Feature Adapted Convolutional Neural Networks for Downbeat Tracking

Simon DURAND[1], Juan Pablo BELLO[2], Bertrand DAVID[1], Gaël RICHARD[1]
[1]LTCI, CNRS, Télécom Paristech, Université Paris-Saclay, France
[2]Music and Audio Research Laboratory - New York University, USA
simon.durand@telecom-paristech.fr

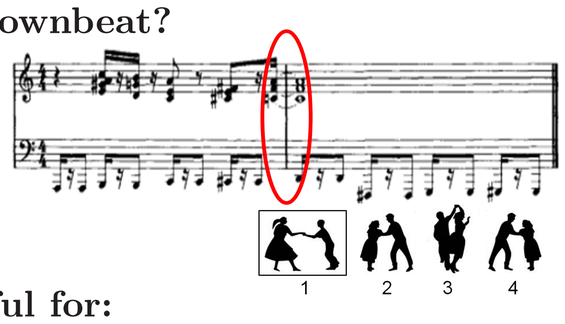
Introduction

What is our aim?

- **Recover downbeat time instants from music audio signals.**

What is a downbeat?

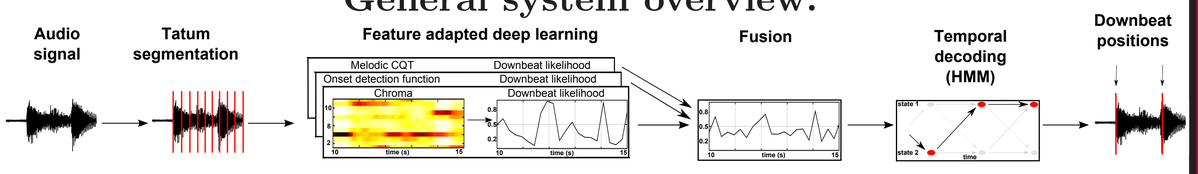
- Bar boundaries.
- First beat of a bar.



It is useful for:

- Automatic sheet-music transcription.
- Genre, chord or structure recognition.

General system overview:

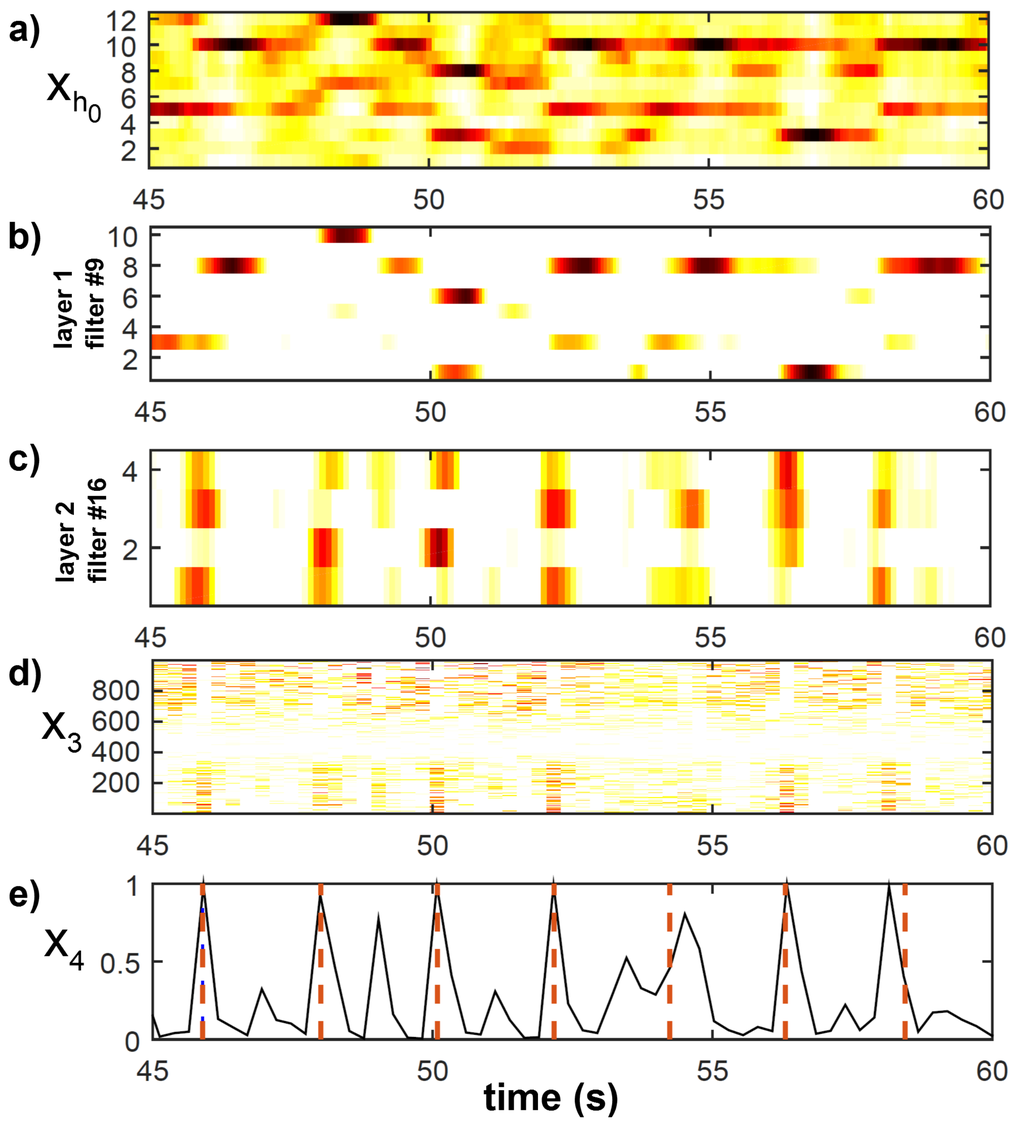


Focus of this work:

- To design **adapted convolutional neural network** (CNN) architecture to each feature characteristic.

1) Harmonic Network (HCNN)

- Highlight instantaneous harmonic change around downbeats.
→ Use small filter receptive fields and input temporal dimension.
- A song transposition shouldn't change our downbeat perception.
→ Implement circular shifting data augmentation.
- Visualization of the harmonic network:

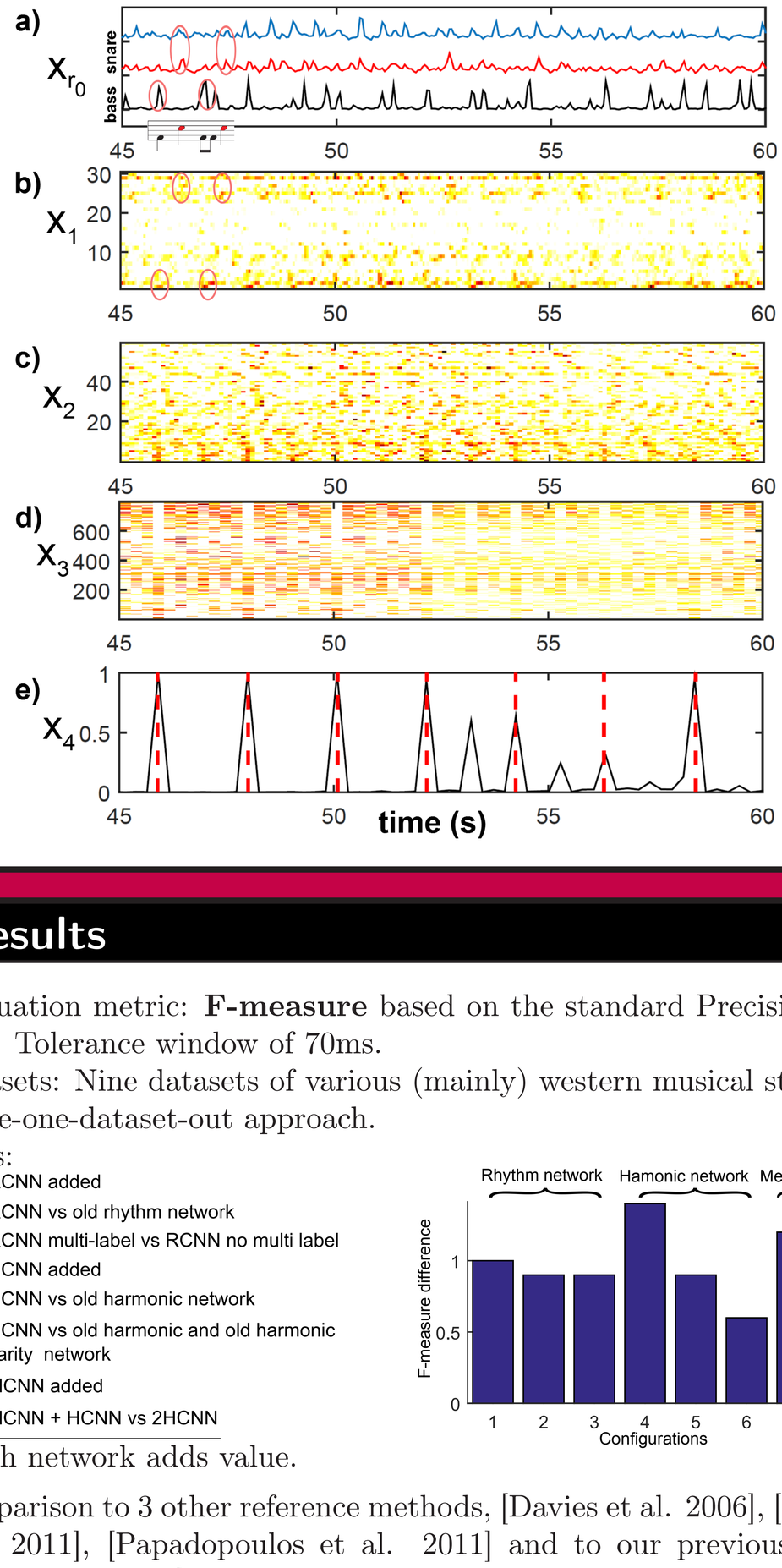


2) Melodic network (MCNN)

- Melody contour plays a role in perceiving rhythm hierarchies, but it is difficult to derive high level heuristics.
→ Design a low-level representation of melodic contour based on the constant-Q transform and a salience function
→ Use large filter receptive fields to find a melodic pattern as a first layer.
- Melody contour is pitch invariant.
→ Perform max pooling on the whole frequency range of this layer output to keep the most salient melodic pattern.

3) Rhythmic network (RCNN)

- Highlight bar-long pattern.
→ Use large filter receptive fields and input temporal dimension.
- Can encode the length of the bar.
→ Output different labels for different bar length and downbeat positions.
- Visualization of the rhythmic network:



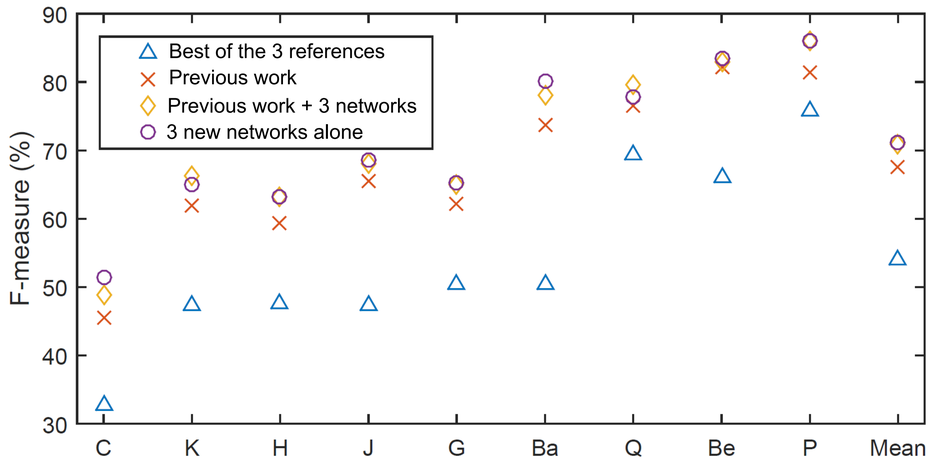
4) Results

- Evaluation metric: **F-measure** based on the standard Precision and Recall. Tolerance window of 70ms.
- Datasets: Nine datasets of various (mainly) western musical styles.
- Leave-one-dataset-out approach.
- Tests:

Configurations:
(1) RCNN added
(2) RCNN vs old rhythm network
(3) RCNN multi-label vs RCNN no multi label
(4) HCNN added
(5) HCNN vs old harmonic network
(6) HCNN vs old harmonic and old harmonic similarity network
(7) MCNN added
(8) MCNN + HCNN vs 2HCNN

→ Each network adds value.

- Comparison to 3 other reference methods, [Davies et al. 2006], [Peeters et al. 2011], [Papadopoulos et al. 2011] and to our previous work, [Durand et al. 2015].



Main ideas and conclusion

- **Use melody, rhythm and harmony to characterize downbeats.**
- **Take advantage of the high level and continuous aspect of downbeats with convolutional neural networks.**
- **Adapt the network architecture to each feature.**
- **Significantly outperforms the previous state of the art.**