Multiple F0 Estimation in Vocal Ensembles using Convolutional Neural Networks

Helena Cuesta[1], Brian McFee[2], Emilia Gómez[3,1]

[1] Music Technology Group, Universitat Pompeu Fabra (Barcelona)
[2] Music and Audio Research Lab & Center for Data Science, New York University (New York)
[3] Joint Research Centre, European Commission (Sevilla)

Listen to the results!

We present and evaluate a set of **CNNs** for **multiple F0 estimation in vocal quartets**. We use the **magnitude** and **phase differentials** of the **HCQT** as input to the networks and build upon an existing system to produce a **pitch salience** representation of the input signal. We construct a dataset that comprises several **multi-track polyphonic singing** datasets for training and evaluation. Our model can be used with polyphonic recordings in the wild and outperforms two baseline methods on the same data.

1 Motivation

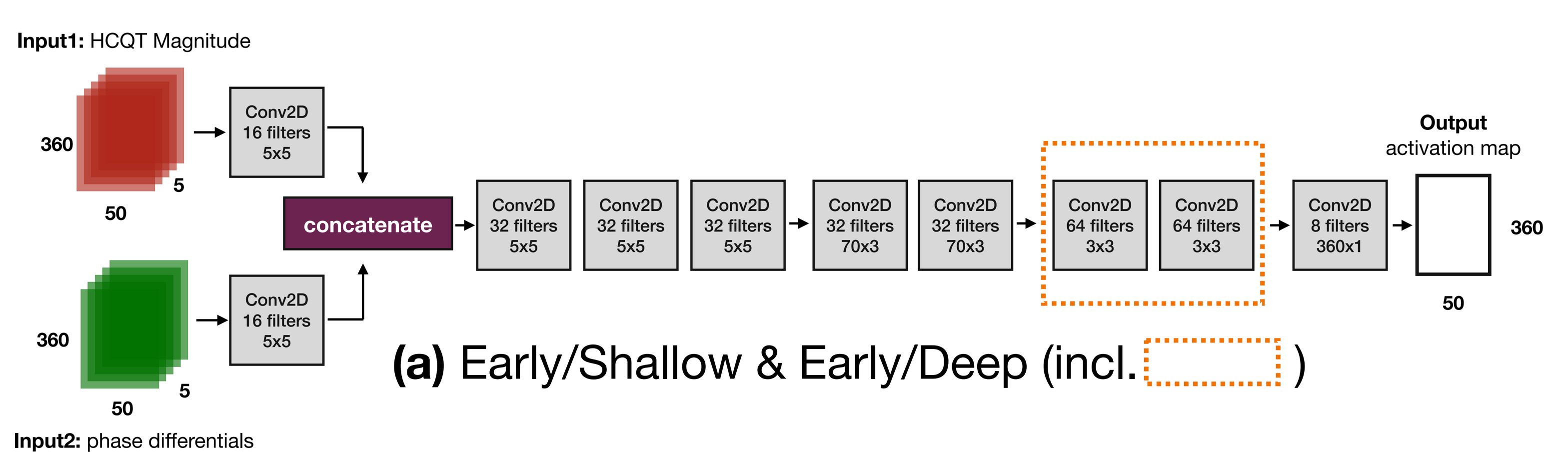
- Analysis of **ensemble singing** commonly requires individual recordings of each voice and/or individual F0 curves.
- Intonation analysis, source separation, and automatic transcription benefit from multi-F0 estimation.
- We can obtain individual F0 curves from mixed, polyphonic recordings of vocal quartets.
- Work based on DeepSalience [1].

3 Data collection

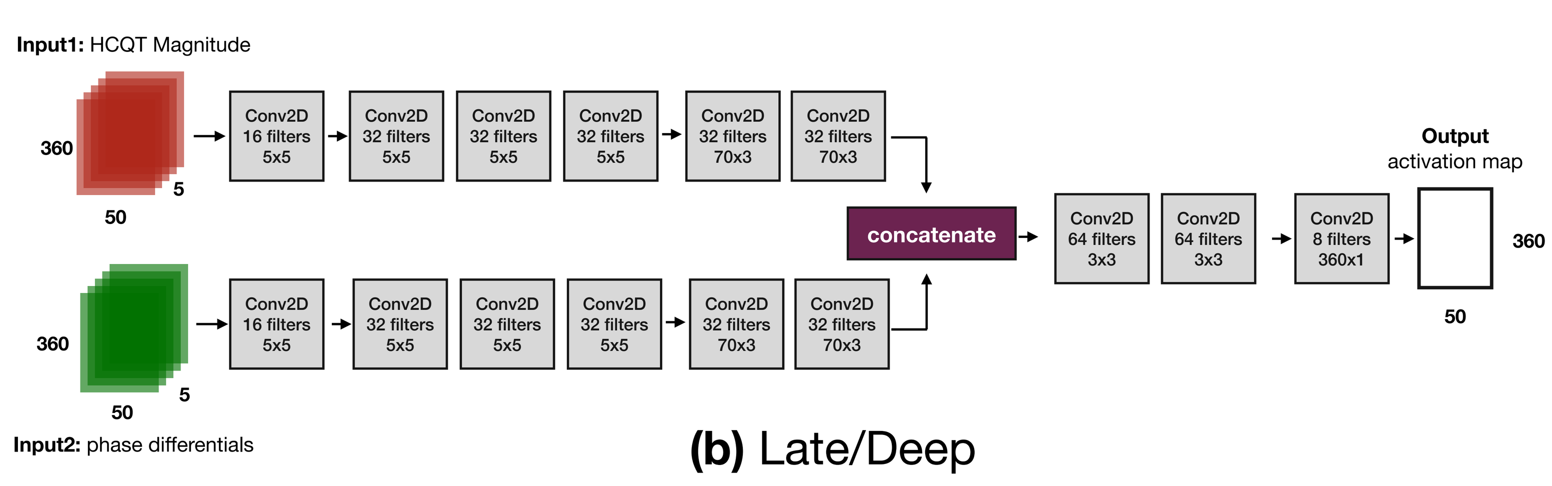
Dataset	Availability	Configuration	Duration (mm:ss)
Choral Singing Dataset [2]	Public	16 singers, SATB	07:14
Dagstuhl ChoirSet [3]		13 singers, SATB	55:30
ESMUC Choir Dataset	Private	13 singers, SATB	21:08
Barbershop Quartets		4 singers, LTBB	42:10
Bach Chorales		4 singers, SATB	58:20

- Compile 5 multi-track datasets of polyphonic vocal music.
- Public + proprietary datasets.
- F0 annotations for each voice in the ensemble.
- Combine voices to create all possible SATB quartets (intra-dataset).

5 CNN architectures



(a) Early/Shallow & Early/Deep (incl. [dashed box])



(b) Late/Deep

7 Conclusions & limitations

- Late concatenation of magnitude & phase information works better than early concatenation.
- Our models look robust to **increased pitch resolution** (100 vs. 20 cents).
- We need **further experiments on unisons and commercial recordings**.
- Post-processing of the outputs is necessary!
- Additional steps: voice tracking and assignment.

8 References

[1] R. M. Bittner et al. **"Deep salience representations for F0 tracking in polyphonic music,"** in Proc. of ISMIR, 2017.

[2] H. Cuesta et al. **"Analysis of intonation in unison choir singing,"** in Proc. of ICMPC, 2018.

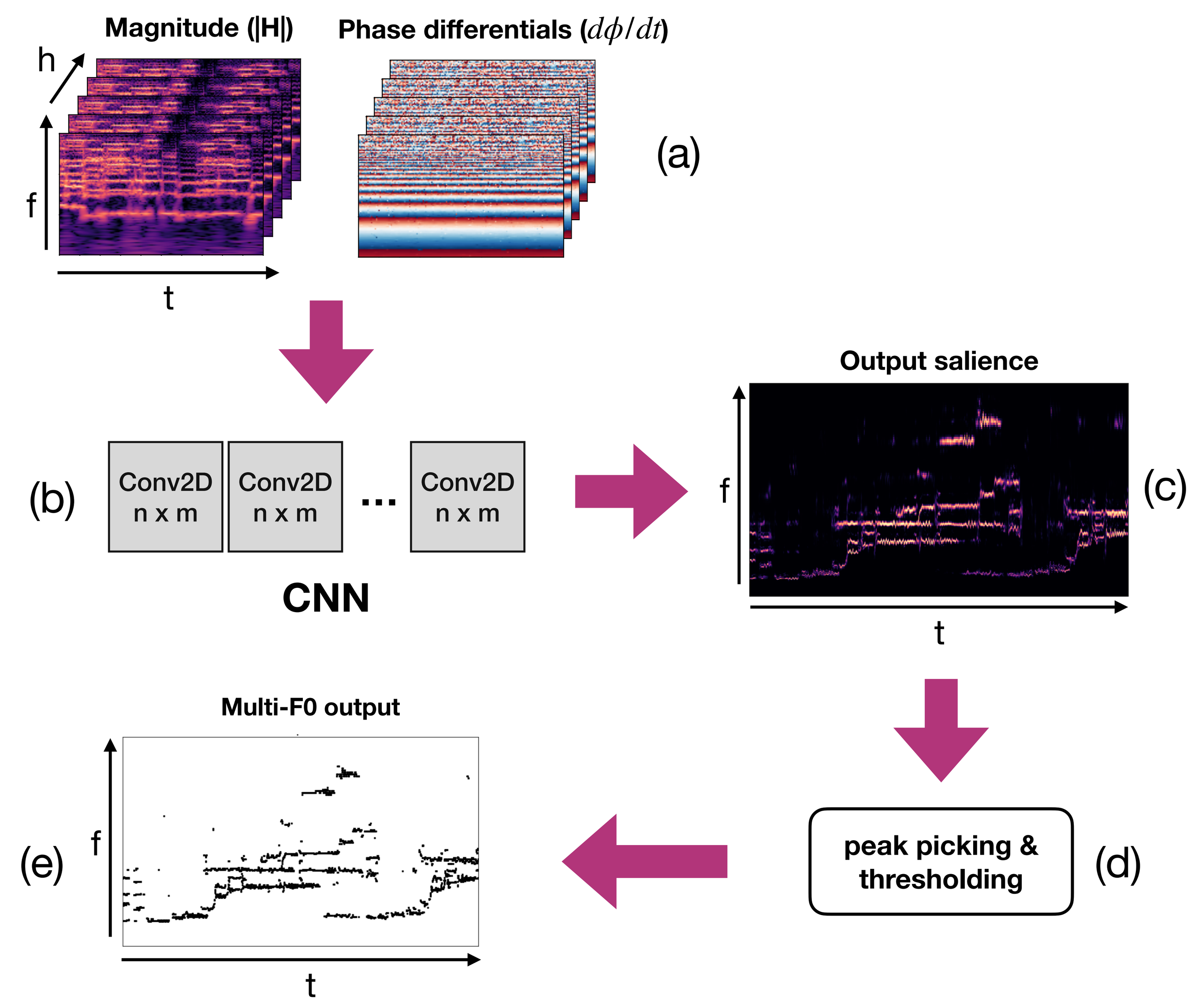
[3] S. Rosenzweig et al. **"Dagstuhl ChoirSet: A Multitrack Dataset for MIR Research on Choral Singing,"** in TISMIR, vol. 3, no. 1, pp. 98– 110, 2020.

[4] R. Schramm & E. Benetos. **"Automatic transcription of a cappella recordings from multiple singers,"** in Proc. of the AES Conference. 2017.

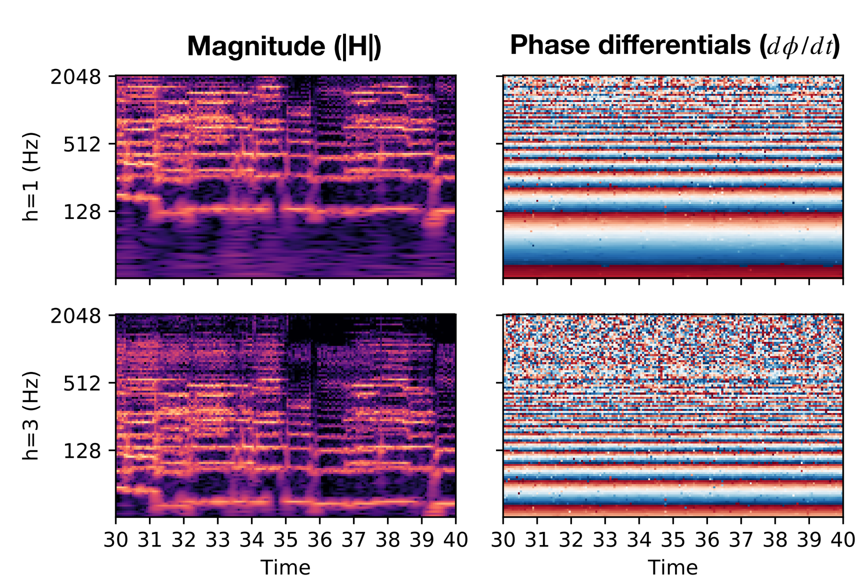
[5] A. McLeod et al. **"Automatic transcription of polyphonic vocal music"**. In Applied Sciences, vol. 7, np. 12. 2017.

[6] L. Su et al. **"Exploiting frequency, periodicity and harmonicity using advanced time-frequency concentration techniques for multipitch estimation of choir and symphony,"** in Proc. of ISMIR, 2016.

2 System overview



4 Input features

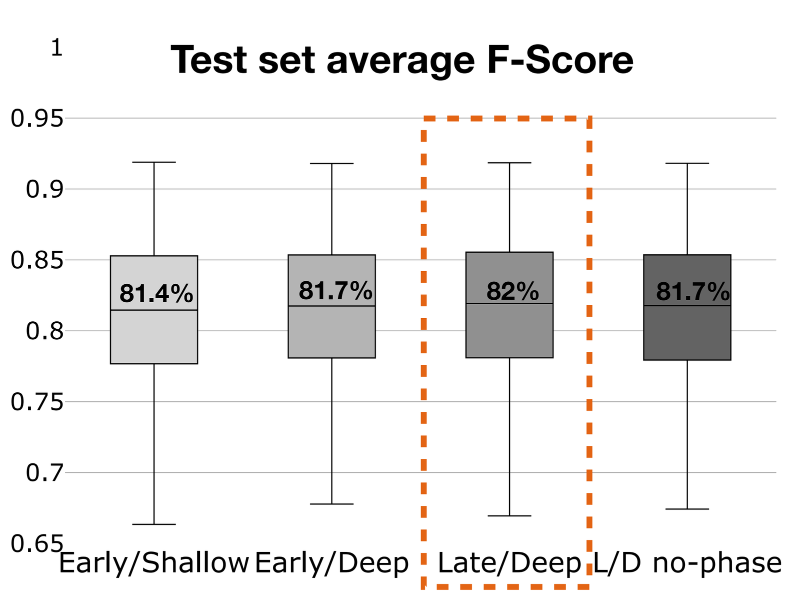


- Harmonic constant-Q transform (HCQT).
- 60 bins/octave, 20 cents/bin, 6 octaves, h=[1,2,3,4,5].
- Magnitude & phase differentials
- Targets are Gaussian-blurred binary time-frequency representations.

6 Experimental results

Experiment 1: **fusion strategy / depth of the network / phase differentials**

- Best performance Late/Deep with phase differentials.
- Very similar overall results (F-Score).
- Precision increases with phase information.



Experiment 2: **comparison to baseline / pitch tolerance**

Method	100 cents			20 cents		
	F	P	R	F	P	R
MSINGERS [4]	0.708 *(0.06)*	0.685 *(0.06)*	0.736 *(0.07)*	0.537 *(0.07)*	0.620 *(0.07)*	0.477 *(0.08)*
VOCAL4-VA [5]	0.757 *(0.06)*	-	-	0.490	-	-
Late/Deep	**0.846** *(0.03)*	0.812 *(0.03)*	0.884 *(0.04)*	**0.831** *(0.03)*	0.797 *(0.03)*	0.868 *(0.04)*

- Late/Deep outperforms baselines on the Barbershop dataset.
- Robust to smaller pitch tolerances - higher pitch resolutions.

Experiment 3: **generalization**

- Outperform baseline on a small dataset of commercial choir recordings [6]: **Late/Deep 70%** F-Score / **Baseline [6] 65%** F-Score.
- Training set w/ dry & reverb signals increases the generalization capabilities of the models in several recording conditions.

Universitat Pompeu Fabra Barcelona · MTG Music Technology Group · MARL NYU Music and Audio Research Laboratory · European Commission · TROMPA