NYU

CREPE:
A Convolutional Representation for Pitch Estimation

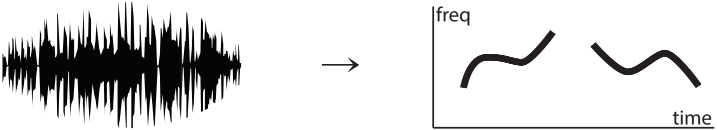
ICASSP 2018 Lecture Session AASP-L4.3: Music Signal Analysis and Processing

April 19, 2018

Jong Wook Kim, Justin Salamon, Peter Li, Juan Pablo Bello

Music and Audio Research Laboratory, New York University

Task: Monophonic Pitch Estimation



- A long-standing topic in audio signal processing
- A fundamental problem in music information retrieval
- With many aplications
  - In a melody extraction system[Bosch and Gómez 2014]
  - Annotating multi-track datasets[Salamon et al. 2017]
  - Analyzing intonations in speech analysis

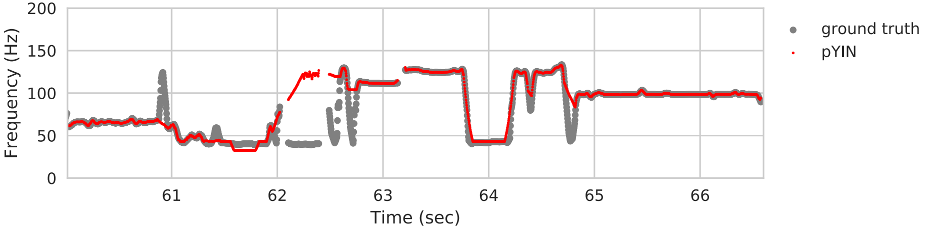
Background on Monophonic Pitch Estimation

A History of Engineering Heuristic Feature Extractor Functions

- Frequency-domain methods
  - Cepstrum[Noll 1967], SWIPE[Camacho and Harris 2008]
- Time-domain methods
  - $f_{ACF}(\tau) = \sum x_t x_{t+\tau}$, $f_{AMDF}(\tau) = \sum |x_t - x_{t_\tau}|$, $f_{ASDF}(\tau) = \sum (x_t - x_{t-\tau})^2$
  - YIN[De Cheveigné and Kawahara 2002]: cumulative mean normalized difference function
  - pYIN[Mauch and Dixon 2014]: a probabilistic extension to YIN – **the state of the art**
- These are *all* based on hand-crafted features and heuristics

Motivation

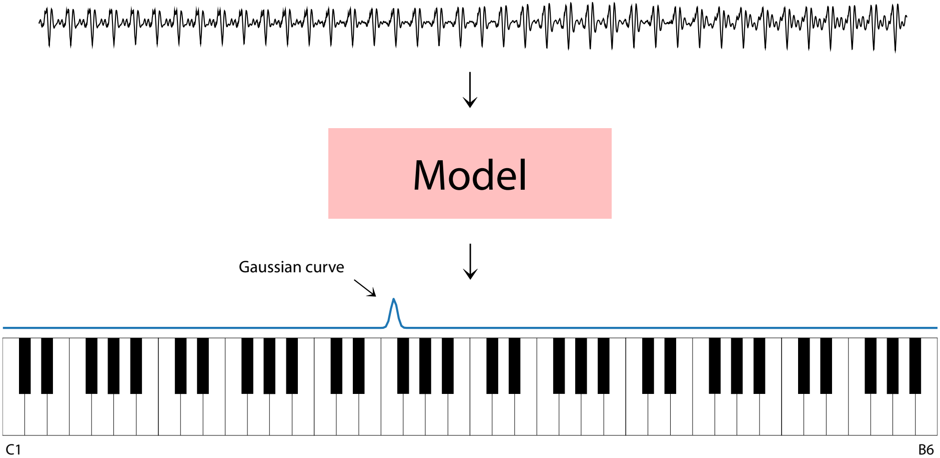
- Reported near-perfect accuracies are based on simplistic datasets
- We encountered many cases where the SoTA doesn't do well: original pYIN



- Should benefit from **data-driven methods**, just like many other MIR tasks

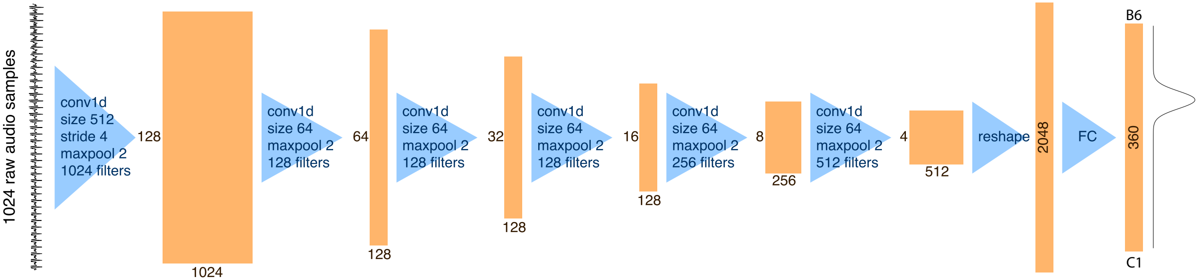
Problem Formulation

1024-sample segment from 16,000 Hz audio



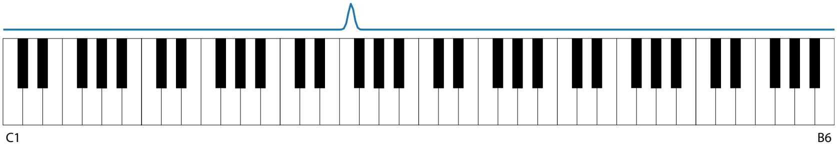
360-D activation over 6 octaves, 20-cents-wide bins

Deep Model Architecture



Post Processing and Optimization

- The model produces a 360-D activation vector for each input frame: Bittner et al. 2017



- Estimated pitch is then given as the (local) weighted average of the weights

$$\hat{\text{¢}} = \sum_{i=1}^{360} \hat{y}_i \text{¢}_i \Bigg/ \sum_{i=1}^{360} \hat{y}_i$$

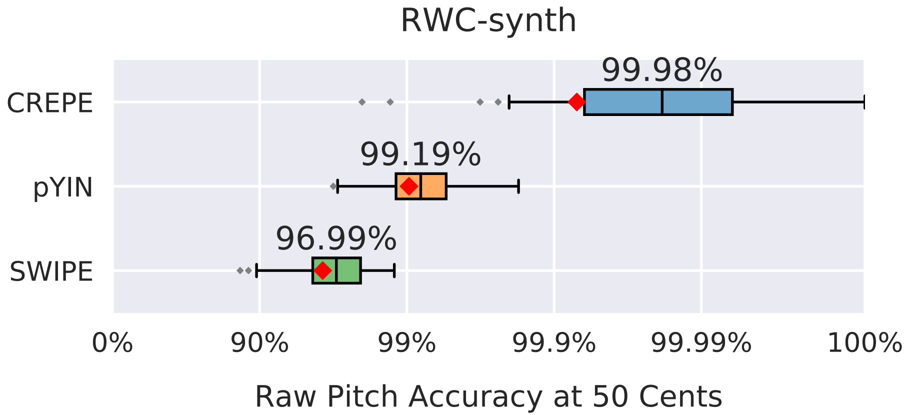
- Optimization target: minimize the binary cross entropy:

$$\mathscr{L}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{i=1}^{360} \Big(-y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i) \Big)$$

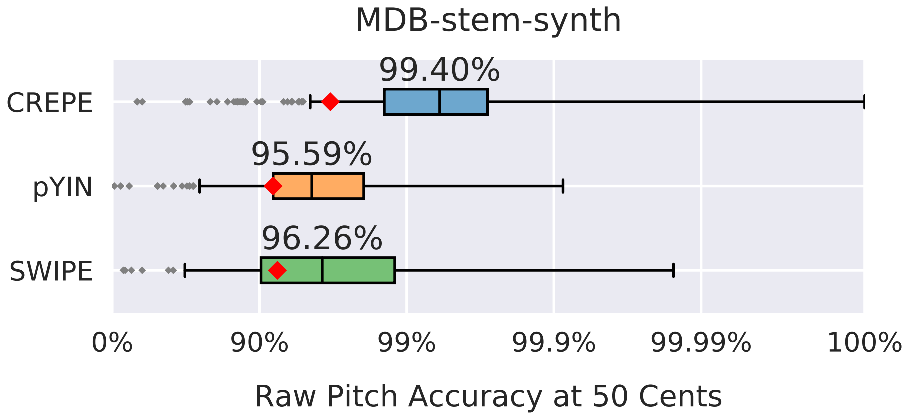
Datasets and Evaluation

- For objective evaluation, we need a dataset with perfect pitch annotations
- The datasets:
  - **RWC-synth**[Mauch and Dixon 2014]: 6.16h, one timbre, on which pYIN was evaluated
  - **MDB-stem-synth**[Salamon et al. 2017]: 15.36h, 25 instruments from MedleyDB[Bittner et al. 2014]
  - Listen: RWC-synth 1 | RWC-synth 2 | MDB-stem-synth 1 | MDB-stem-synth 2
- 5-fold cross validation and artist-conditional splits
- Reporting pitch accuracies using `mir_eval`[Raffel et al. 2014]:
  - Raw Pitch Accuracy (RPA)
  - Raw Chroma Accuracy (RCA)

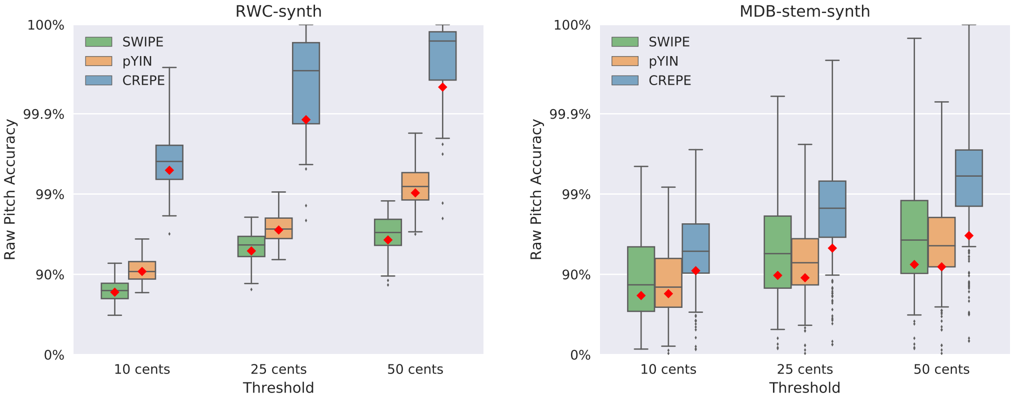
Results: Pitch and Chroma Accuracy on RWC-synth



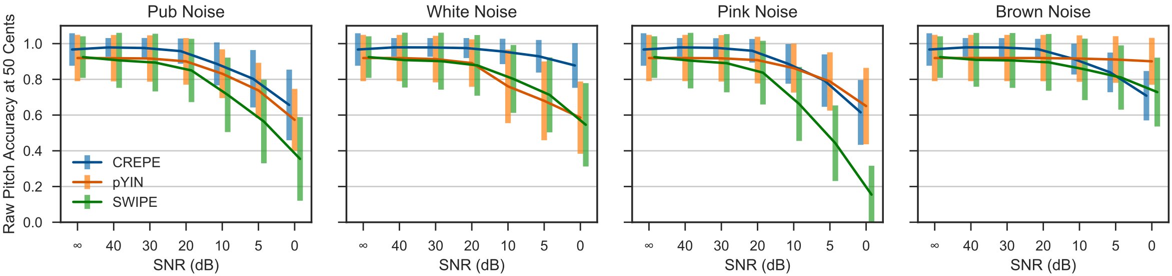
Results: Pitch and Chroma Accuracy on MDB-stem-synth



Results: Thresholds

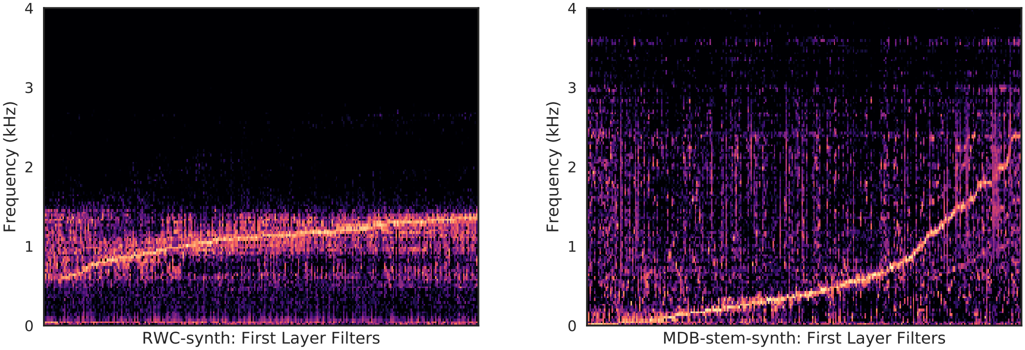


Results: Noise Robustness



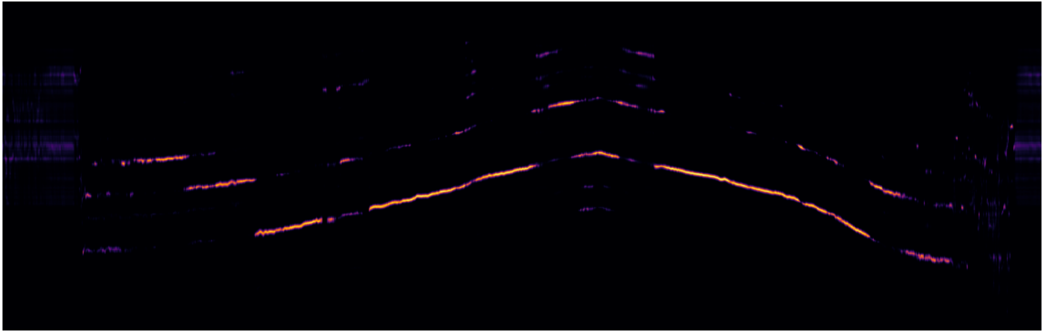
Results: First-Layer Filters

- The filters **adapt** to the timbre distribution of the dataset

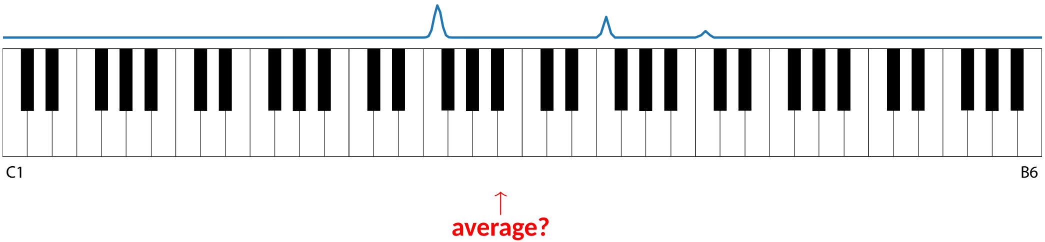


The Generalization Problem

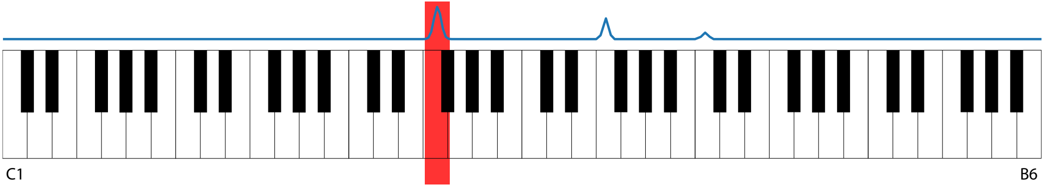
- When the model trained on MDB-stem-synth is tested on my voice:



Fix 1: Argmax-Local Averaging



Fix 1: Argmax-Local Averaging

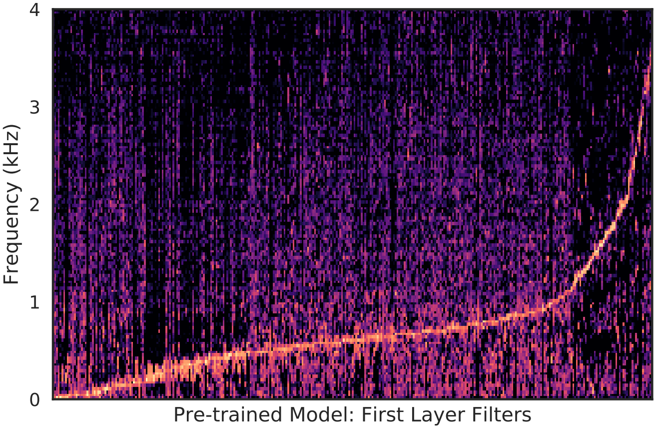


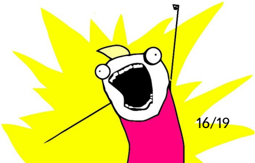
- Local weighted average around the highest activation:

$$\hat{¢} = \sum_{i=m-4}^{m+4} \hat{y}_i ¢_i \Big/ \sum_{i=m-4}^{m+4} \hat{y}_i, \qquad m = \mathrm{argmax}_i\, \hat{y}_i$$

Fix 2: Train with ALL THE DATA!

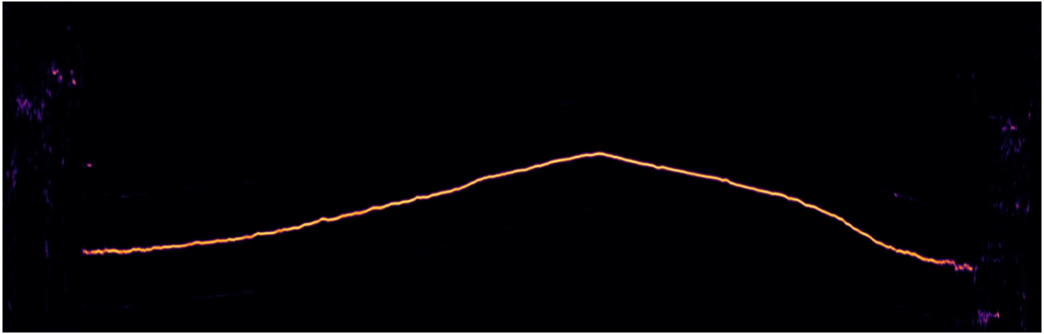
- MIR-1K, Bach10, RWC-synth, MedleyDB, MDB-stem-synth, NSynth





The Pre-trained Model Release

- Fixed the generalization problem:



- The highest activation is a good heuristic for voice activity detection (VAD)
- An interactive demo: `https://marl.github.io/crepe/`

Summary

- Presented a data-driven neural network model as a state of the art method
  - Runs directly on time-domain audio signal
  - Robust with heterogeneous timbre and additive noise
  - Stays highly accurate, even with 10 cents threshold

- Try it today!

```
$ pip install tensorflow   # or tensorflow-gpu
$ pip install crepe        # install the CREPE package
$ crepe audio.wav          # run pitch estimation on audio.wav
```

References

Bittner, R. M. et al. (2014). “MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research.” In: *Proceedings of the 15th ISMIR Conference*. Vol. 14, pp. 155–160.

Bittner, R. M. et al. (2017). “Deep Salience Representations for F0 tracking in Polyphonic Music”. In: *Proceedings of the 18th ISMIR Conference*.

Bosch, J. and E. Gómez (2014). “Melody Extraction in Symphonic Classical Music: a Comparative Study of Mutual Agreement Between Humans and Algorithms”. In: *Proceedings of the 9th Conference on Interdisciplinary Musicology (CIM14)*.

Camacho, A. and J. G. Harris (2008). “A sawtooth waveform inspired pitch estimator for speech and music”. In: *The Journal of the Acoustical Society of America* 124.3, pp. 1638–1652.

De Cheveigné, A. and H. Kawahara (2002). “YIN, a fundamental frequency estimator for speech and music”. In: *The Journal of the Acoustical Society of America* 111.4, pp. 1917–1930.

Mauch, M. and S. Dixon (2014). “pYIN: A fundamental frequency estimator using probabilistic threshold distributions”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, pp. 659–663.

Mauch, M. and S. Ewert (2013). “The Audio Degradation Toolbox and its Application to Robustness Evaluation”. In: *Proceedings of the 14th ISMIR Conference*. accepted. Curitiba, Brazil.

Noll, A. M. (1967). “Cepstrum pitch determination”. In: *The journal of the acoustical society of America* 41.2, pp. 293–309.

Raffel, C. et al. (2014). “mir_eval: A transparent implementation of common MIR metrics”. In: *Proceedings of the 15th ISMIR Conference*.

Salamon, J. et al. (2017). “An Analysis/Synthesis Framework for Automatic F0 Annotation of Multitrack Datasets”. In: *Proceedings of the 18th International Society for Music Information Retrieval (ISMIR) Conference*.