# Main Melody Extraction with Source-Filter NMF and CRNN

Dogac Basaran, Slim Essid, Geoffroy Peeters
dogac.basaran@ircam.fr, slim.essid,geoffroy.peeters@telecom-paristech.fr

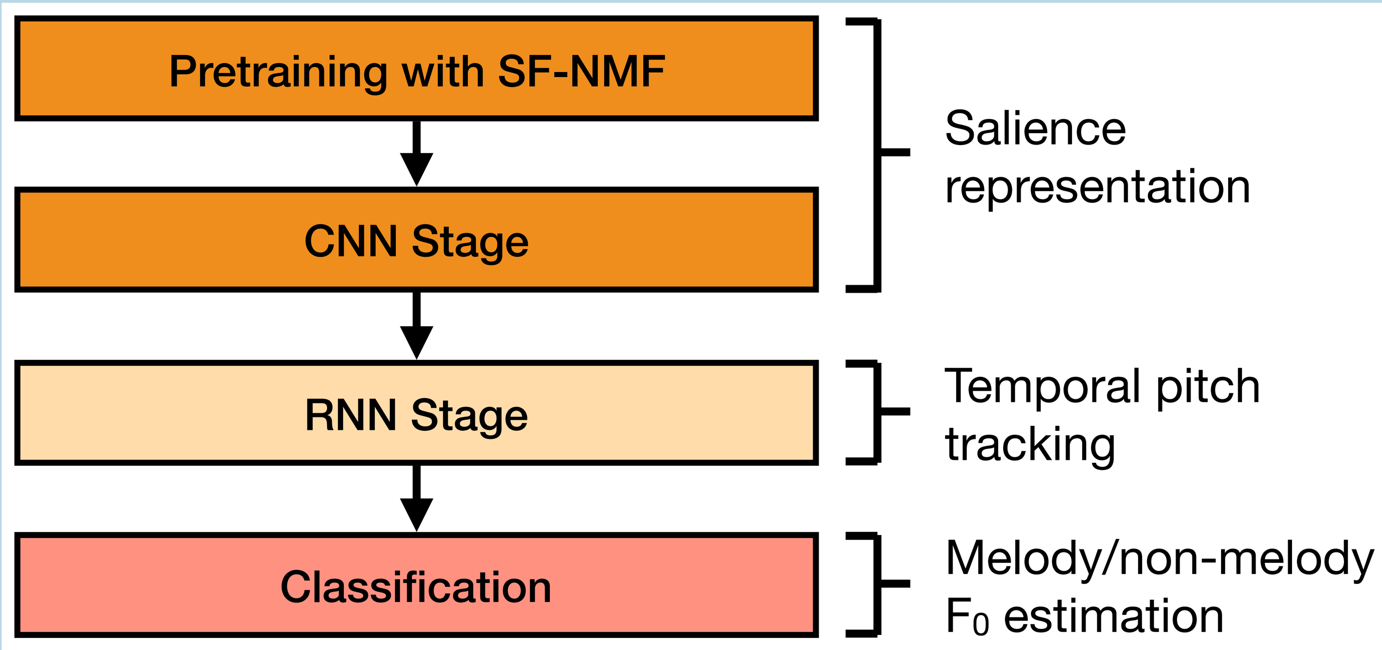
## Introduction

We propose a Convolutional-Recurrent Neural Network (CRNN) model whose pretraining is based on the SF-NMF model [1].

**Contributions:**

- State-of-the-art performance achieved without large training datasets or data augmentation.
- Results on MedleyDB demonstrate the usefulness of a good input salience representation to the network.

## System Overview




- Initial saliency $\mathbf{H}^{F_0}$ + Deep Convolutive network for enhanced salience representation

- RNN: **Bidirectional GRU**

- 62 classes: 1 non-melody class, 61 target $F^0$ classes in semitone resolution

## Pretraining with SF-NMF

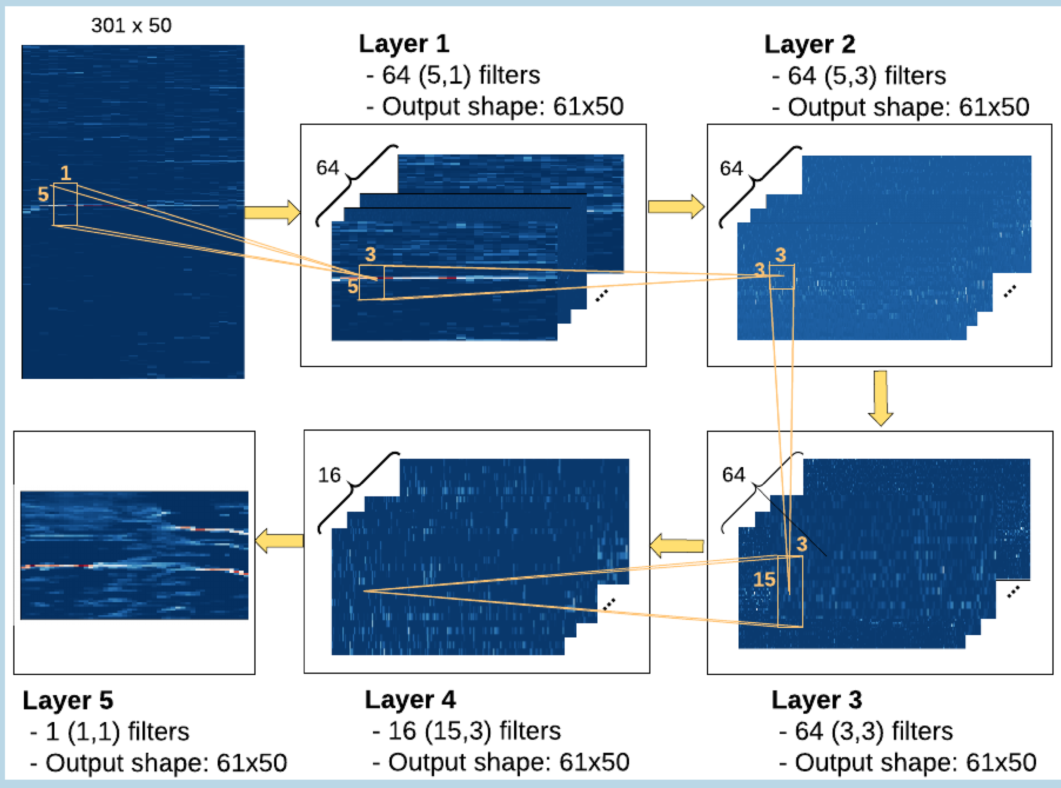
Source Filter - Nonnegative Matrix Factorization (SF-NMF) model:

$$\begin{aligned}\mathbf{V} \approx \hat{\mathbf{V}} &= \mathbf{V}^{F_0} \odot \mathbf{V}^{\Phi} + \mathbf{V}^{B} \\ &= \underline{\mathbf{W}}^{F_0}\mathbf{H}^{F_0} \odot \mathbf{W}^{\Phi}\mathbf{H}^{\Phi} + \mathbf{W}^{B}\mathbf{H}^{B} \\ &= \underline{\mathbf{W}}^{F_0}\mathbf{H}^{F_0} \odot \underline{\mathbf{W}}^{\Gamma}\mathbf{H}^{\Gamma}\mathbf{H}^{\Phi} + \mathbf{W}^{B}\mathbf{H}^{B}\end{aligned}$$

$\mathbf{W}^{F_0}$: Preconstructed basis, each column represents the harmonic structure of an $F^0$

$\mathbf{H}^{F_0}$: Each row represents the activation of an $F^0$ → A Salience Representation

## CNN Architecture 1 (CNN1)




- **Input resolution**: 5 $F^0$s per semitone, from A1 (55Hz) to A6 (1760Hz) → 301 features per frame

- **Layer 1**: Focuses the energy around semitones on top of them (conv. with strides of 5), decrease frequency resolution to semitone.

- **Layers 2** & **3**: For learning to overcome one-tone and one-semitone confusion errors.

- **Layer 4**: For learning octave error patterns.

## Experimental Setup

- Evaluation on MedleyDB: *Melody 2 definition.*
- Models trained on 67 tracks of MedleyDB

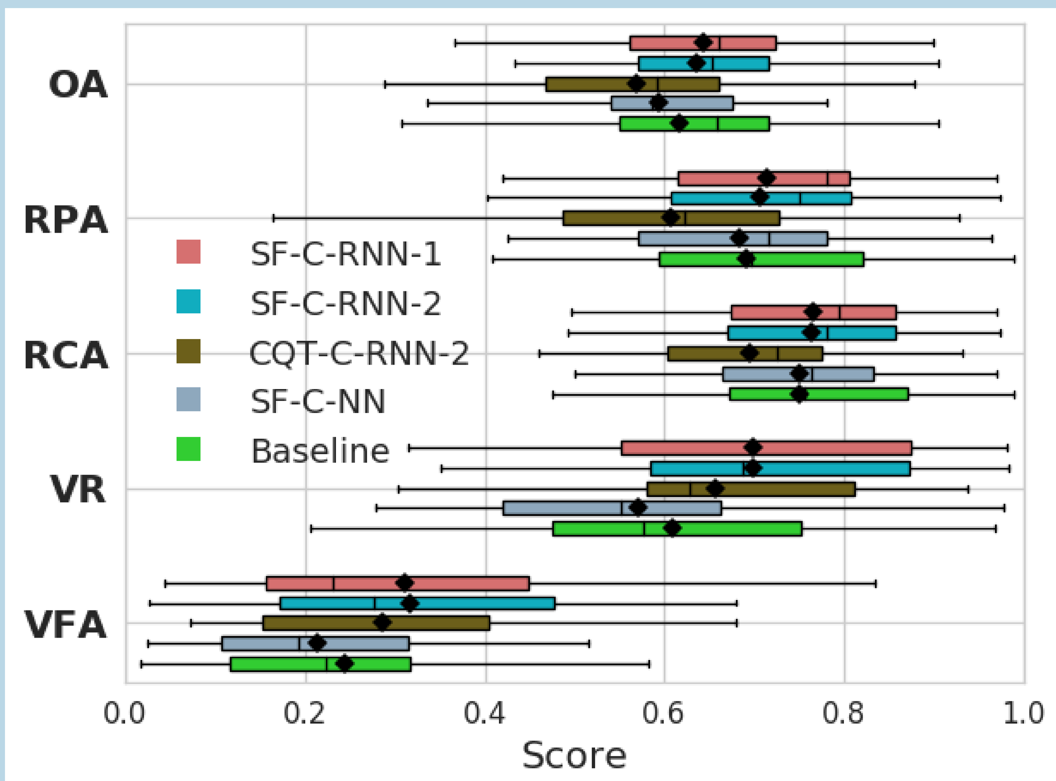
*Metrics*:
Overall Accuracy (OA), Raw Pitch Accuracy (RPA), Raw Chroma Accuracy (RCA), Voicing Recall (VR), Voicing False Alarm (VFA)

*Network variants*:

**SF-CRNN-1**: CNN1 + 1 layer BiGRU (128 Units) + Classification layer (307,199 params)
**SF-CRNN-2**: CNN2 + 1 layer BiGRU (160 Units) + Classification layer (854,319 params)
**CQT-CRNN-2**: same as above but with CQT input.
**SF-CNN**: CNN2 + Classification layer
**Baseline**: CNN2 with Harmonic-CQT input (406,253 parameters) [2]

- No MaxPooling or Dropout
- With early stopping
- **CNN**: trained on 0.29-sec (25-frame) patches
- **RNN**: trained on 5.8-sec (500-frame) patches

## Experimental Results




- SF-CRNN1 beats the baseline on OA, RPA, RCA and VR with 1/3 amount of training data and less parameters.
- A better initial saliency representation results in better performance (SF-CRNN2 vs. CQT-CRNN2).
- Temporal tracking with RNN significantly improves the performance of the system (SF-CNN vs. SF-CRNN2).
- CNN1 (low resolution) performs better than CNN2 (high resolution) (SF-CRNN1 vs. SF-CRNN2)

### $\mathbf{H}^{F_0}$ vs. CQT as salience

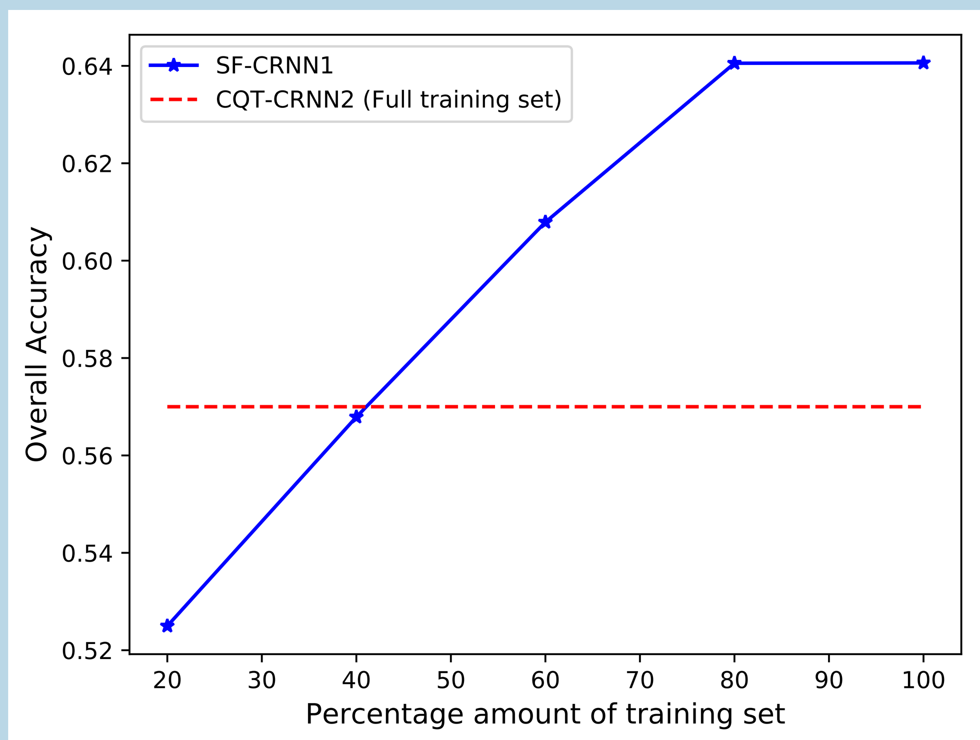
- $\mathbf{H}^{F_0}$ has much higher RPA and RCA than CQT
- $\mathbf{H}^{F_0}$ is better initial salience representation.

| | $\mathbf{H}^{F_0}$ | CQT |
|---|---|---|
| RPA | $0.538 \pm 0.141$ | $0.210 \pm 0.16$ |
| RCA | $0.648 \pm 0.127$ | $0.411 \pm 0.15$ |

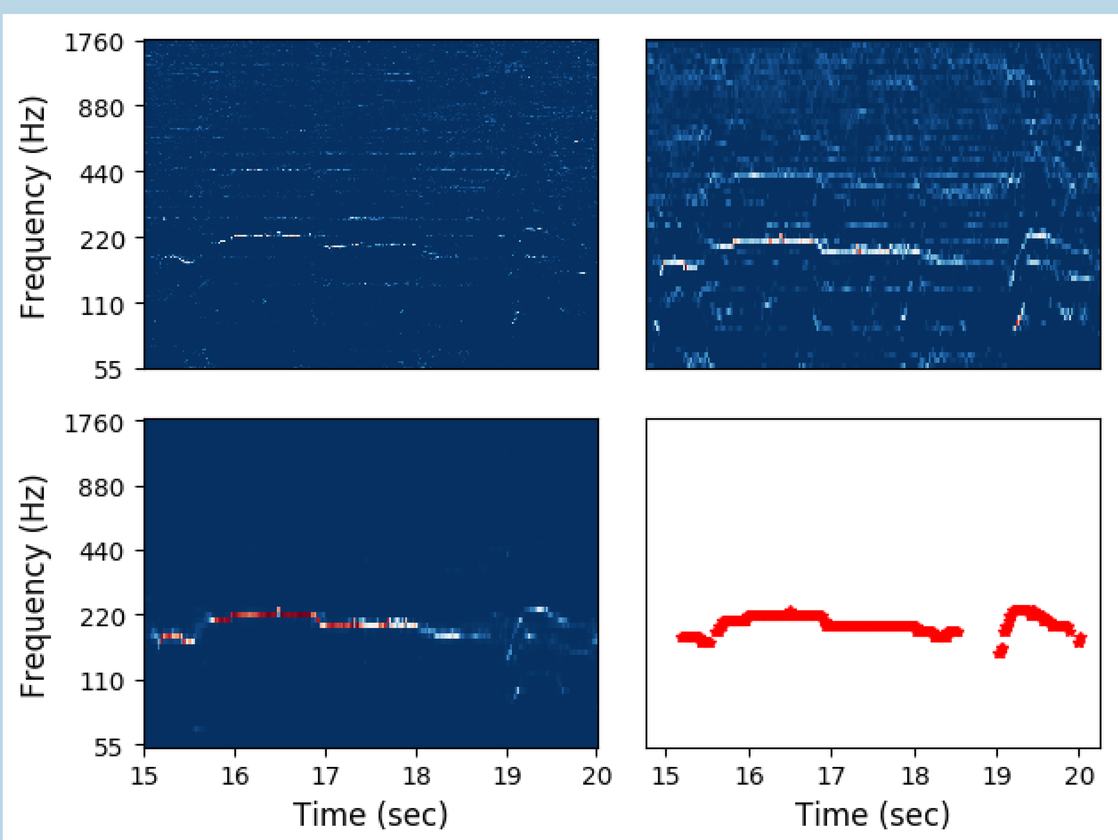
### Singing voice vs. instrument

| | SF-CRNN-1 | | Baseline | |
|---|---|---|---|---|
| | S.V. | Ins. | S.V. | Ins. |
| OA | 0.638 | 0.466 | 0.598 | 0.424 |
| RPA | 0.791 | 0.647 | 0.784 | 0.619 |
| RCA | 0.804 | 0.726 | 0.823 | 0.717 |

40% of training data is enough for SF-CRNN1 to reach OA of CQT-CRNN2 with full training data.




## CRNN Activations




**(Top-left)** $\mathbf{H}^{F_0}$ as input to CRNN
**(Top-right)** CNN1 activations
**(Bottom-left)** Classifier activations
**(Bottom right)** Ground-truth annotations.

## Conclusion and Future Work

- Pretraining stage has proven very effective.
- Proposed system achieves the state-of-the-art with lower complexity and less training data.

- Future goals: to jointly train SF-NMF and CRNN models, and to improve $\mathbf{H}^{F_0}$ for more discriminative salience representation.

## References

[1] J. L. Durrieu, B. David, and G. Richard. A musically motivated mid-level representation for pitch estimation and musical audio source separation. IEEE Journal of Selected Topics in Signal Processing, 2011.

[2] R.M. Bittner, B. McFee, J. Salamon, P. Li, and J.P. Bello. Deep salience representations for f0 estimation in polyphonic music. In 18th International Society for Music Information Retrieval Conference, ISMIR, 2017.

## Acknowledgements

This research was partially supported by the *DigThatLick* project (http://dig-that-lick.eecs.qmul.ac.uk)