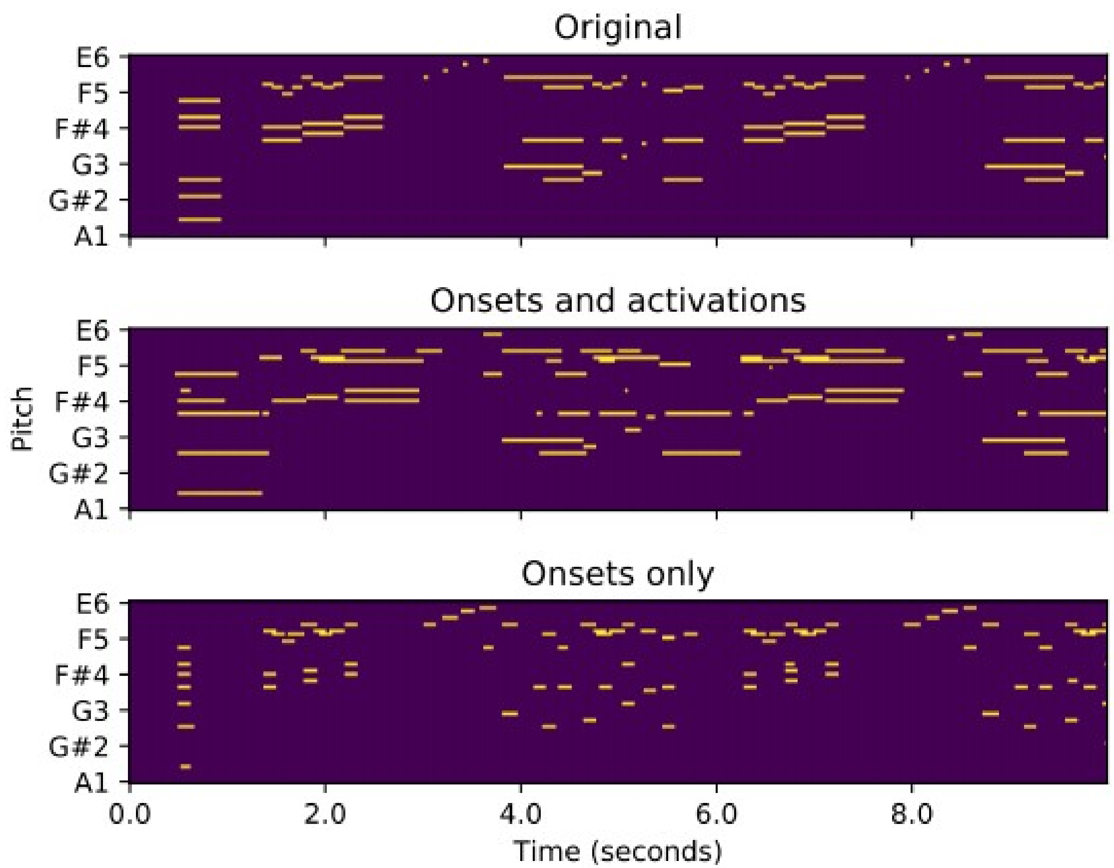
Results

- Transcription of Mozart Sonata from test set



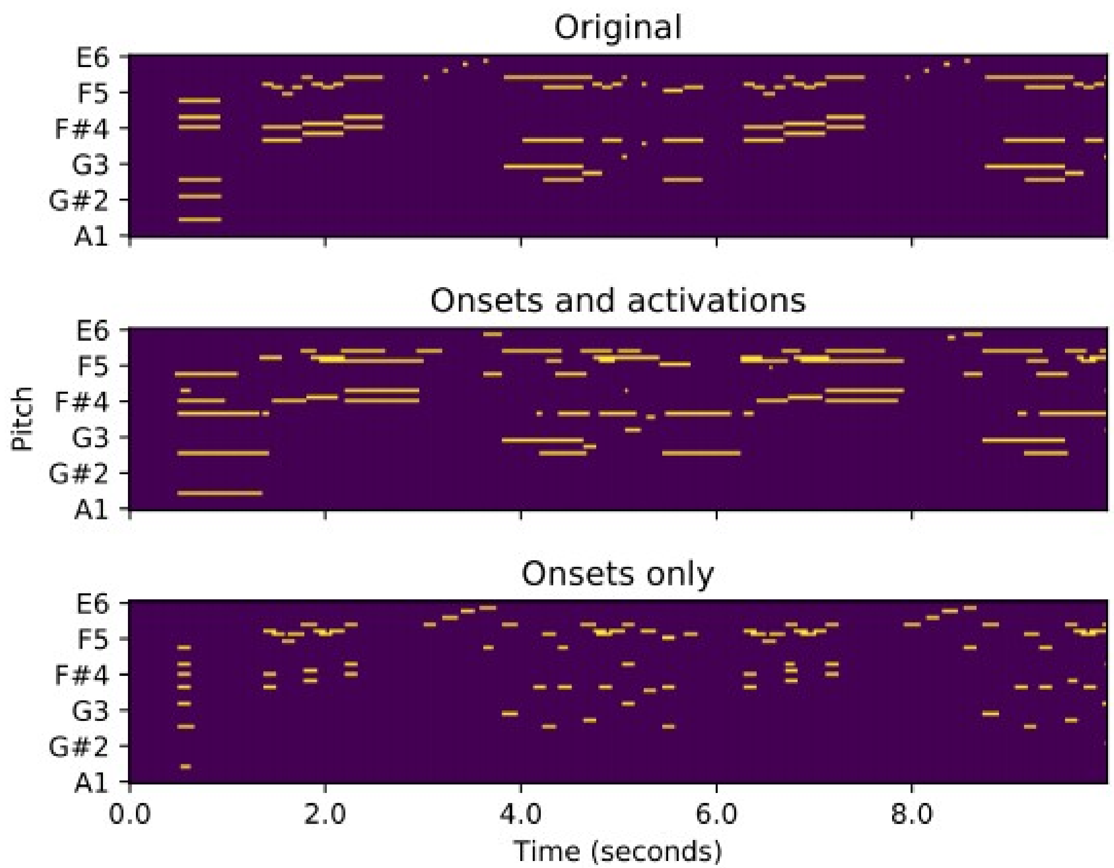
Results

- Transcription of Mozart Sonata from test set



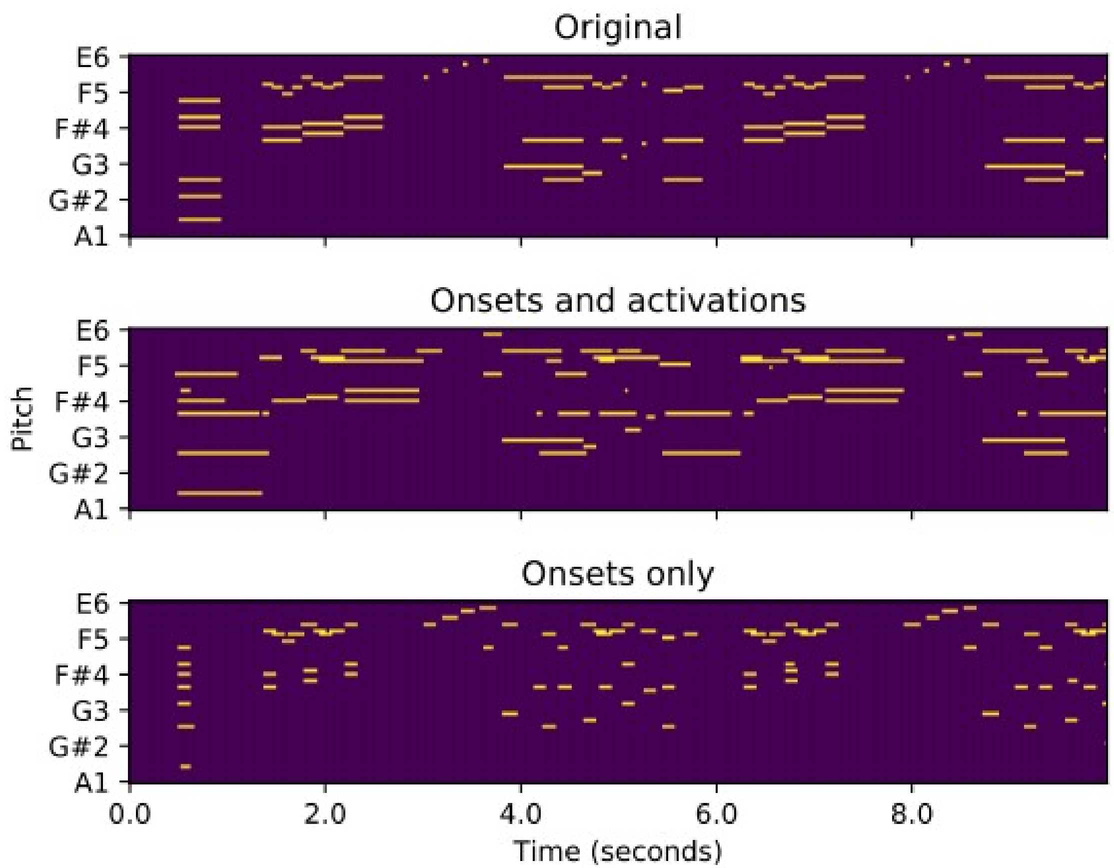
Results

- Transcription of Mozart Sonata from test set



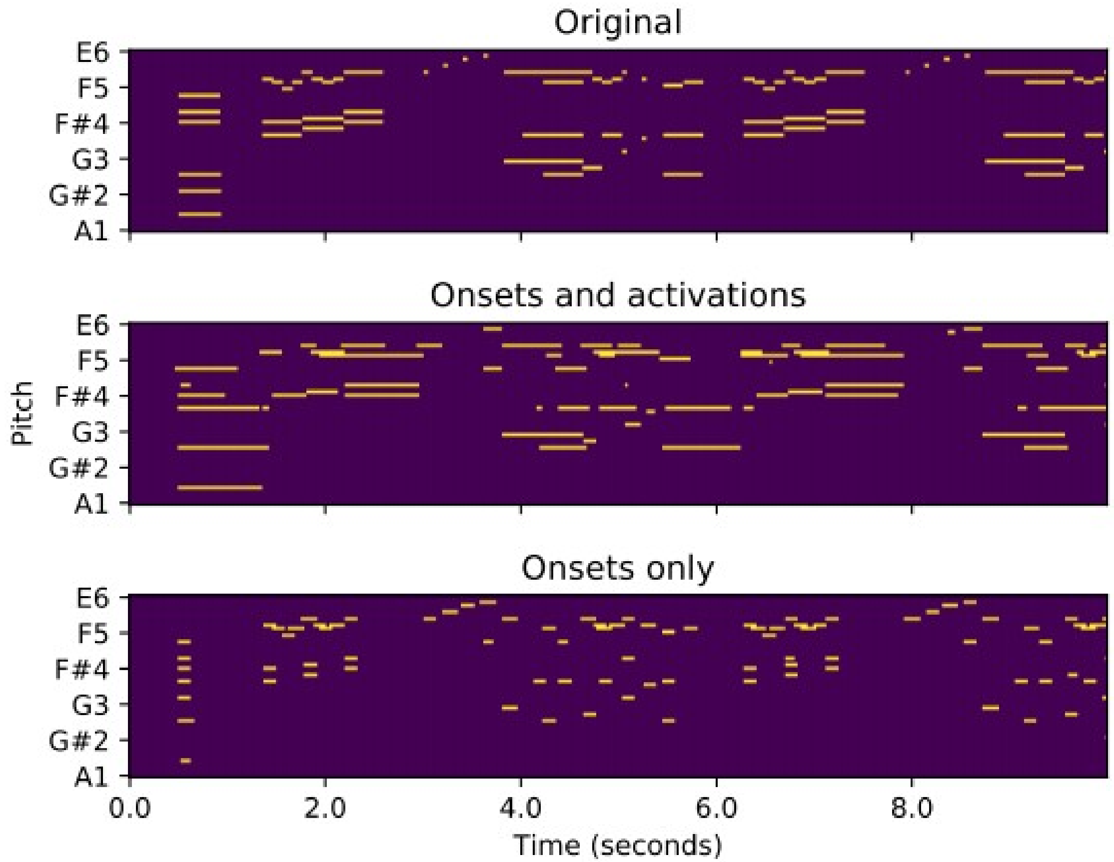
Results

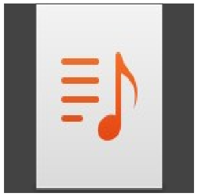
- Transcription of Mozart Sonata from test set



Results

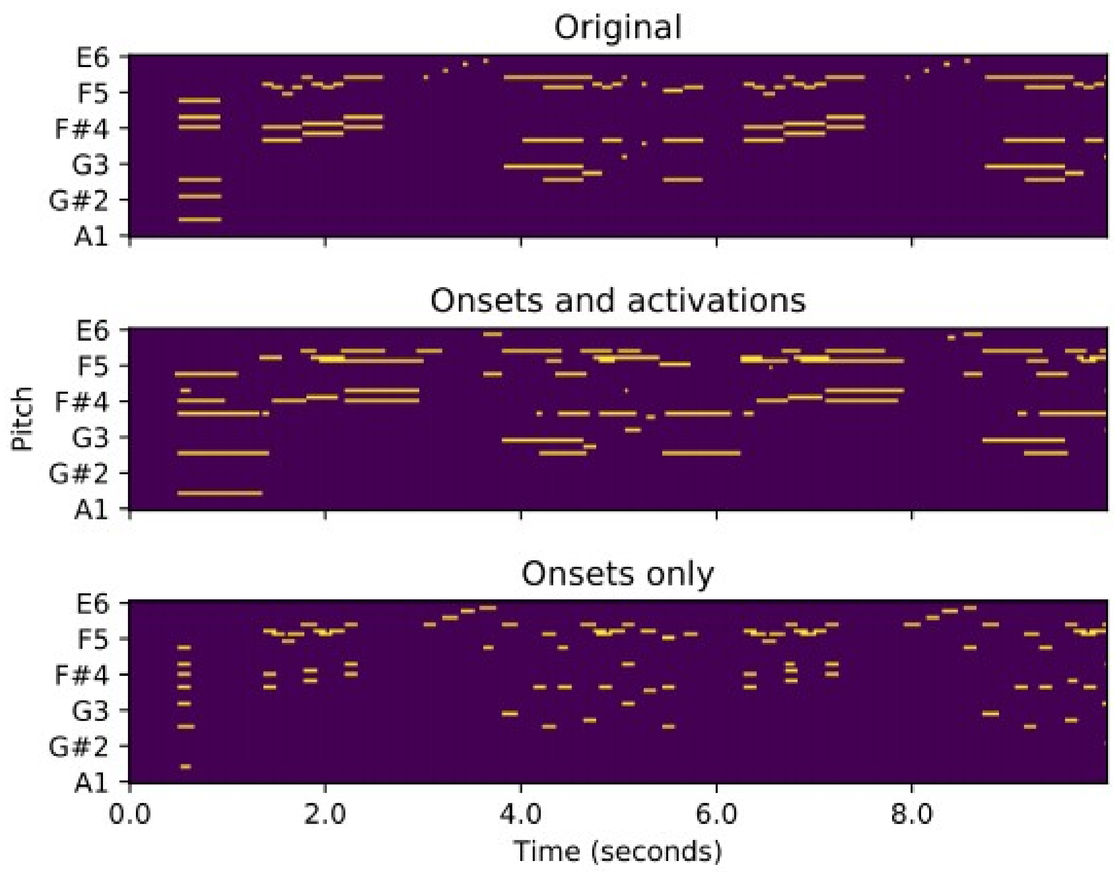
- Transcription of Mozart Sonata from test set





Results

- Transcription of Mozart Sonata from test set



Results

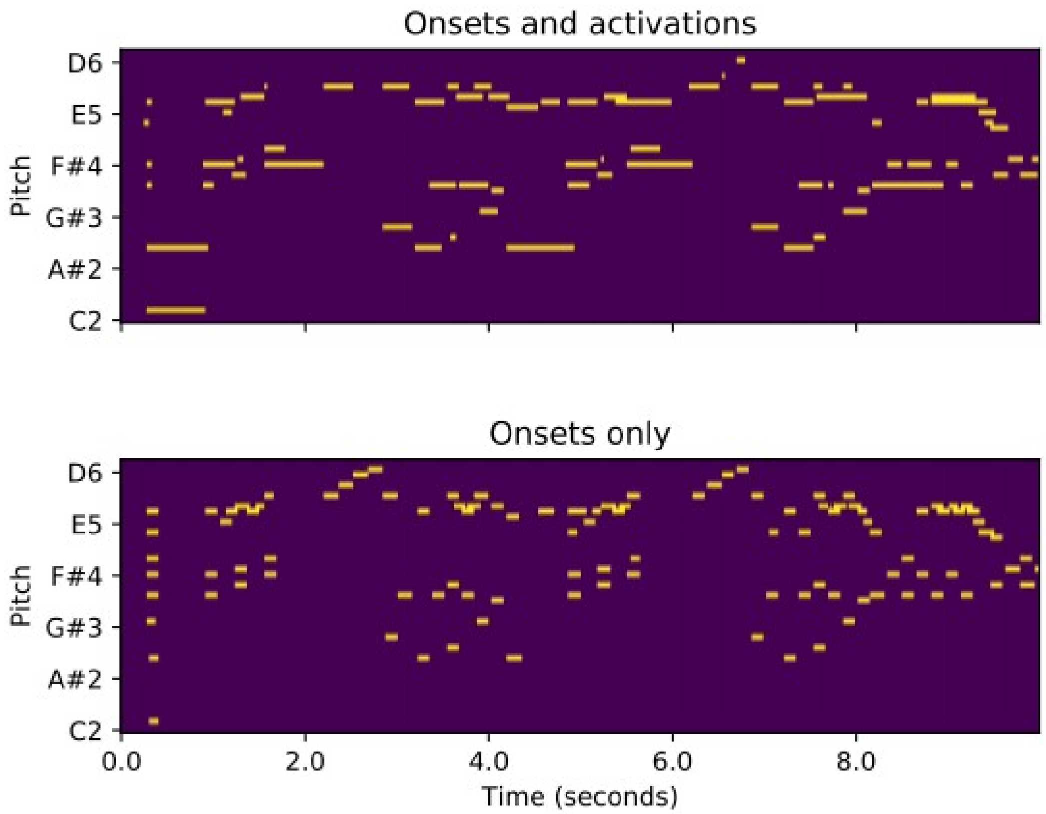
- Mozart Sonata performed by Glenn Gould

Results

- Mozart Sonata performed by Glenn Gould

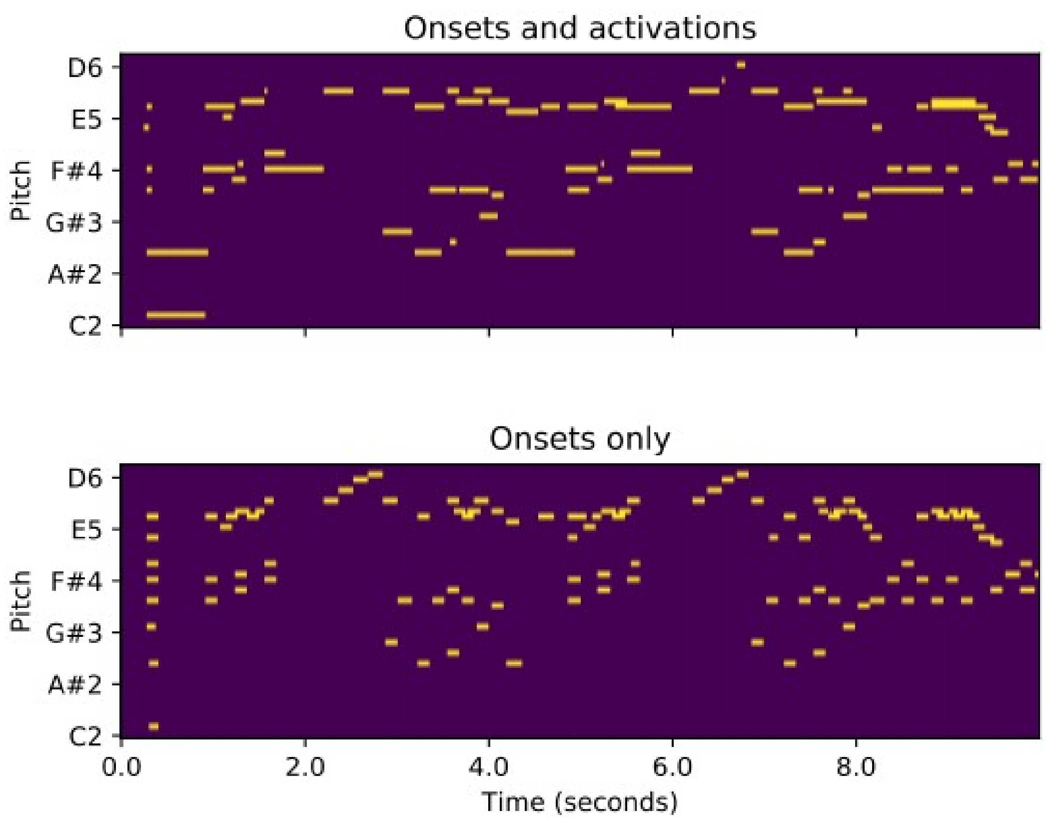
Results

- Transcription of Mozart Sonata performed by Glenn Gould



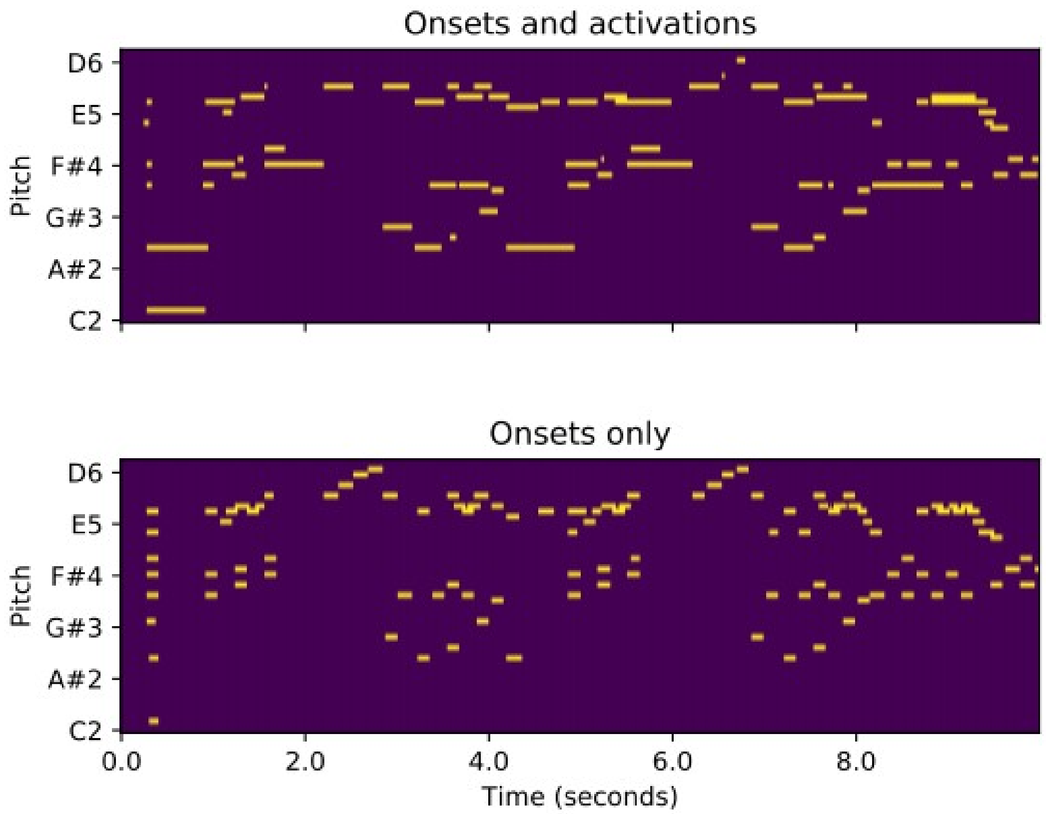
Results

- Transcription of Mozart Sonata performed by Glenn Gould



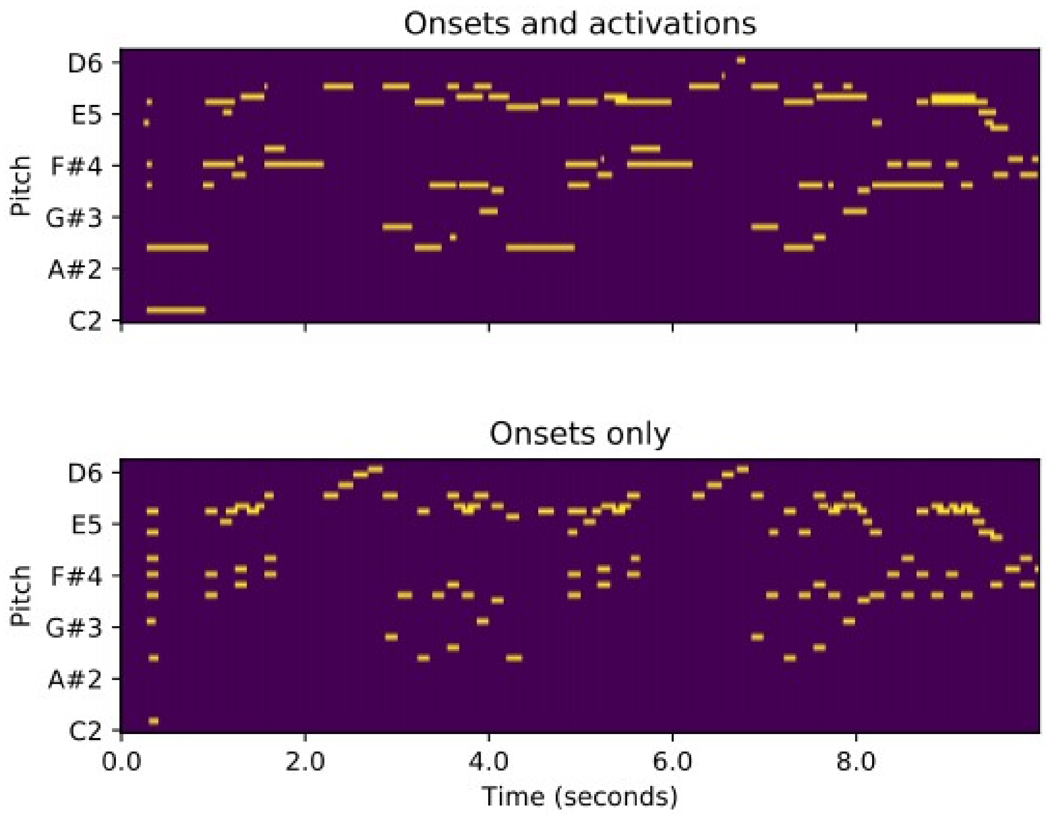
Results

- Transcription of Mozart Sonata performed by Glenn Gould



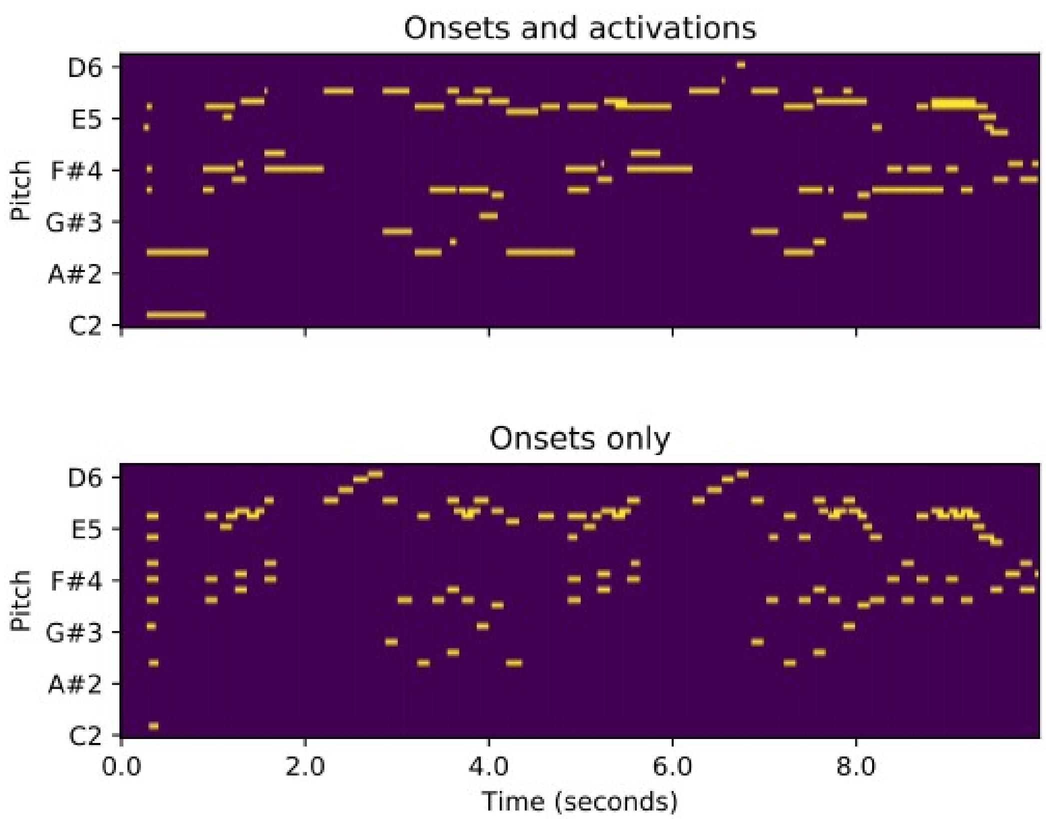
Results

- Transcription of Mozart Sonata performed by Glenn Gould



Results

- Transcription of Mozart Sonata performed by Glenn Gould



Augmentation

- Notable difference between example in our dataset and the recording performed by Gould

Augmentation

- Recording from dataset

Augmentation

- Recording from dataset

Augmentation

- Gould recording

Augmentation

- Gould recording

Augmentation

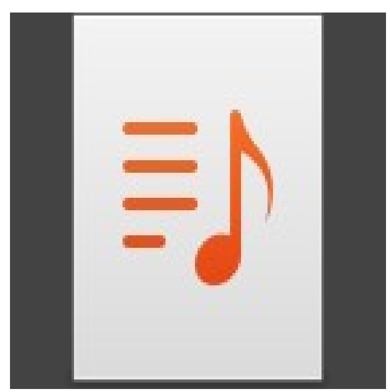
- **Gould recording**
  - Piano tuning is sharper
  - Played at a faster tempo
  - More room sound, different microphone techniques

Augmentation

- **We can alter our data to be qualitatively like Gould's recordings**
  - Pitch shifting (here +25 cents)
  - Time compression (here 25% faster)
  - Reverb

Augmentation

- **We can alter our data to be qualitatively like Gould's recordings**
  – Pitch shifting (here +25 cents)
  – Time compression (here 25% faster)
  – Reverb

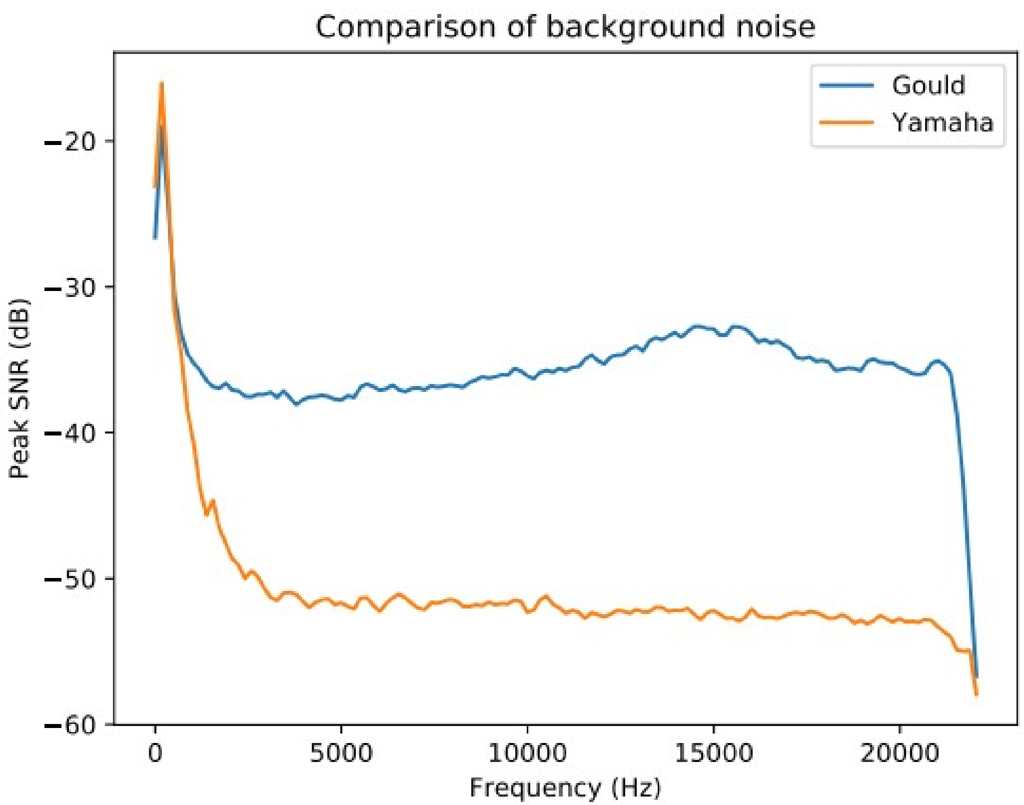


Augmentation

- **We can alter our data to be qualitatively like Gould's recordings**
  - Pitch shifting (here +25 cents)
  - Time compression (here 25% faster)
  - Reverb

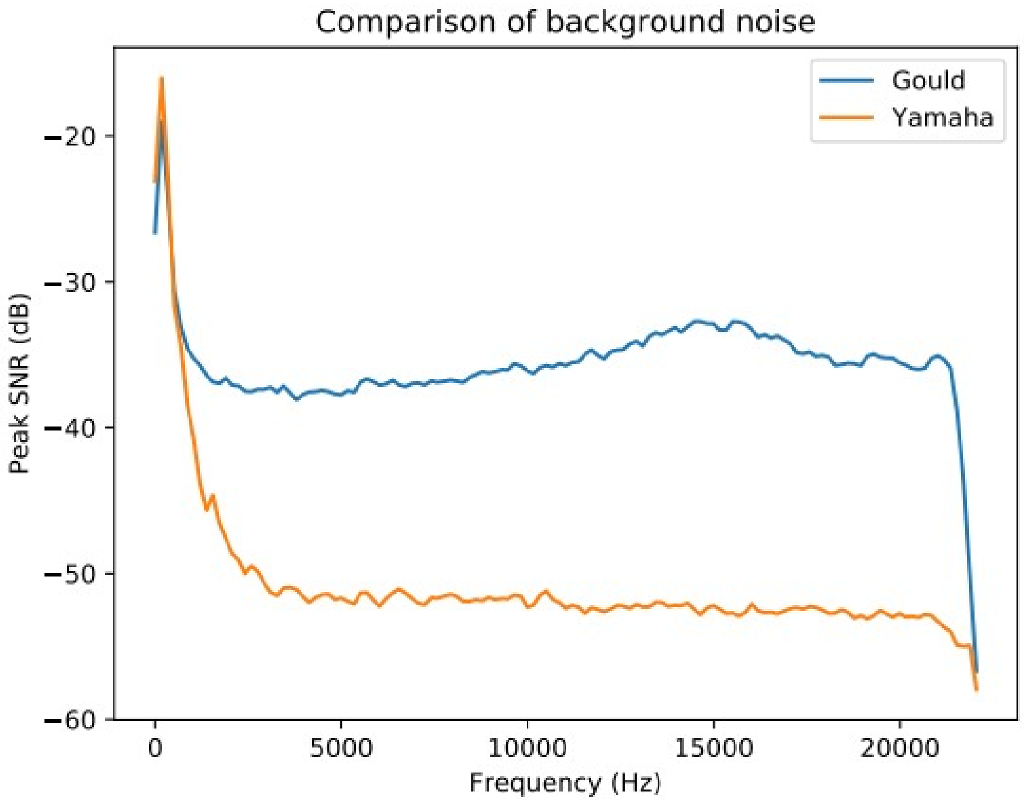
Regularization

- Some Gould recordings are much noisier than the Yamaha and MAPS datasets



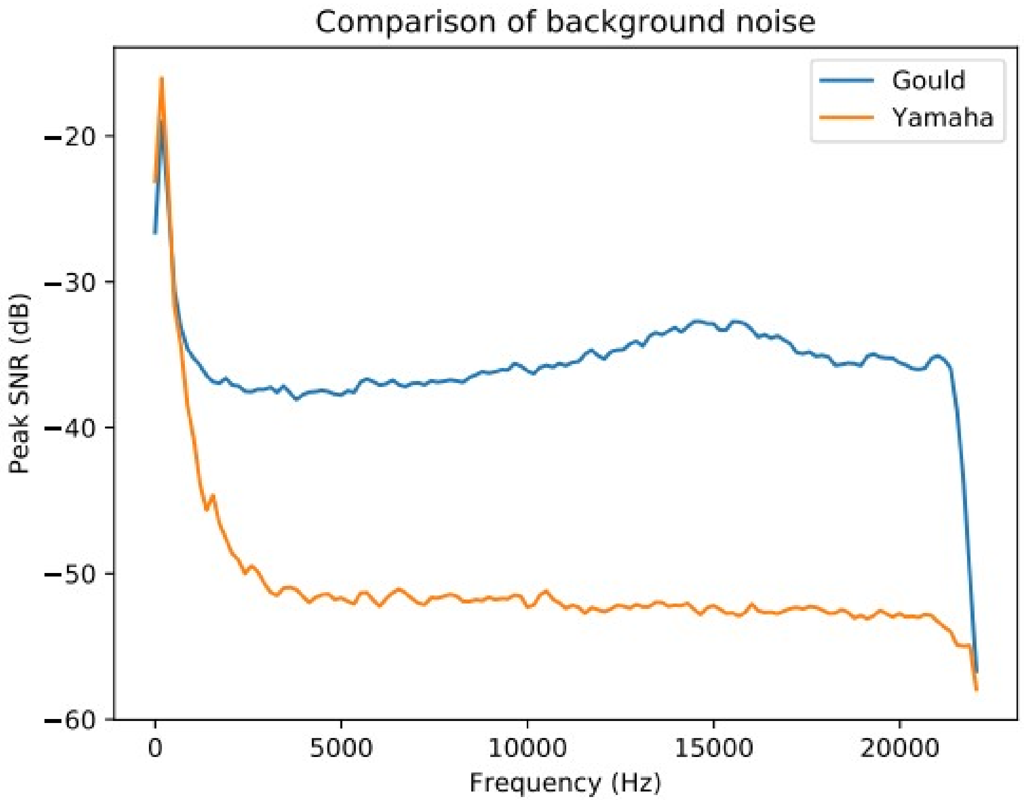
Regularization

- Some Gould recordings are much noisier than the Yamaha and MAPS datasets



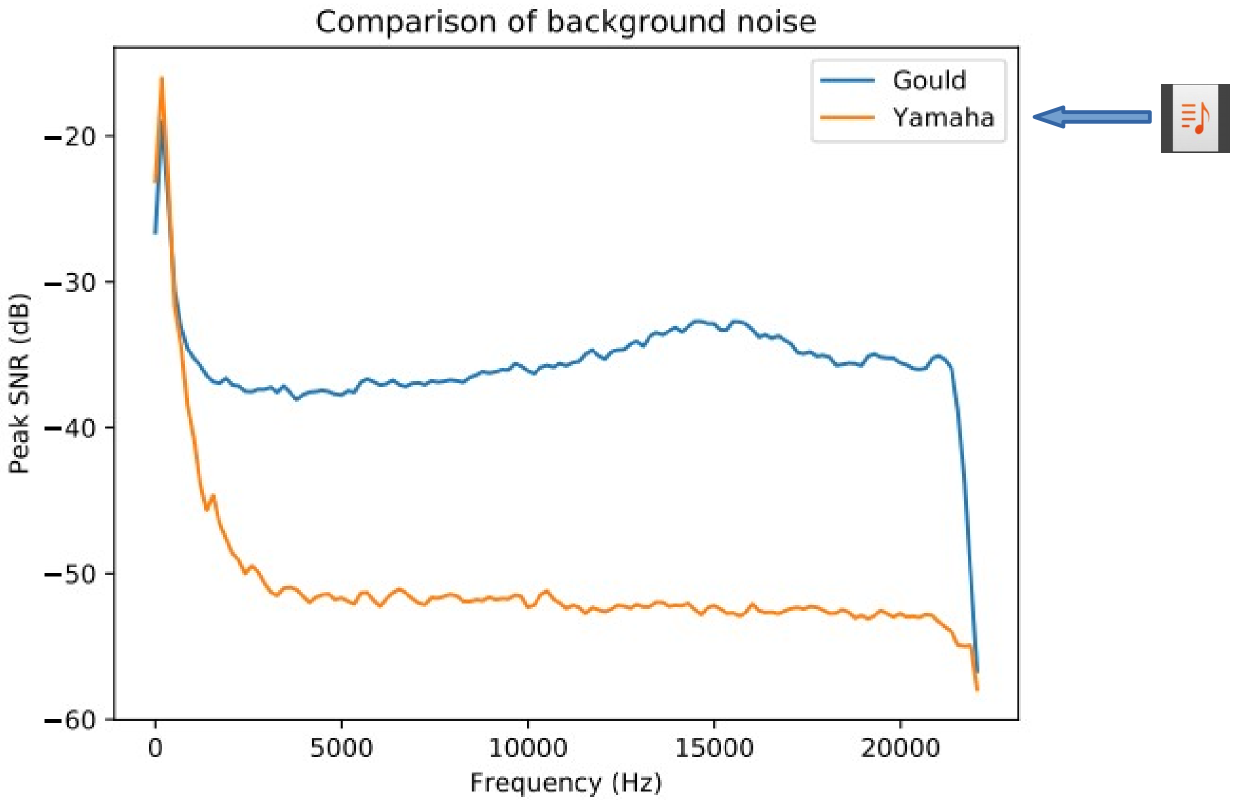
Regularization

- Some Gould recordings are much noisier than the Yamaha and MAPS datasets



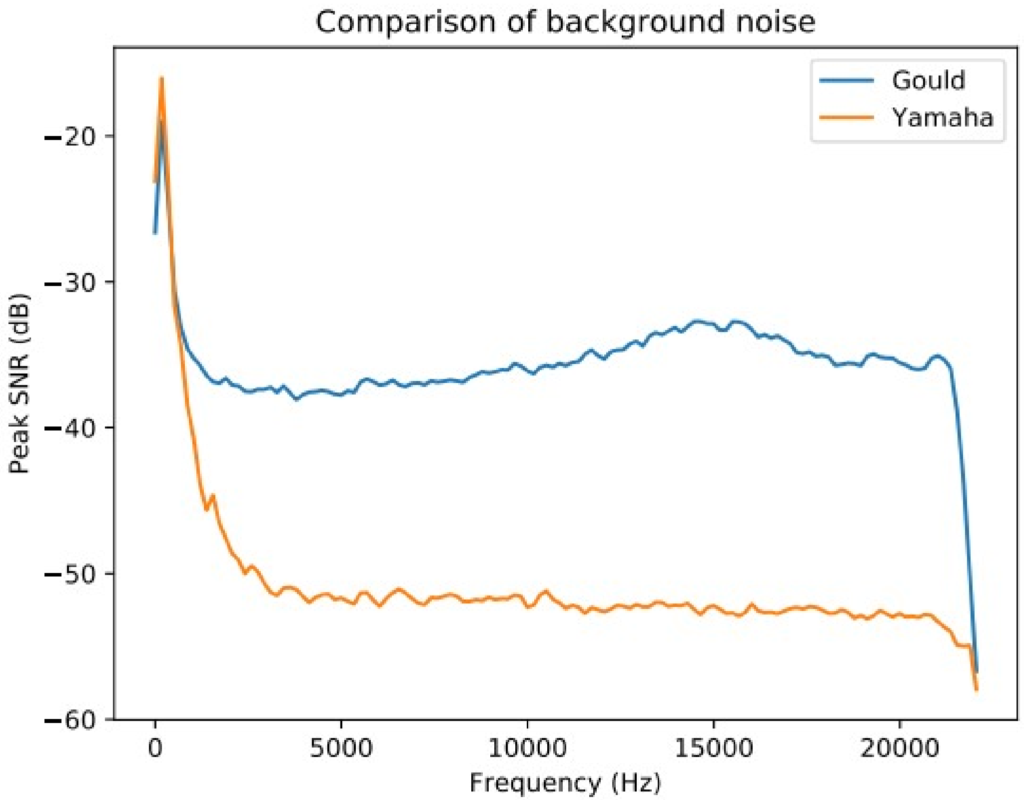
Regularization

- Some Gould recordings are much noisier than the Yamaha and MAPS datasets



Regularization

- Some Gould recordings are much noisier than the Yamaha and MAPS datasets



Regularization

- Transcription attempt without regularization

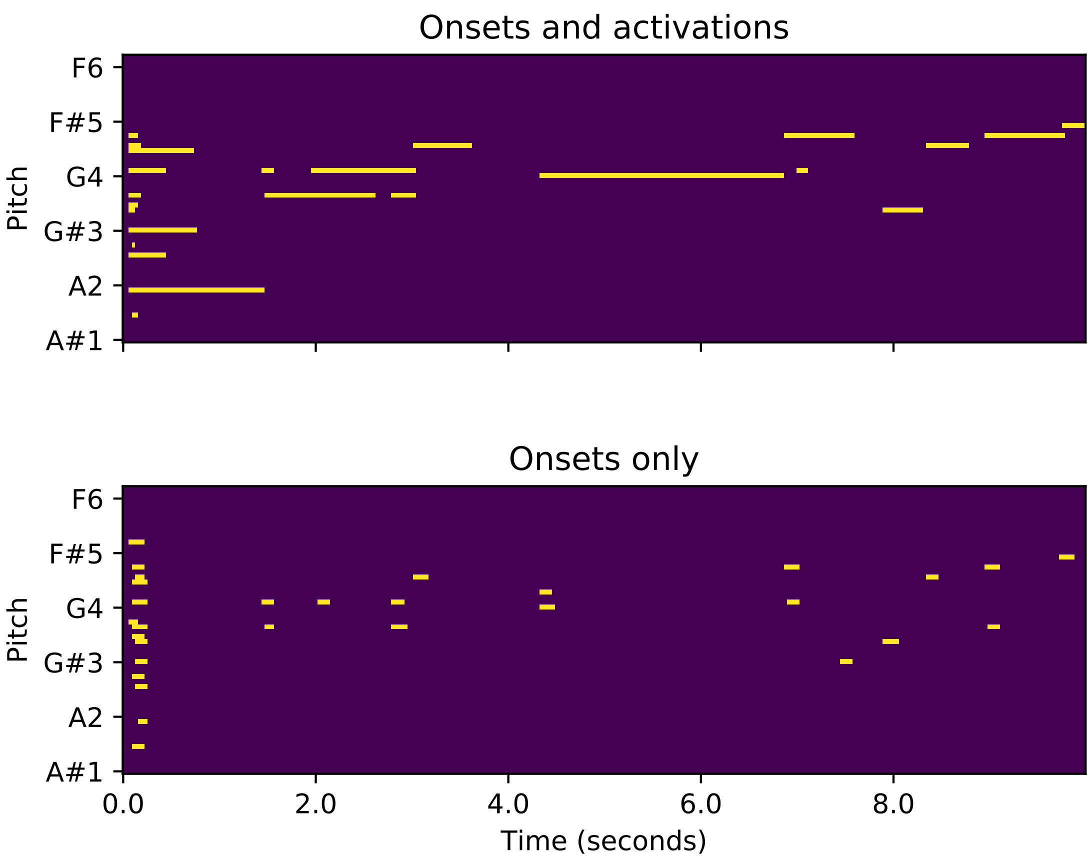
Regularization

- Transcription attempt without regularization

Original

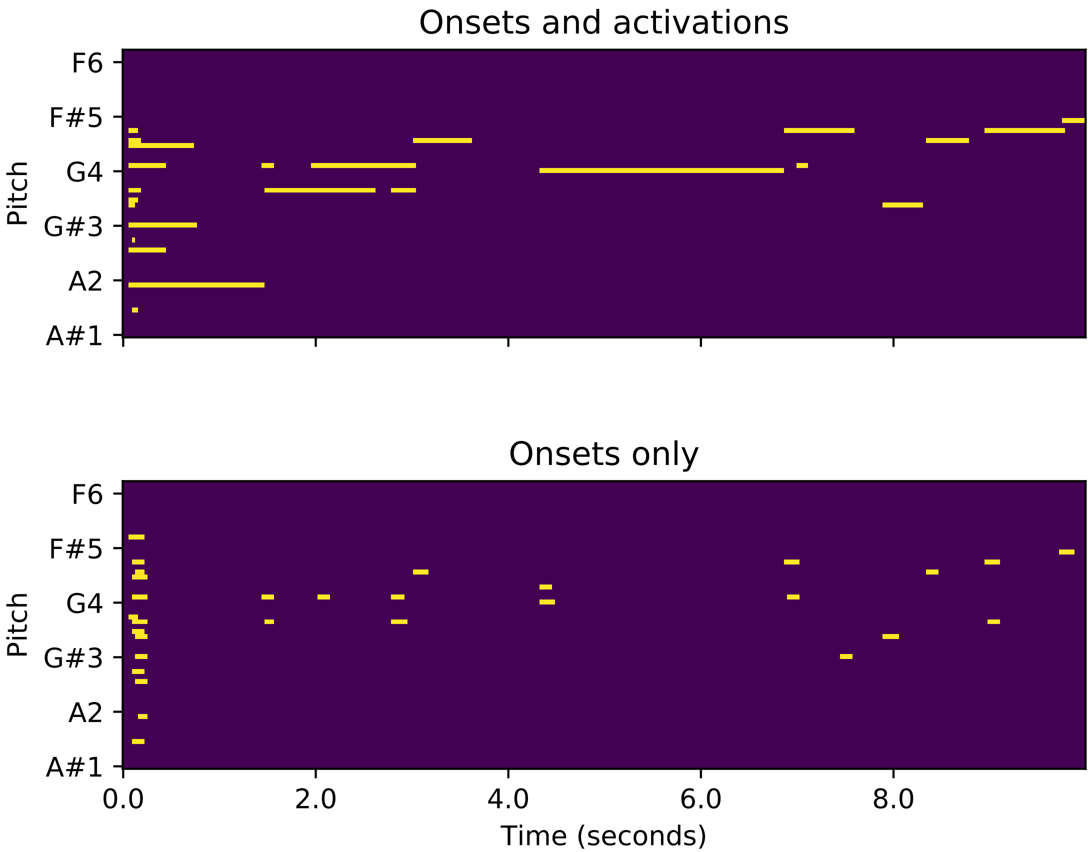
Regularization

- Transcription attempt without regularization



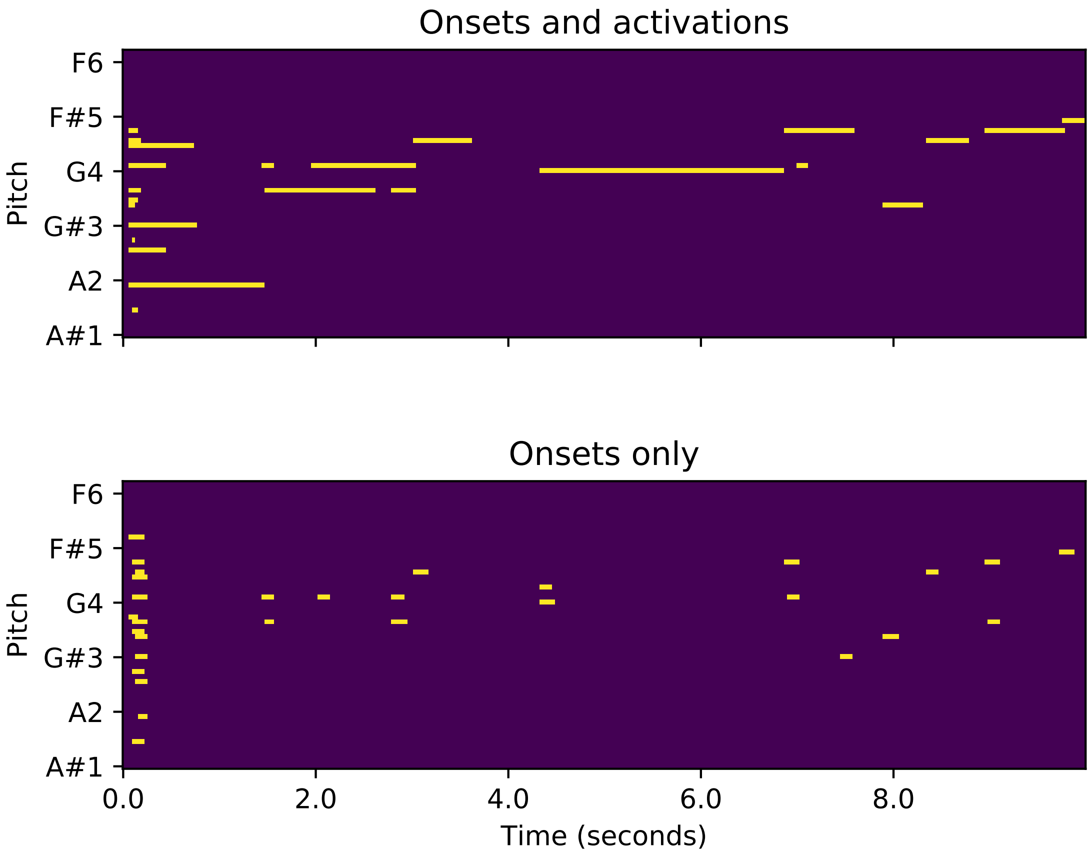
Regularization

- Transcription attempt without regularization



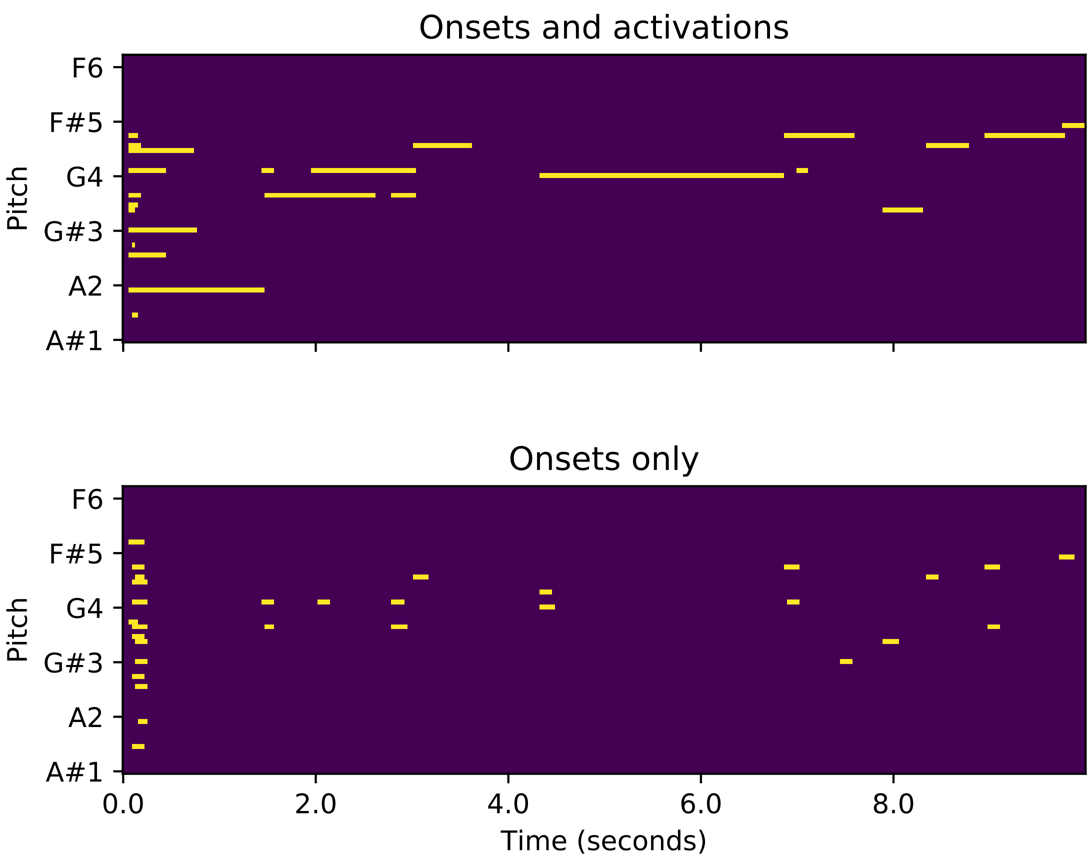
Regularization

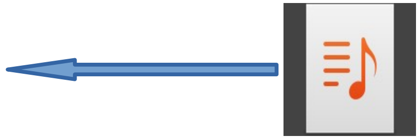
- Transcription attempt without regularization



Regularization

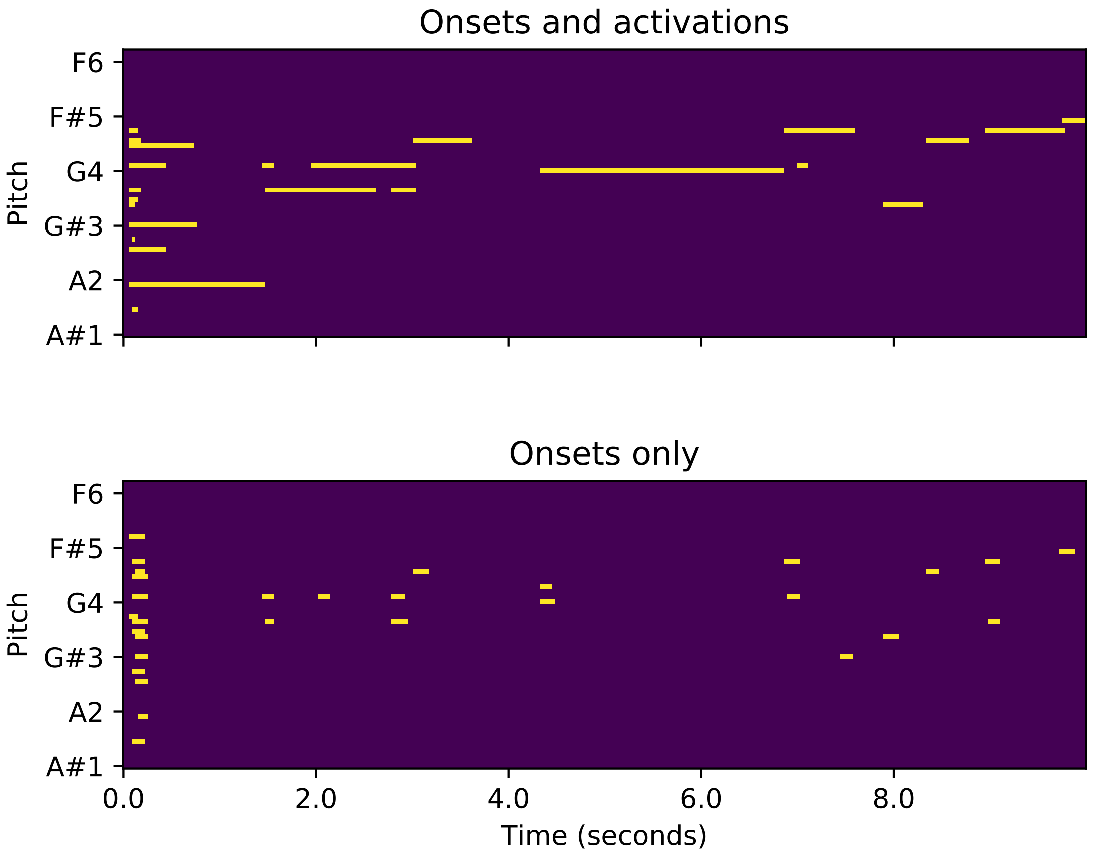
- Transcription attempt without regularization





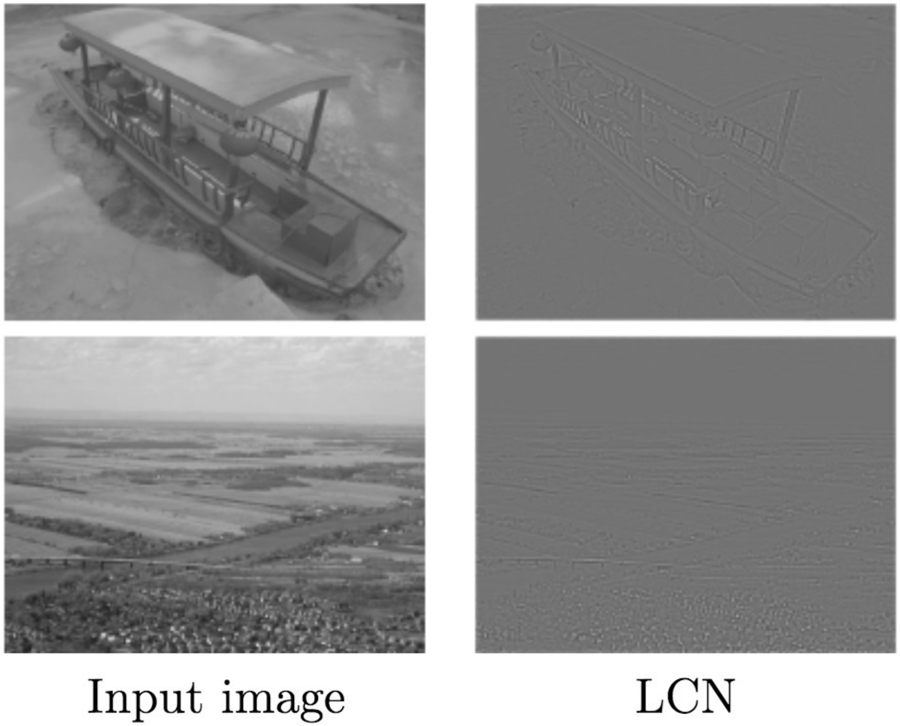
Regularization

- Transcription attempt without regularization



Regularization

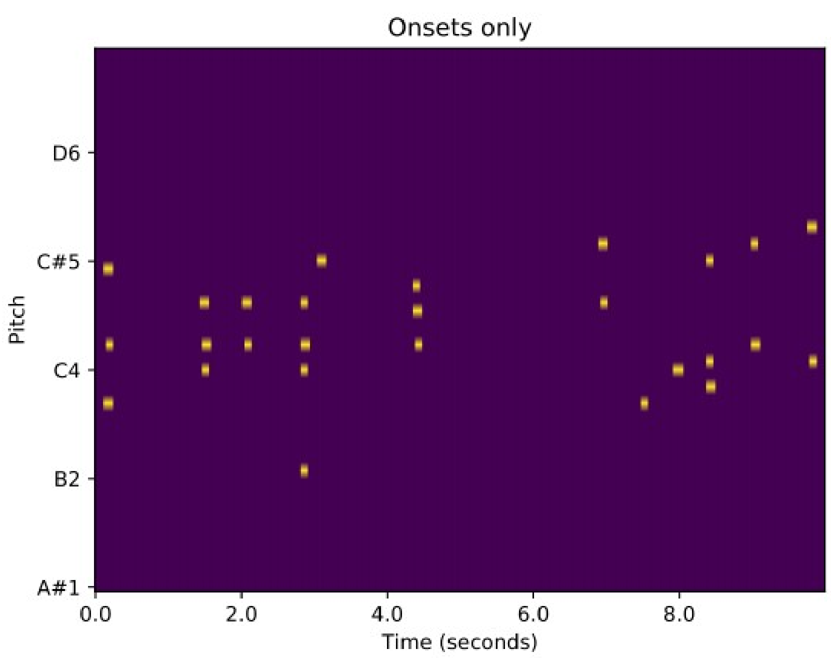
- **Prevent algorithm from over-fitting to noise-free data**
  - Add <u>unique</u> noise to each training example
- **Make algorithm invariant to recording's dynamic range**
  - Perform local contrast normalization (LCN)



Goodfellow et al (2016)

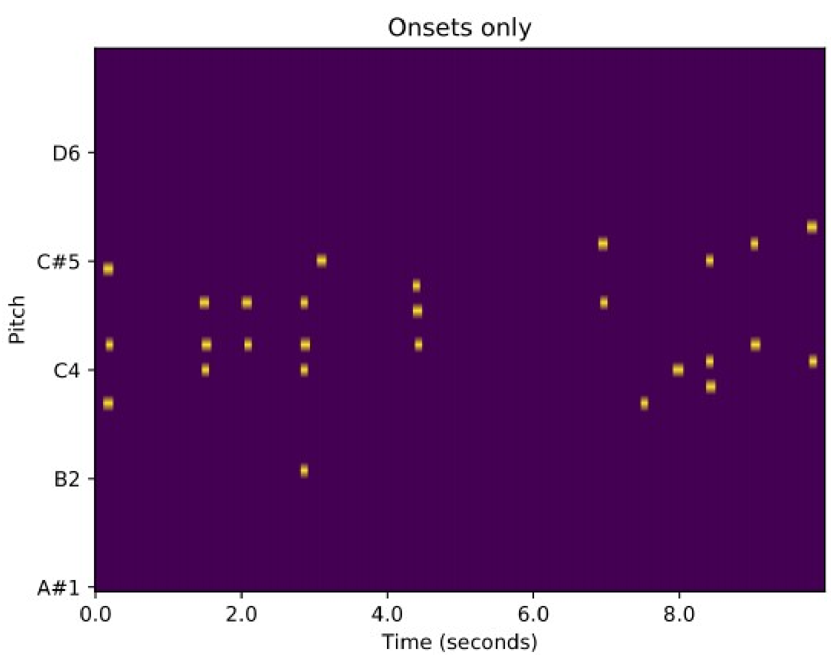
Regularization

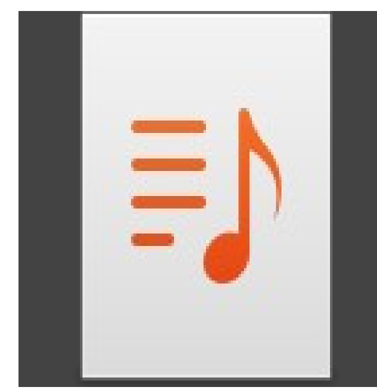
- **Transcription attempt with regularization**
  - Added unique noise to data points giving SNR of -40 dB
  - Applied LCN



Regularization

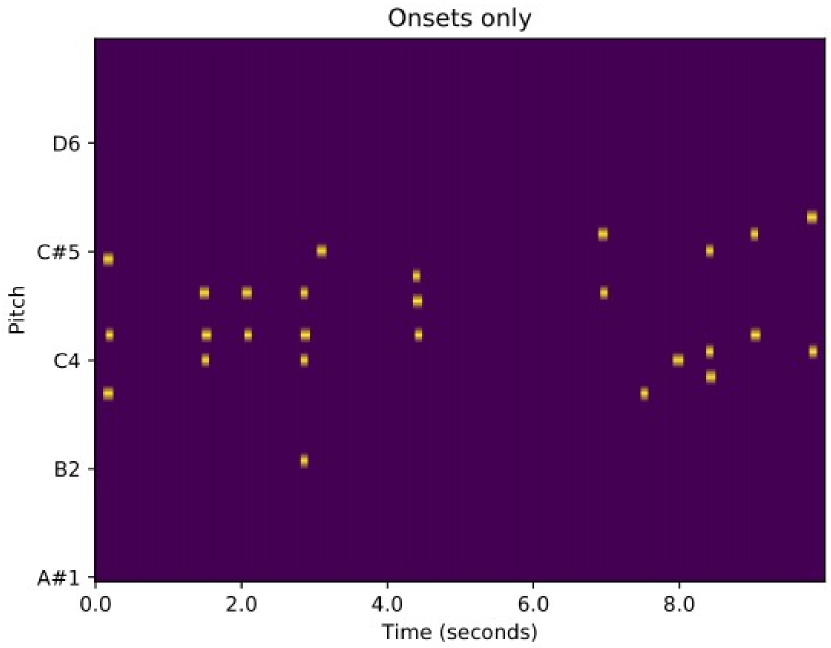
- **Transcription attempt with regularization**
  - Added unique noise to data points giving SNR of -40 dB
  - Applied LCN





Regularization

- **Transcription attempt with regularization**
  - Added unique noise to data points giving SNR of -40 dB
  - Applied LCN



Evaluation

F-Measure Scores: With and without regularization

	Activation	Onset	Note with offset
YAMAHA DKV & MAPS	.6451	.8268	.2847
YAMAHA DKV & MAPS with augmentations and NR	.6069	.7651	.2279
YAMAHA DKV & MAPS with augmentations NR and LCN	.5973	.7367	.2127
Google O&F 2018 (MAPS and offline)	.7830	.8229	.5022
Kelz et al 2016 (MAPS)	.7160	.5094	.2314

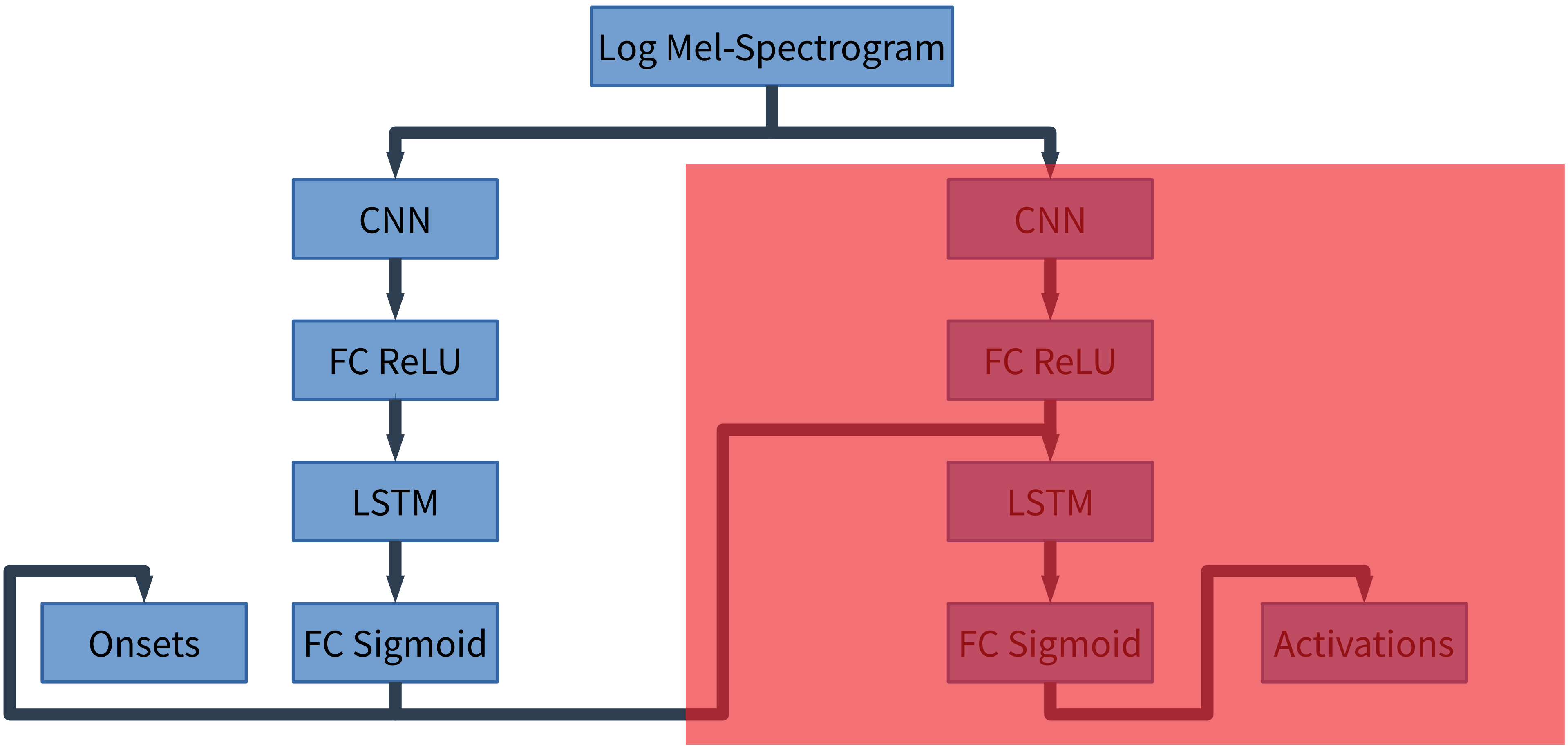
Evaluation

F-Measure Scores: evaluated on augmented data only

	Activation	Onset	Note with offset
YAMAHA DKV & MAPS	.5594	.7052	.1886
YAMAHA DKV, MAPS with augmentations	.5236	.7252	.1894
YAMAHA DKV & MAPS with augmentations and NR	.5509	**.7429**	**.1952**
YAMAHA DKV & MAPS with augmentations, NR and LCN	**.5604**	.7189	.1879

Model Revision

- Remove activations detector when only looking into the past



Architecture evaluation

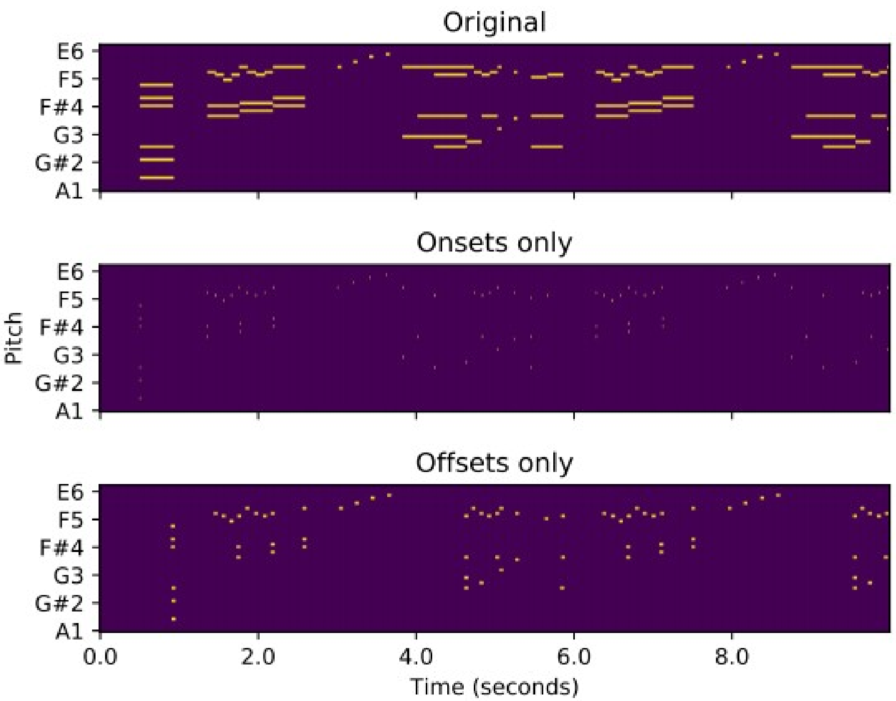
- **Inference complexity (con)**
  - $O(n^2)$ where n ~ 7000 ... expensive!
- **Model size (con)**
  - Contains about 80 million parameters
- **Algorithm performance (pro)**
  - State of the art (Hawthorne et al 2018)

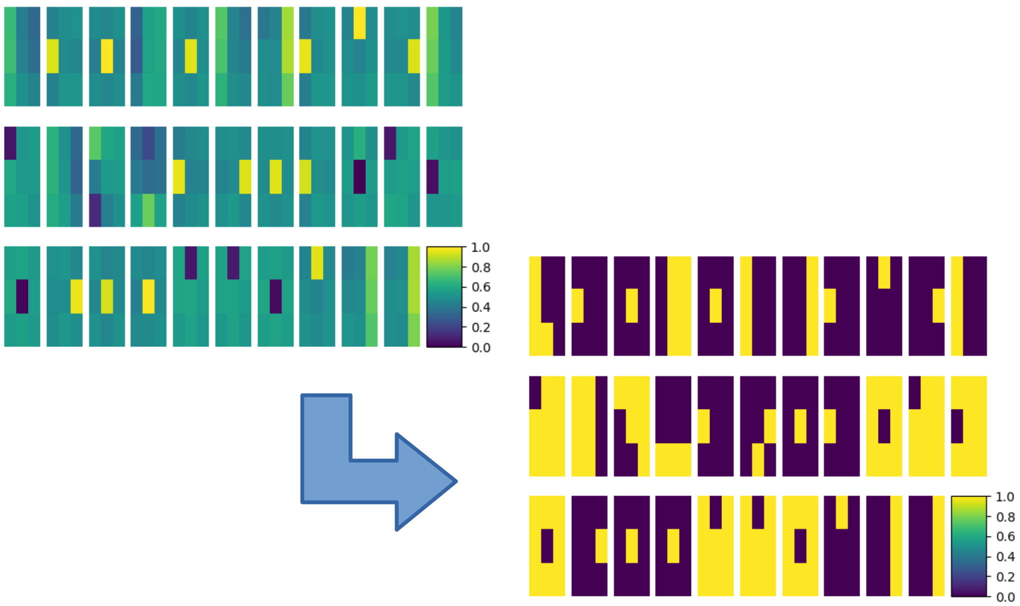
Conclusion

- **Onset estimation much easier than activation estimation**
  - Especially for "online" algorithms
  - Activation estimation perhaps ill-defined
- **The fewer assumptions about recording conditions, the better**
  - Training on noisy data can help
  - Evaluate on recordings closer to true recordings

Future Work

- Investigate offset detector
- Diminish model size and inference complexity

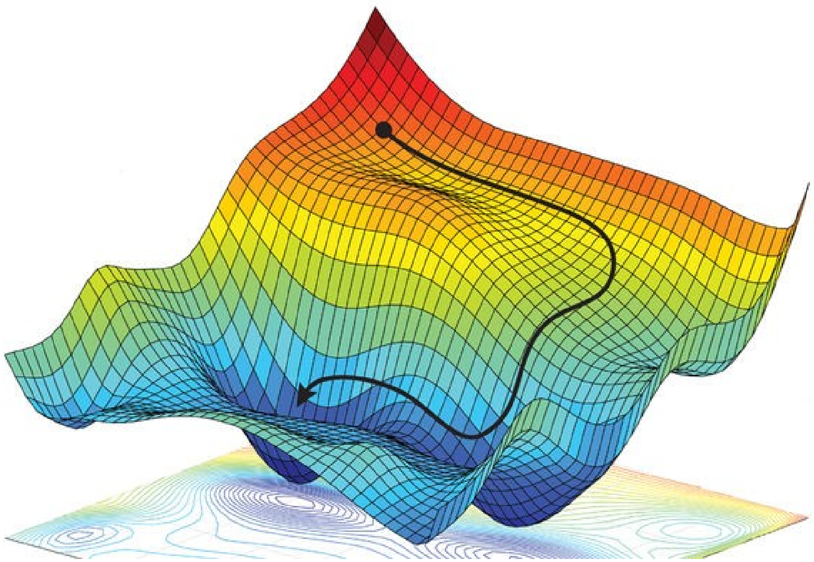




Thank You

Training Strategy

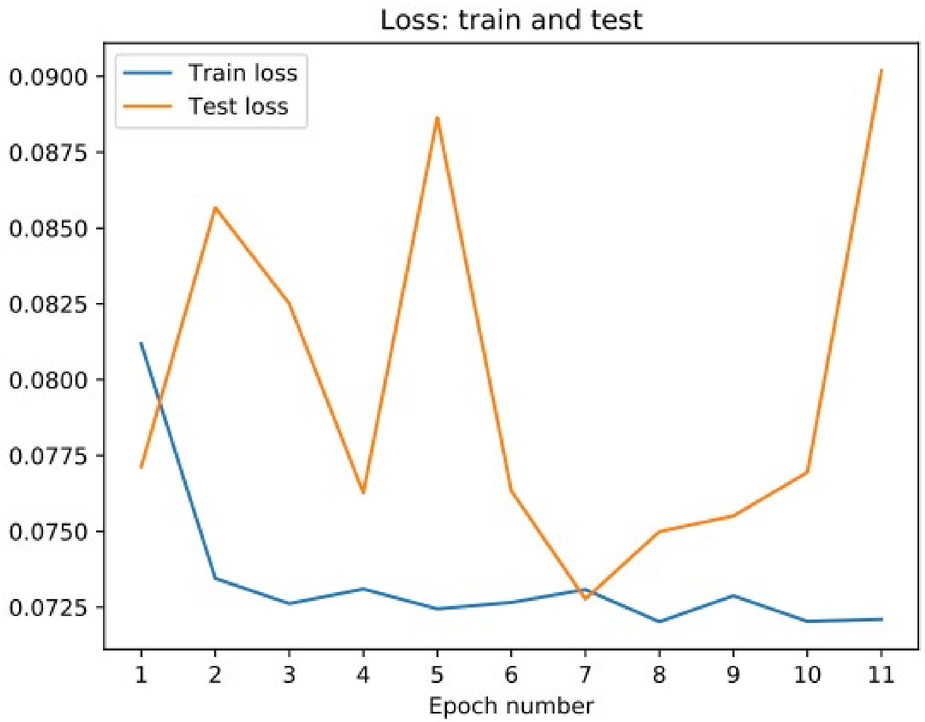
- **Training using Adam optimizer**
  - Adaptive gradient descent algorithm
  - Learning rate scaling
  - Gradient clipping

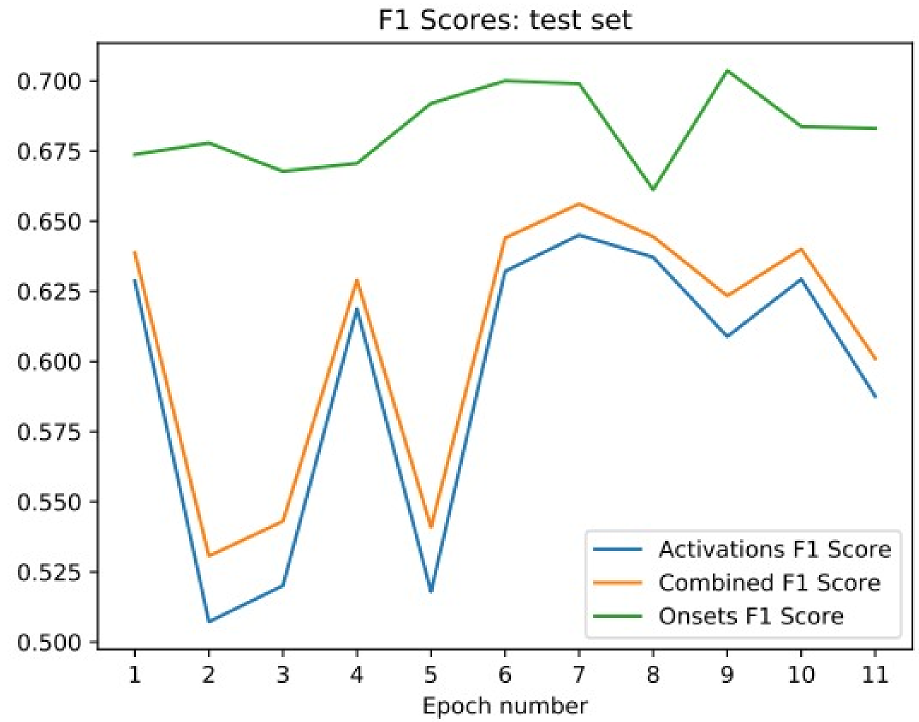


https://www.sciencemag.org/

Training Strategy

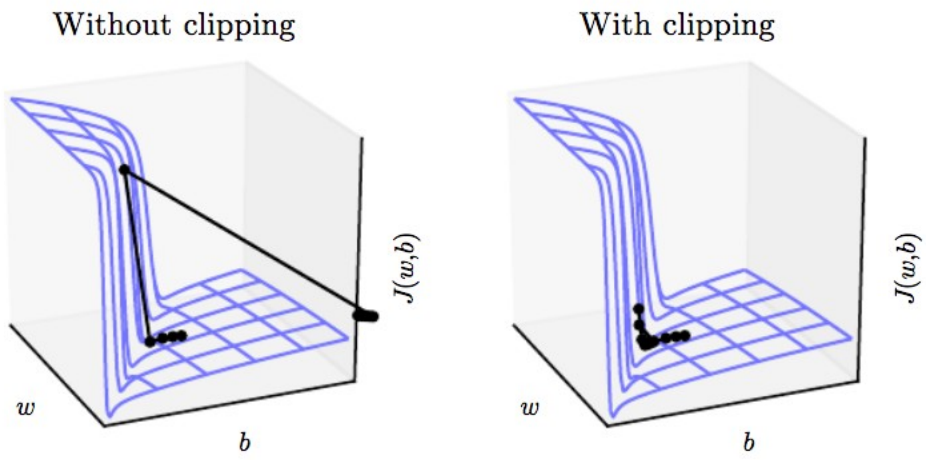
- Train until test loss reaches minimum
- Often corresponds to a desirable model





Advice for Training

- RNN needs gradient clipping
- Adaptive optimizers (e.g., Adam) need learning rate decay
- Data normalization can actually be hurtful



Goodfellow et al (2016)