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

Satwinder Singh Ruili Wang* Yuanhang Qiu

School of Natural and Computational Sciences, Massey University, Auckland, New Zealand

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

Abstract

We propose a novel fundamental frequency (f_0) estimation technique called DeepF0, which leverages the available annotated data to directly learn from the raw audio in a data-driven manner. f_0 estimation is important in various speech processing and music information retrieval applications. Existing deep learning models for pitch estimation have relatively limited learning capabilities due to their shallow receptive field. The proposed model addresses this issue by extending the receptive field of a network by introducing the dilated convolutional blocks into the network. The dilation factor increases the network receptive field exponentially without increasing the parameters of the model exponentially. To make the training process more efficient and faster, DeepF0 is augmented with residual blocks with residual connections. Our empirical evaluation demonstrates that the proposed model outperforms the baselines in terms of raw pitch accuracy and raw chroma accuracy even using 77.4% fewer network parameters. We also show that our model can capture reasonably well pitch estimation even under various levels of accompaniment noise.

Introduction

The fundamental frequency often called as pitch is the lowest and predominant frequency in a complex periodic signal.

- Pitch estimation has been studied for the last 5 decades due to its central importance in range of speech processing and music information retrieval applications.
- Pitch estimation approaches are categorized into two broad categories: digital signal processing (DSP) based approaches, and data-driven approaches.
- The DSP based methods mostly based on auto-correlation, cross-correlation function and thier variants, which calculate the self similarity between original and the lagged version of the signal.
- These approaches are computationally intensive, not robust in noisy environments, fail when the pitch is rapidly changing, and do not learn anything from available data.
- On the other hand, data-driven approaches take full advantage of the available data and learn the estimation model based on the data itself.

Problem Statement

Although these data driven deep learning based models can outperform digital signal processing-based methods. However, they still have limitations:

- Shallow receptive fields (as illustrated in Fig. 1)
- Large number of network parameters

Proposed Architecture

- To deal with these issues, we proposed dilated temporal convolutional neural networks (Fig.2).
- Dilation introduces the holes in a convolution kernel, which skip certain values, making receptive field bigger than the filter size [1].
- As a result, the receptive field grows exponentially while the number of parameters grows linearly (as illustrated in Fig. 1).
- We incorporated residual blocks and residual connections (Fig. 3) between these blocks, which helped in efficient training of our model [2].
- The raw 16kHz audio input is resampled into 1024 samples in a frame with 160 samples of overlap.
- The model outputs 360-dimensional vector, which represents pitches on the logarithmic scale measured in terms of cents (a unit to measure small musical intervals).

Proposed Architecture

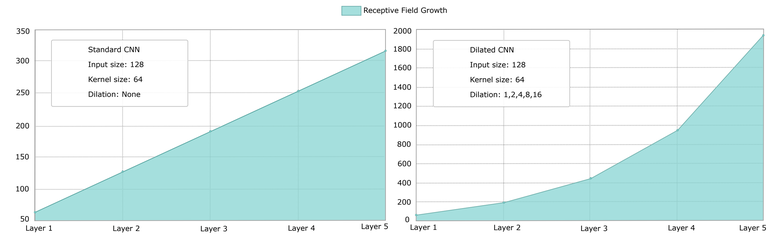

Figure 1. Receptive field size comparison between Standard CNN and Dilated CNN. With 5 layers and kernel size 64 standard CNN achieves receptive field around 300, where as with same number of layers and kernel size dilated CNN with dilation rate upto 16 achieves receptive fields as large as ≈2000.

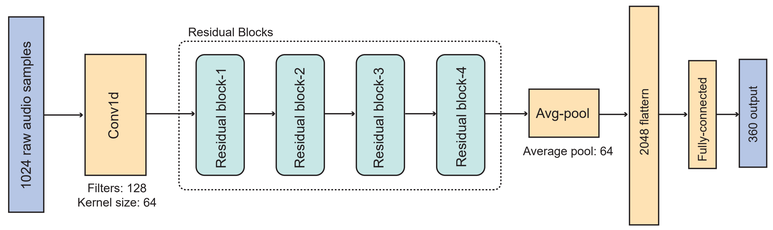

Figure 2. Network architecture of DeepF0.

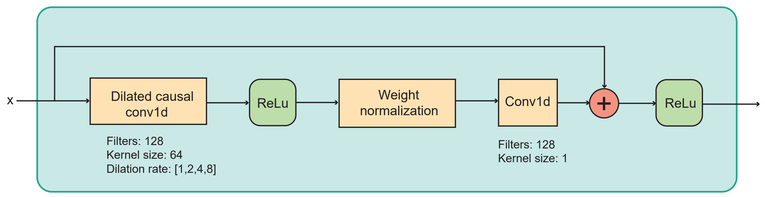

Figure 3. Internal view of a residual block of DeepF0.

Experimental Setup

- The proposed model is trained using 5-fold cross-validation with 60/20/20 split of train, validation, and test, respectively.
- The split is carried out in such a way that no artist/speaker/instrument overlaps with train and test splits.
- The proposed model is also tested on accompaniment noise with SNR levels of 20dB, 10dB, and 0dB.

Datasets

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

Evaluation metrics

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

Baselines

- Convolution Representation for Pitch Estimation (CREPE) [3]
- Sawtooth Waveform Inspired Pitch Estimator (SWIPE) [4]

Results

Pitch Accuracy on Clean Datasets

Table 1. Average raw pitch accuracy and raw chroma accuracy and their standard deviation (±) 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	97.55 ± 3.43	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**

Pitch Accuracy on Noisy Dataset

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

Model	Metrics	Noise Profile			
		Clean	20dB	10dB	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

Model Analysis

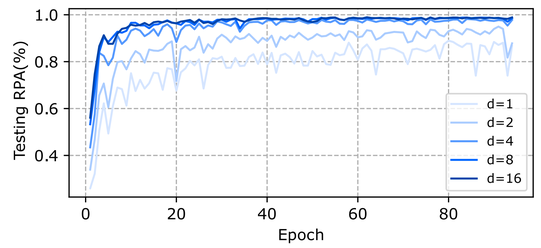

Figure 4. Evaluation results of the proposed model with different dilation rates on the MDB-stem-synth dataset. Dilation rate $d = 8$ shows the best results.

Conclusions

- Our proposed model outperforms the baselines in terms of raw pitch and chroma accuracy.
- DeepF0 is also efficient in terms of network parameters used. It uses 77.4% fewer network parameters as compared to the CREPE model.
- Further, we find that the larger size of the receptive field of the network is very significant in pitch estimation, which aids in achieving excellent results with consistently low variance.

References

[1] Aaron van den Oord et al. "WaveNet: A generative model for raw audio". 2016.

[2] Kaiming He et al. "Deep residual learning for image recognition". In "IEEE CVPR", pp. 770–778. 2016.

[3] Jong Wook Kim et al. "CREPE: A convolutional representation for pitch estimation". In "IEEE ICASSP", pp. 161–165. 2018.

[4] Arturo Camacho et al. "A sawtooth waveform inspired pitch estimator for speech and music". In "The Journal of the Acoustical Society of America", vol. 124, pp. 1638–1652. 2008.