DeepF0: End-to-end Fundamental Frequency Estimation for Music and Speech Signals

Satwinder Singh, Ruili Wang*, Yuanhang Qiu

s.singh4@massey.ac.nz, ruili.wang@massey.ac.nz, y.qiu1@massey.ac.nz

School of Natural and Computational Sciences

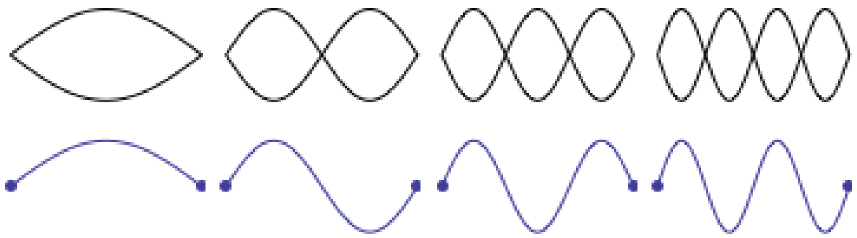
Massey University, Auckland, New Zealand

Outline

- Introduction and background
- Existing research
- Problem statement
- Proposed approach
- Experimental setup and results
- Conclusions

What is Pitch?

- Also known as Fundamental frequency (f0), is a lowest and predominant frequency in complex audio signal.
- Fundamental frequency is regarded as physical property of the signal
- Whereas Pitch is more often used to refer to how the fundamental frequency is perceived.

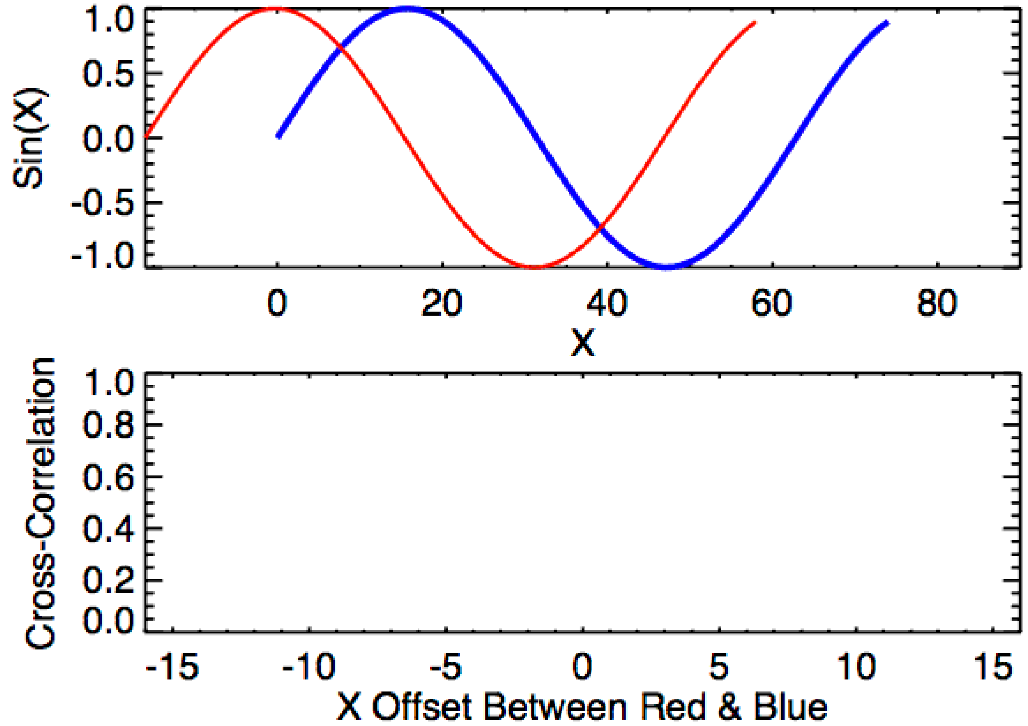


Why it is important?

- Melody extraction
- Gender identification
- Environmental sound classification
- Speech recognition
- Speech synthesis

What has been done?

- Digital Signal Processing (DSP) based methods
  - Mostly based on auto-correlation or cross-correlation function and their variants

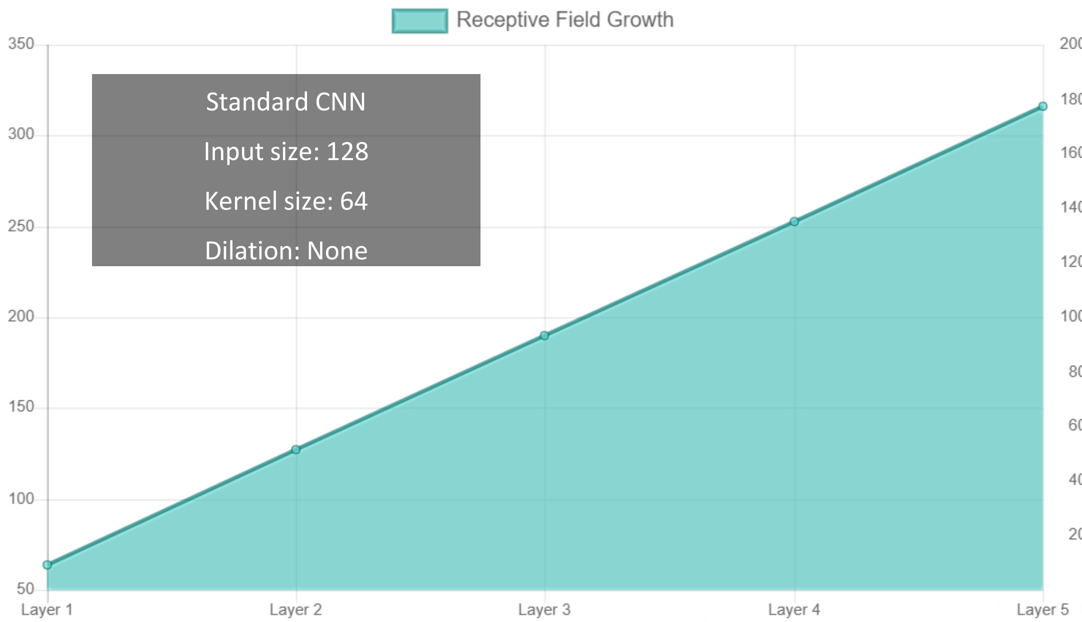


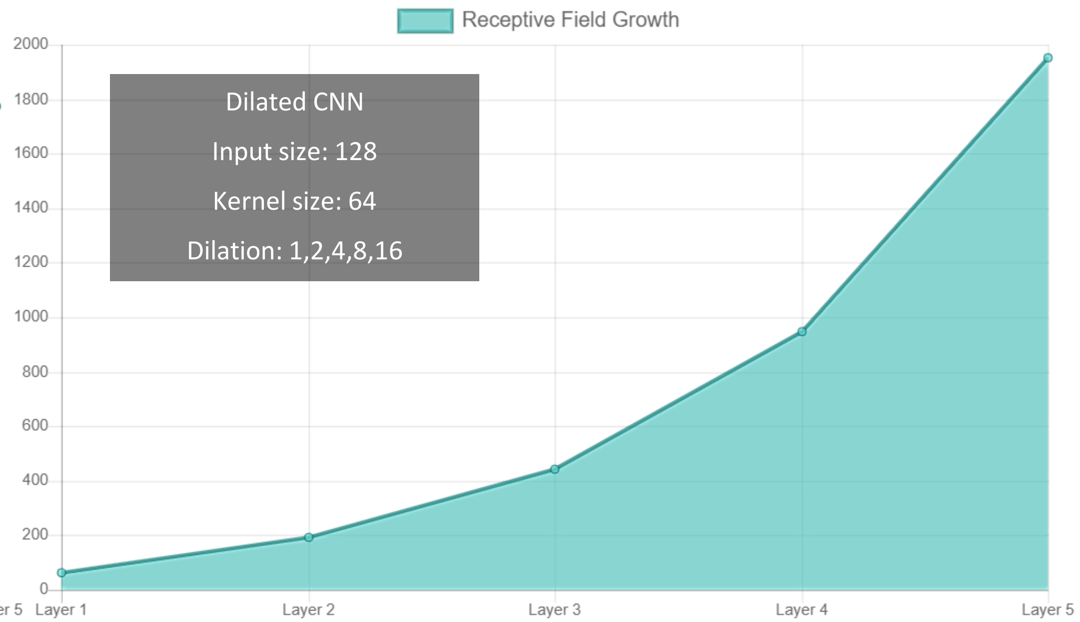
What has been done?

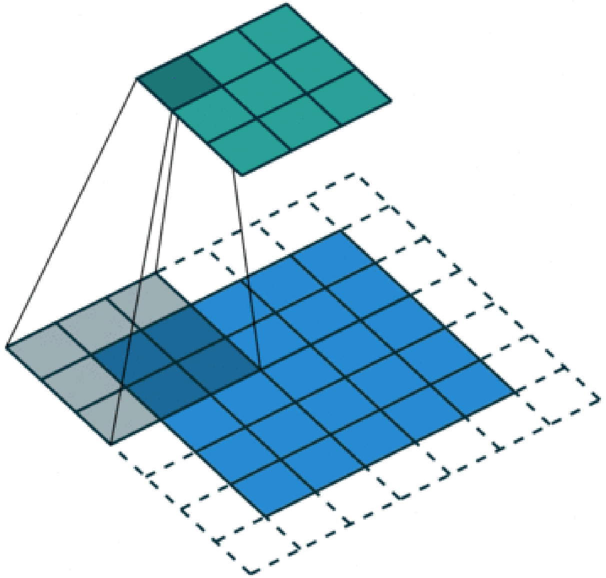
- Data-driven based methods
  - Deep leaning
    - CREPE (Convolution Representation of Pitch Estimation) [Kim et al., 2018]
    - CRN-Raw (Convolution Residual Network) [Dong et al., 2019]
    - SPICE (Self-Supervised Pitch Estimation) [Gfeller et al., 2020]

Problem Statement

- Shallow receptive fields
- Large number of network parameters







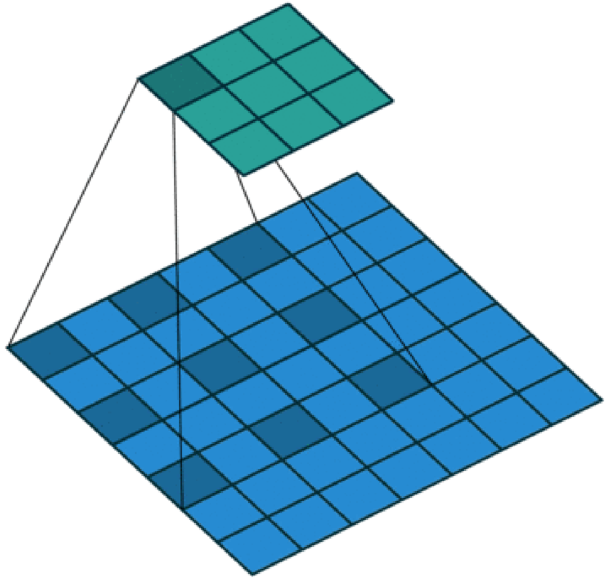
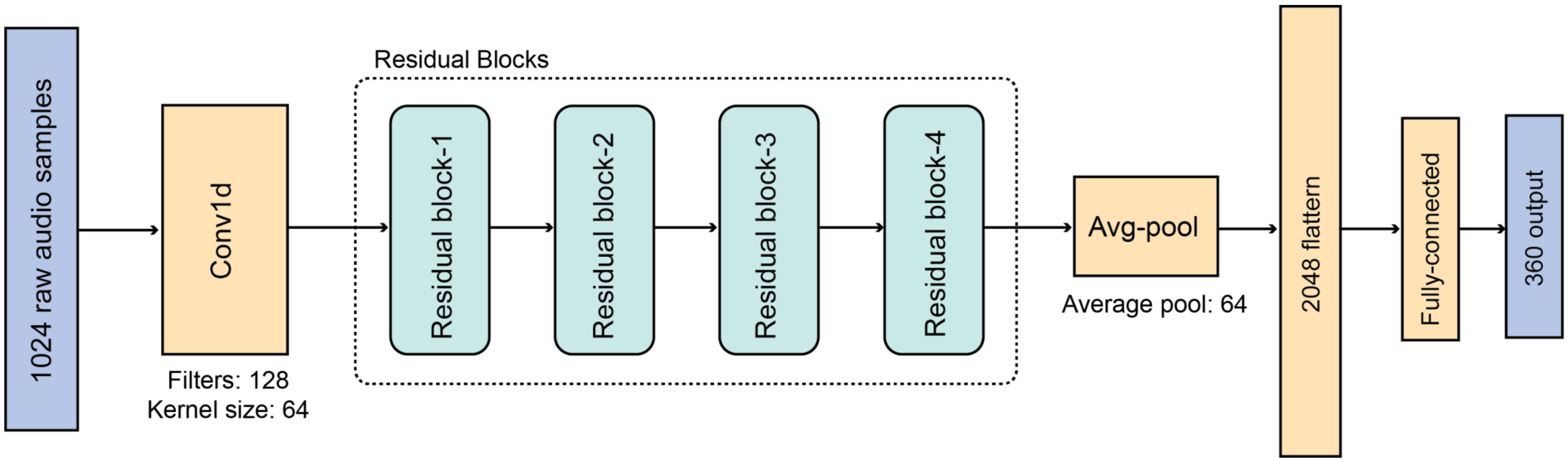


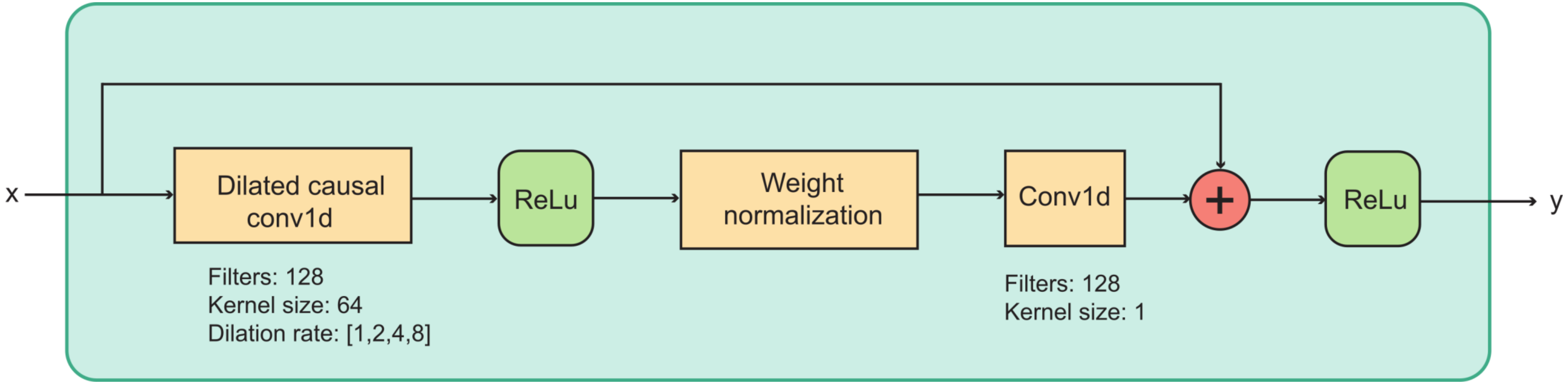
Image source: Dumoulin et al.

Receptive field size comparison (Standard CNN vs Dilated CNN)

Proposed Architecture



Residual Block



Experimental Setup

Datasets

- MIR-1k (Singing Voices)
- MDB-stem-synth (Musical instruments)
- PTDB-TUG (Speaking Voices)

Evaluation measures

- Raw Pitch Accuracy (RPA)
- Raw Chroma Accuracy (RCA)

Baselines

- Convolution Representation for Pitch Estimation (CREPE)
- Sawtooth Waveform Inspired Pitch Estimator (SWIPE)

Experimental Results (Clean audio)

Table 1: Average raw pitch accuracy and raw chroma accuracy and their standard deviation ($\pm$) tested on three different test datasets.

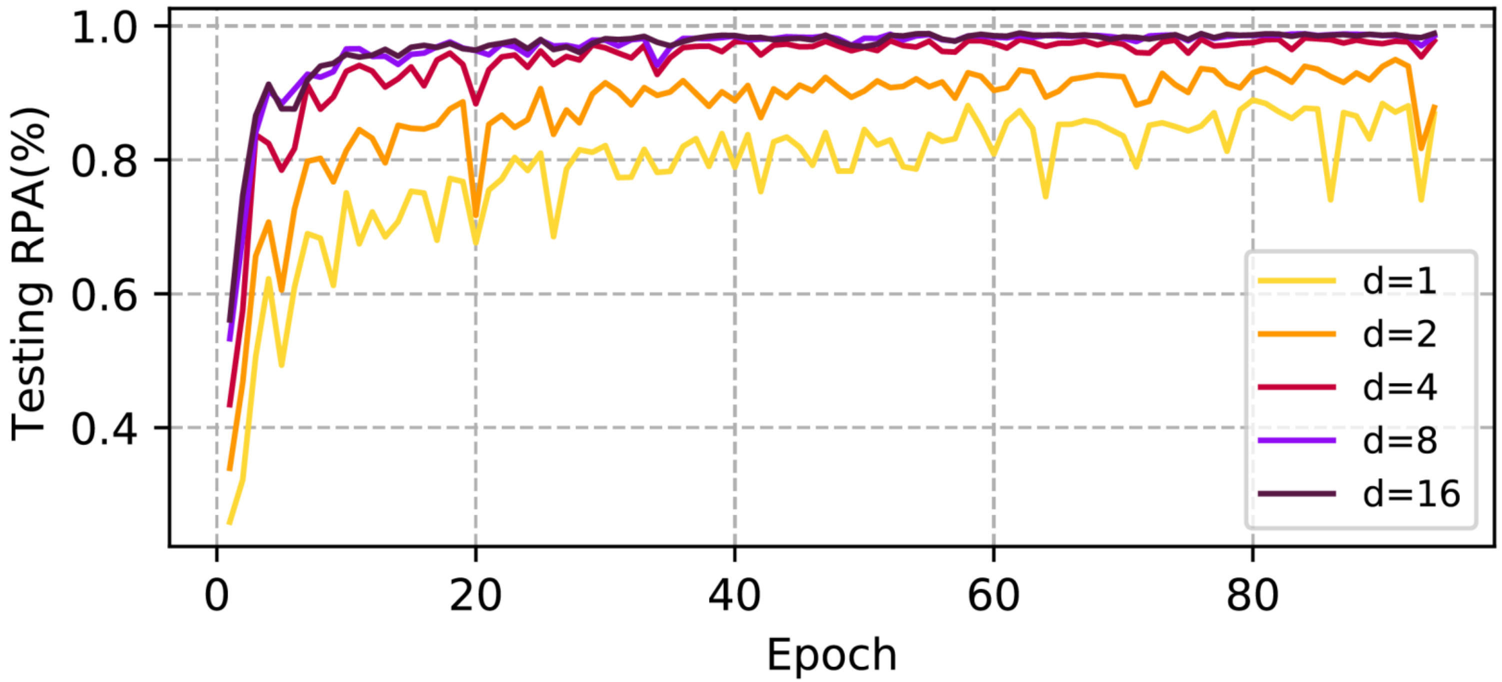
Model	Params	Metrics	Datasets		
			MIR-1k	MDB-Stem-synth	PTDB-TUG
SWIPE	-	RPA (%)	88.73 ±5.43	92.84 ±9.59	87.74 ±7.17
		RCA (%)	89.24 ±5.28	93.83 ±7.69	88.93 ±6.12
CREPE	22.2M	RPA (%)	96.51 ±3.23	97.22 ±4.12	78.18 ±10.07
		RCA (%)	96.84 ±2.56	97.50 ±2.97	79.81 ±9.39
DeepF0	5M	RPA (%)	**97.82 ± 3.34**	**98.38 ± 2.97**	**93.14 ± 3.32**
		RCA (%)	**98.28 ± 1.94**	**98.44 ± 2.87**	**93.47 ± 3.41**

Experimental Results (Noisy audio)

Table 2: Average raw pitch accuracy and raw chroma accuracy and their standard deviation ($\pm$) on MIR-1k dataset with added noise on various levels of SNR.

Model	Metrics	Noise Profile			
		Clean	20dB	10 dB	0dB
SWIPE	RPA (%)	88.73 ± 5.43	84.45 ± 5.64	59.78 ± 11.58	32.04 ± 11.84
	RCA (%)	89.24 ± 5.28	85.31 ± 5.19	62.85 ± 11.07	37.31 ± 12.93
CREPE	RPA (%)	96.51 ± 3.23	96.49 ± 3.32	**95.11 ± 4.65**	**84.92 ± 10.70**
	RCA (%)	96.84 ± 2.56	96.96 ± 2.63	96.18 ± 3.35	**87.85 ± 8.82**
DeepF0	RPA (%)	**97.82 ± 3.34**	**97.39 ± 3.76**	94.77± 6.03	79.52± 14.0
	RCA (%)	**98.28 ± 1.94**	**98.09 ± 2.10**	**96.35 ± 3.72**	84.37 ± 10.71

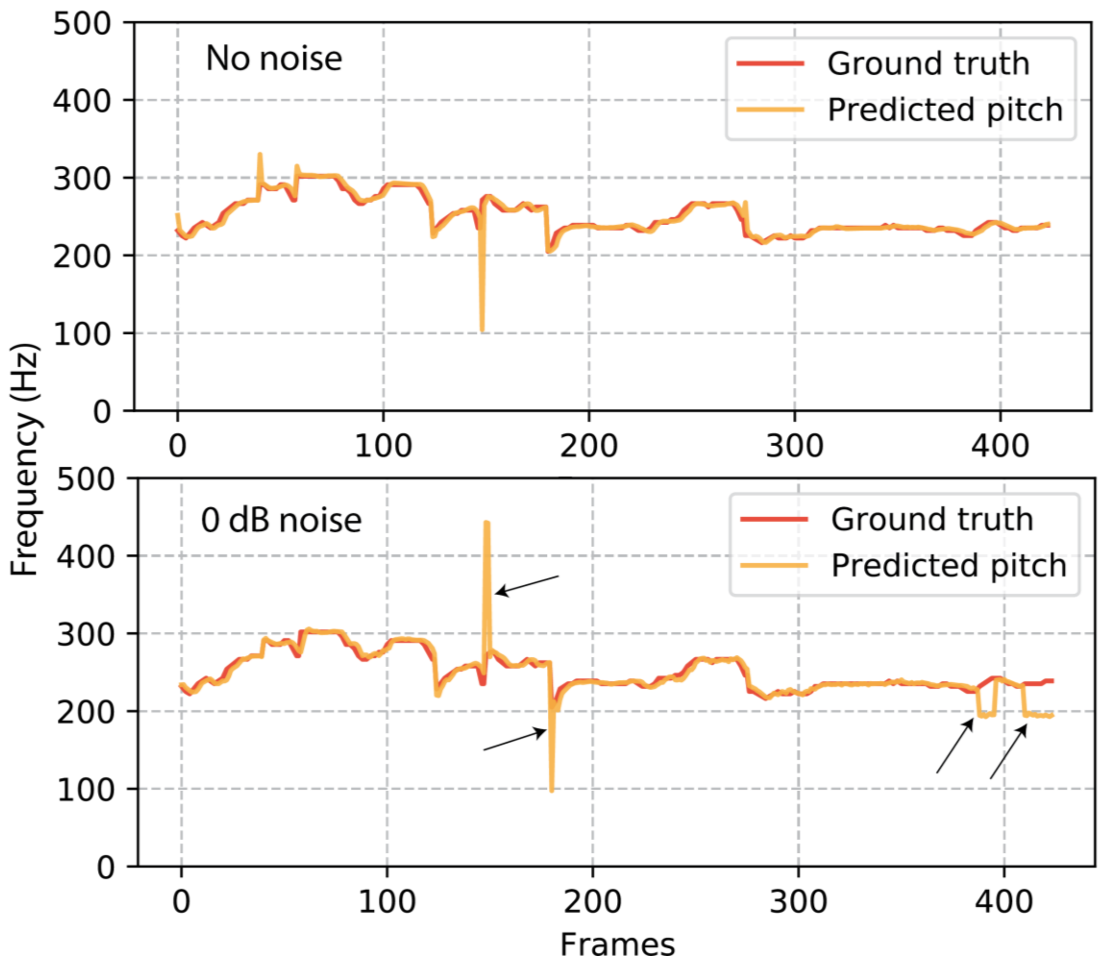
Experimental Results (with different dilation rates)



Pitch Trajectories

(ground truth vs. predicted pitch)

- The estimated pitch trajectories of DeepF0 in comparison with ground truth under clean (top) and 0dB noise (bottom).
- Under no noise scenario DeepF0 produces near perfect pitch estimation, while under noise there are few errors here and there.



Conclusions

- Our proposed model with 77.4% fewer parameters can still perform better than CREPE model.
- Larger receptive field is indeed very important in pitch estimation model.
- We also show that our model can capture reasonably well pitch estimation even under the various levels of accompaniment noise.

References

1. Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello, “CREPE: A convolutional representation for pitch estimation,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 161–165.

2. Mingye Dong, Jie Wu, and Jian Luan, “Vocal pitch extraction in polyphonic music using convolutional residual network,” in 20th Annual Conference of the International Speech Communication Association, 2019, pp. 2010–2014.

3. Beat Gfeller, Christian Frank, Dominik Roblek, Matt Sharifi, Marco Tagliasacchi, and Mihajlo Velimirovic, “SPICE: Selfsupervised pitch estimation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020.

Thank you!

Questions?