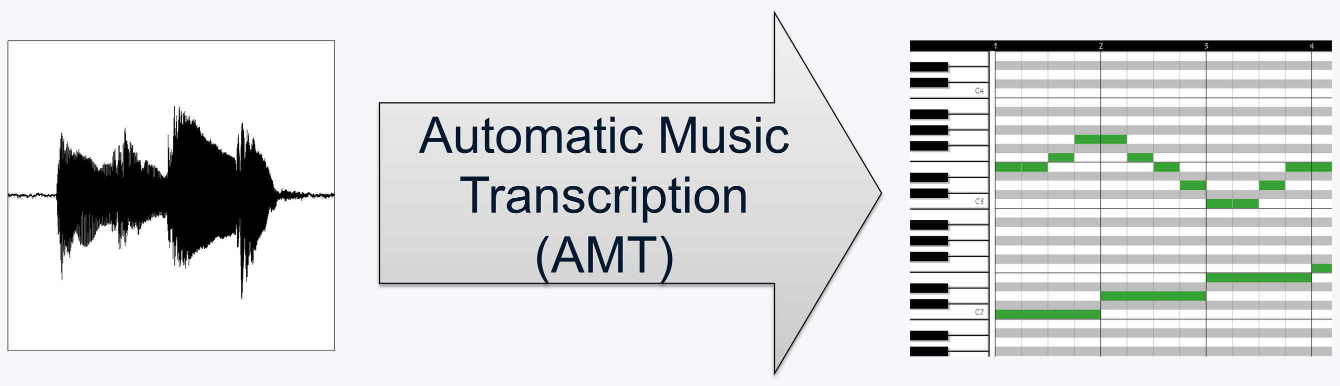
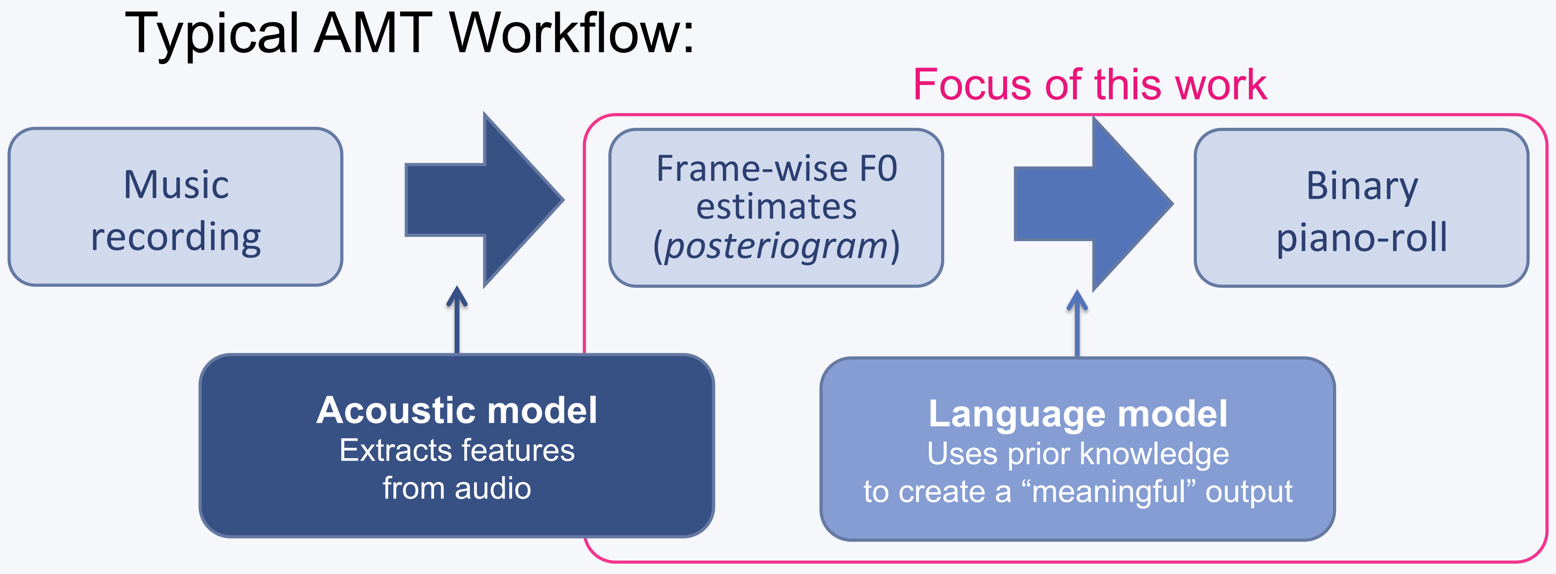
POLYPHONIC MUSIC SEQUENCE TRANSDUCTION WITH METER-CONSTRAINED LSTM NETWORKS

Adrien Ycart, Emmanouil Benetos
Centre for Digital Music, Queen Mary University of London
a.ycart@qmul.ac.uk / emmanouil.benetos@qmul.ac.uk

1. Introduction



Typical AMT Workflow:



2. State of the Art

Complex Neural Networks… … are not so efficient when used inappropriately!

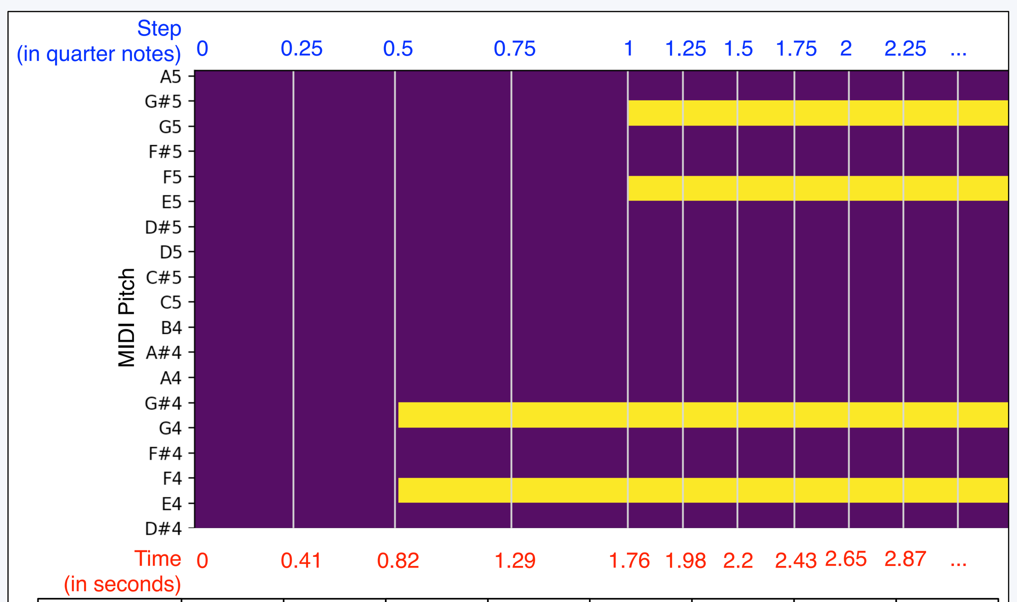
- Boulanger-Lewandowski et al. (2012):
  - RNN-RBM architecture for sequence modelling
  - Time-step: 16th-note
- Sigtia et al. (2015):
  - RNN-RBM integrated with a neural acoustic model
  - Time-step: 32ms

- Kelz et al. (2016):
  - **Outperforms** Sigtia et al. without complex language model
- Korzeniowski & Widmer (2017), Ycart & Benetos (2017):
  - When using a too short time-step, self-transitions predominate → LSTMs only have a **smoothing effect**

Our aim:
Use a simple LSTM network for time-pitch posteriogram post-processing and compare 10ms (time-based) and 16th-note (note-based) time-steps

3. Dataset

- MAPS dataset - Emiya et al. (2010)
  - Aligned MIDI and audio files, played on virtual pianos and on Disklavier
- Rhythm annotations obtained from Piano-midi.de MIDI files
  - Symbolic alignment between Piano-midi.de and MAPS MIDI files
  - Obtain a correspondence table: time position of each 16th-note



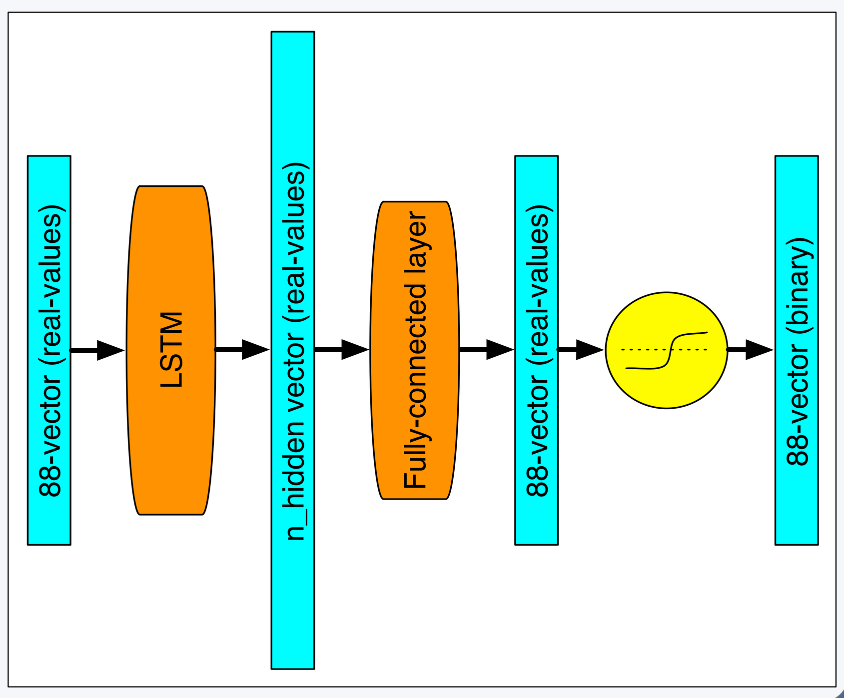
Time	0	0.41	0.82	1.29	1.76	1.98	2.2	2.43
Step	0	0.25	0.5	0.75	1	1.25	1.5	1.75

- **Annotations available at:** http://c4dm.eecs.qmul.ac.uk/ycart/icassp18.html

4. Model

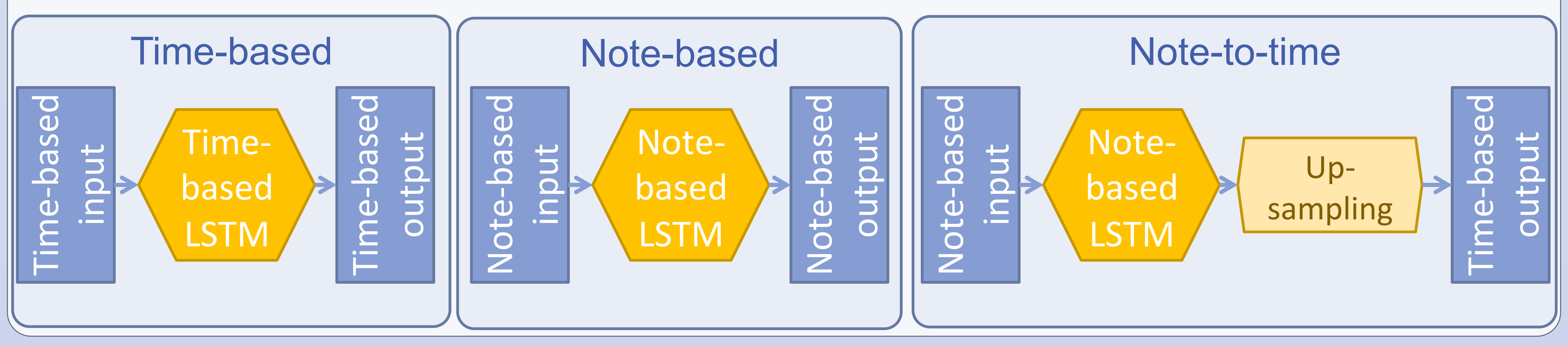
Acoustic Model
- From Benetos and Weyde (2015)
- Based on Probabilistic Latent Component Analysis
- Operates with 10ms time-step: outputs have to be downsampled to 16th note steps

Transduction Model
- 128 hidden nodes, learning rate=0.01
- Adam optimiser on cross-entropy
- Output thresholded (using validation data)



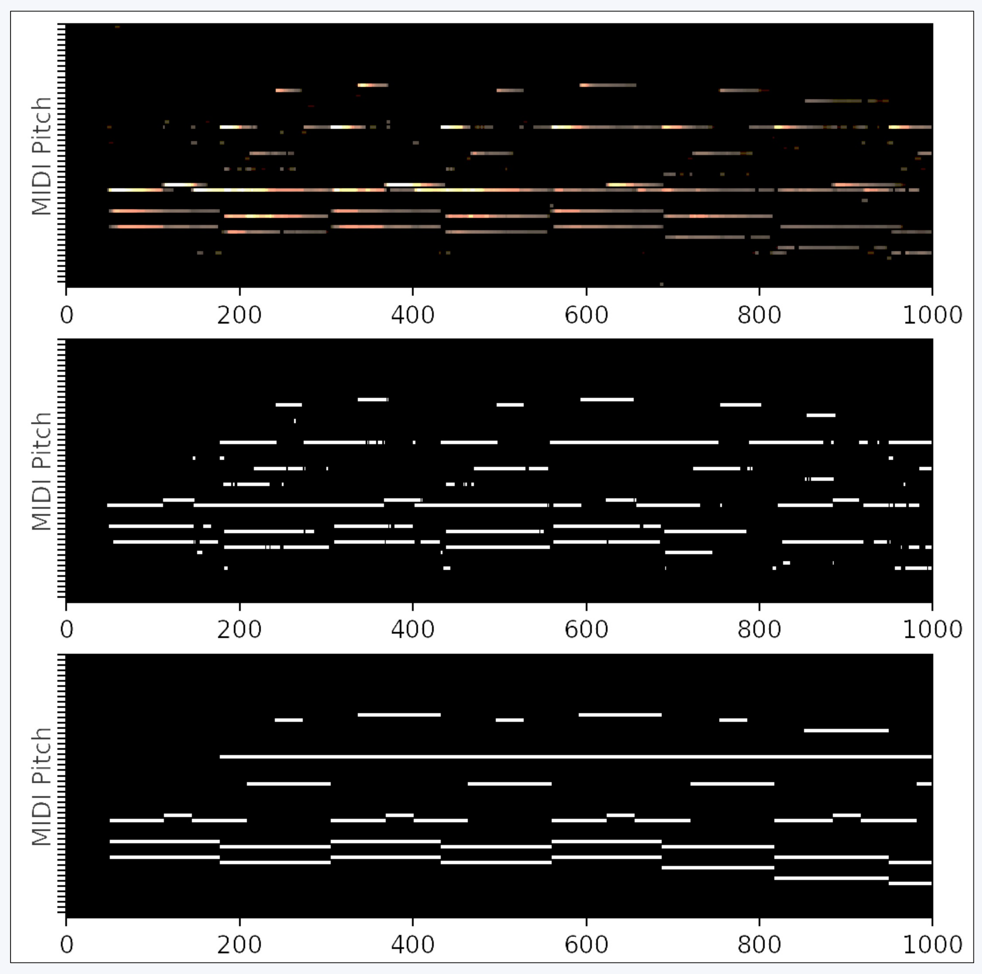
5. Evaluation Metrics

- Two types of metrics
  - Frame metrics: piano-rolls compared frame-by-frame
  - Note metrics: piano-rolls first converted to lists of notes, then compared
  - In both cases we compute: Precision, Recall, F-measure
- Three settings:



6. Experiments

- System compared against:
  - Baseline: median-filtering and thresholding posteriograms
  - HMM: Each pitch is modeled as a 2 state on-off hidden Markov model
- Results:
  - Outperforms both simpler models on frame metrics
  - Outperformed by baseline on note metrics, due to over-fragmentation of notes
  - In every case, better performance in note-to-time setting than in time-based



From top to bottom: posteriogram, LSTM output, ground truth, all in time-based setting

		Time-based setting			Note-based setting			Note-to-time setting		
		$\mathcal{F}(\%)$	$\mathcal{P}(\%)$	$\mathcal{R}(\%)$	$\mathcal{F}(\%)$	$\mathcal{P}(\%)$	$\mathcal{R}(\%)$	$\mathcal{F}(\%)$	$\mathcal{P}(\%)$	$\mathcal{R}(\%)$
Frame metrics	Baseline	63.8	71.0	61.6	69.4	70.5	71.3	65.2	64.8	69.9
	HMM	55.2	**74.1**	48.1	59.5	**76.5**	52.4	56.3	**70.5**	51.4
	LSTM	**66.3**	67.0	**67.8**	**70.2**	70.8	**71.8**	**67.1**	65.9	**71.0**
Note metrics	Baseline	**65.3**	63.2	**70.6**	**72.0**	69.3	**76.5**	**66.3**	66.6	**67.7**
	HMM	61.8	**86.2**	50.9	64.9	**85.9**	54.9	58.5	**81.9**	48.0
	LSTM	57.2	51.1	69.3	65.8	60.5	73.9	62.2	59.6	67.0

7. Discussion

- Two-fold improvement with note-based time steps:
  - Durations are quantised
  - Network better models temporal dependencies

Which one is most important ?

- Compare note-to-time and time-based with quantised durations
  - Equivalent results in both cases: improvement only comes from quantisation
- Downside of note-based time steps:
  - Require beat tracking (rhythm annotations are considered given in this study)
  - Cannot represent extra-metrical notes: trills, ornaments, tuplets…
- Future directions:
  - Replicate experiments with RNN-RBM architecture: a more complex architecture could better model temporal dependencies
  - Use a beat-tracking algorithm instead of ground-truth beat annotations

N. Boulanger-Lewandowski, P. Vincent, and Y. Bengio. "Modeling Temporal Dependencies in High Dimensional Sequences: Application to Polyphonic Music Generation and Transcription." *29th International Conference on Machine Learning*, 2012.

S. Sigtia, E. Benetos, and S. Dixon. "An end-to-end neural network for polyphonic piano music transcription". *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(5):927–939, May 2016.

R. Kelz, M. Dorfer, F. Korzeniowski, S. Bock, A. Arzt, and G. Widmer, "On the Potential of Simple Framewise Approaches to Piano Transcription," 17th *International Conference on Music Information Retrieval (ISMIR*), 2016.

F. Korzeniowski and G. Widmer. "On the Futility of Learning Complex Frame-Level Language Models for Chord Recognition". *In AES International Conference on Semantic Audio*, 2017.

A. Ycart and E. Benetos, "A study on LSTM networks for polyphonic music sequence modelling," in *18th International Society for Music Information Retrieval Conference (ISMIR),* 2017.

E. Benetos and T. Weyde. "An efficient temporally constrained probabilistic model for multiple instrument music transcription". *In 16th International Society for Music Information Retrieval Conference (ISMIR),* 2015.

Queen Mary University of London

centre for digital music