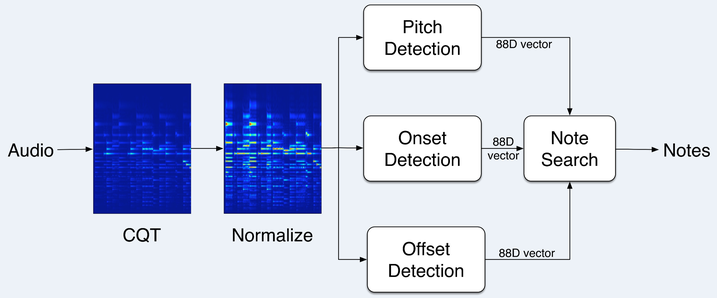
A PARALLEL FUSION APPROACH TO PIANO MUSIC TRANSCRIPTION BASED ON CONVOLUTIONAL NEURAL NETWORK

Fu'ze Cong*, Shuchang Liu*, Li Guo*, and Geraint A. Wiggins#,†
*Beijing University of Posts and Telecommunications
Key Lab of Universal Wireless Communications, Ministry of Education, Beijing, China
School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
† AI Lab, Department of Computer Science, Free University of Brussels, Belgium

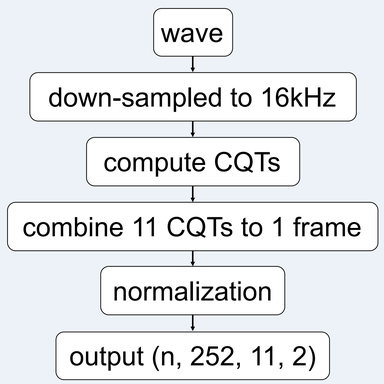
Overview

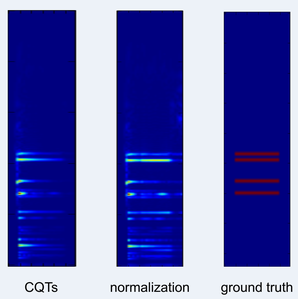
Automatic music transcription (AMT) is the process of generating some form of notation-like musical score from a given acoustic musical signal. In order to improve the performance of a frame-based AMT system in note-based output, we attempt to integrate onset/offset detection model into the current AMT system.



The extracted features are fed to pitch, onset and offset detection models. The output about pitch, onset and offset are integrated by note search model to determine the final note events which are the results of transcription.

Preprocessing



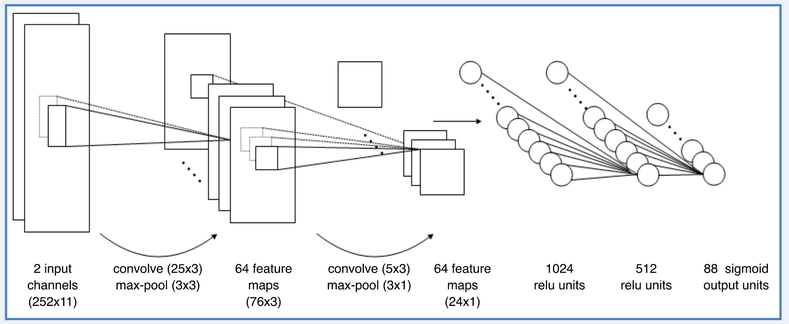


Constant Q transform (CQT)

- 7 octaves with 36 bins per octave, therefore:
  $f_{min} = 32.7Hz\ (C1) \quad f_{max} = 4185.6Hz\ (C8)$
- Length of the input is 1024 and hop size is 512.

■ **To make the offset of each note more clear, we normalize the CQTs per frame (11 CQTS).**

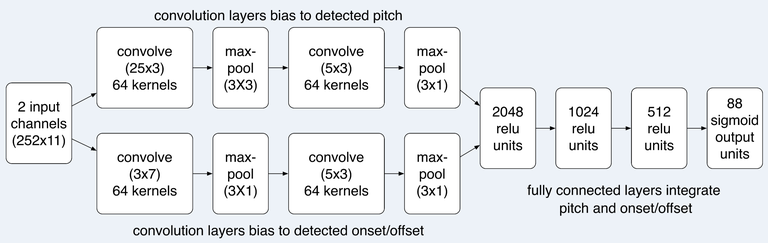
Pitch Detection Model



- The input of the model is (252, 11, 2) tensors standing for 11 frames of 2 channel piano audio.
- The output of the model is an 88-dimensional vector corresponding to the probabilities of pitch in notes A0-C8 on a piano.
- We choose the kernels with shapes 25x5 and 5x3, which have been shown perform better than others.

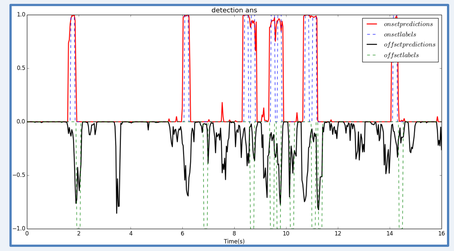
Onset/Offset Detection Model

Compared with other deep learning models, CNNs are good at edge detection, because convolution is an effective way of describing changes by applying the same linear transformation of a small local region across the entire input.

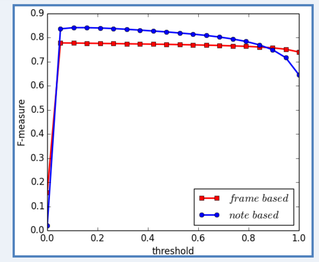


- The kernels with shape 25x3 are designed to estimate pitch in each frame.
- The kernels with shape 3x7 are more sensitive to the changes in time domain which are designed to detect onset and offset.
- The fully connected layers integrate pitch and onset/offset information.

Note Search Model



Outputs of onset/offset detection model



Frame- and note-based F-measures of note search with different thresholds (δ) of onset/offset searching.

- The outputs of onset/offset model are filtered by a threshold, the middle of successive positive frames is viewed as the time point of onset/offset event.
- If there is no offset event between two onset events, more than 3 successive frames with low pitch probability (less than 0.1) are viewed as the end of the note.

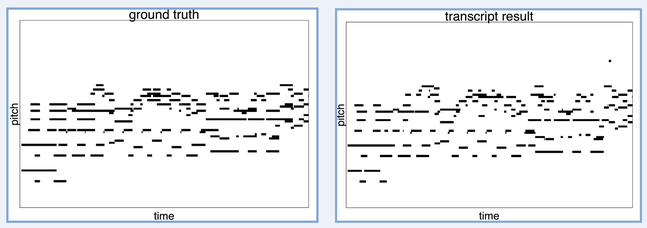
Results and Conclusion

MAPS (MIDI Aligned Piano Sounds)

- Mode I: randomly split, 216 pieces for training, 54 for test.
- Mode II: synthesized music (210 pieces) for training, real piano music (60 pieces) for test.

	Mode I		*Mode II*	
	frame	Note	Frame	Note
Vincent	59.78	69.00	59.60	59.12
Sigtia	74.45	67.05	64.14	54.89
Ours	**77.76**	**84.16**	**65.02**	**68.23**

■ **The note-based results of the system proposed is much better, which indicates our CNN based onset/offset model can improve the note accuracy of AMT system.**



Transcription result and ground truth for the first 30 seconds of track *MAPS_MUS-alb_esp2_AkPnCGdD.*

References

S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 5, pp. 927– 939, May 2016.

Acknowledgements

This work is supported by the 111 Project of China (B16006).