Melody Extraction

Aman Verma
Roll No - 180075
Department of Electrical Engineering
Indian Institute of Technology Kanpur
amanve@iitk.ac.in

Raj Prakash Gohil
Roll No - 19104407
Department of Electrical Engineering
Indian Institute of Technology Kanpur
rgohil@iitk.ac.in

Deshpande Prasad Jayant
Roll No - 17103270
Department of Civil Engineering
Indian Institute of Technology Kanpur
prasadj@iitk.ac.in

April 19, 2021

Abstract

This paper discusses a CREPE: A Convolutional Representation for Pitch Estimation algorithm as a baseline model used for melody extraction. The model algorithm is trained on MIR-1K and tested on MIR-50K dataset. The highlights and limitations of CREPE are discussed. To overcome the limitations, another algorithm will be tested in future which is Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness.

1 Introduction

Music Information Retrieval (MIR) is an interesting yet challenging task in audio signal processing. It includes pitch extraction, beat tracking, tempo estimation, source separation etc. The end-user applications of MIR include searching-by-humming, searching-by-example, searching-by-notes etc. It may also include searching songs based on mood or genre. [1]

CREPE algorithm is a data-driven algorithm developed in 2018, which is highly precise, maintaining over 90 percent raw pitch accuracy (RPA) even for a strict evaluation threshold of just 10 cents. [2]. Its Python implementation along with a pre-trained model makes it easy to utilize and reproducible. In this report, the ready made library and pre-trained model is not used. The further details about CREPE are mentioned in Methodology section.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

1.1 Highlights of CREPE algorithm

• CREPE outperforms popular algorithms e.g. pYIN (combination of signal proecessing pipelines and heuristics) [3] and SWIPE (a sawtooth waveform inspired pitch estimator for speech and music) [4]

• CREPE algorithm operates directly on the time-domain of the audio signal, and hence does not need any separate feature extraction method.

The CREPE algorithm will act as a reference i.e. baseline model for further algorithm development.

2 Problem definition

Melody extraction problem can be defined in 2 ways -

a) Identification of dominant frequency of the music (which in measured in Hz). This problem is mathematically modelled as a regression problem.

b) Identification of Swaras/Notes of the music. In this case, the western notes in terms of C, G etc. are found out in terms of absolute frequency. Indian music Swaras in terms of Sa, Re, Ga etc. are found out w.r.t. a suitable reference frequency.

CREPE algorithm is originally designed for classification problem, but we are also testing its utility for regression problem also.

3 Dataset description

3.1 MIR-1K dataset

MIR-1K is an acronym of Multimedia Information Retrieval (MIR) lab's 1000 song clips (1K) dataset. The dataset consists of clips extracted from 110 karaoke songs which contain a mixture track and a music accompaniment track. The duration of each clip ranges from 4 to 13 seconds, and the total length of the dataset is 133 minutes.

MIR-1K dataset provides the timing of ground truth of pitch frequency at a sampling interval of 0.01 seconds. Hence, this dataset is used for modelling the melody extraction problem as a regression problem. Before using the MIR-1K dataset for classification problem, the pitch frequencies need to be converted into corresponding notes. This procedure is described in the Preprocessing section of the Methodology.

3.2 Bach10 dataset

Bach10 dataset is a versatile polyphonic music dataset.This dataset is widely used for research problems, such as multi-pitch estimation and tracking, audio-score alignment, source separation, etc. The dataset consists of MIDI score of polyphonic dataset, the audio files of the separate instruments and the text files containing the onset timings of the different notes of the instruments.

Since this dataset provides the notes, this dataset is used for modelling the melody extraction problem as a classification problem.

We are also searching for Indian classical music datasets, or creating a simple classical music dataset by our own.

The original CREPE paper uses RWC-synth dataset and MDB-stem-synth datasets , in this report MIR-1K dataset is used for training and MIR-50K is used for testing.

4 Methodology

4.1 Feature extraction

Several conventional algorithms make use of hand-designed features, which are further used by used classifiers. These feature extraction methods include SFFT (Short Term Fourier Transform), MFCC

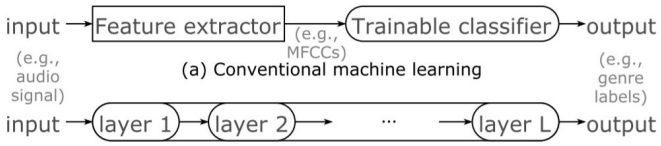

Figure 1: Feature extraction for pitch estimation by different approaches. [5]

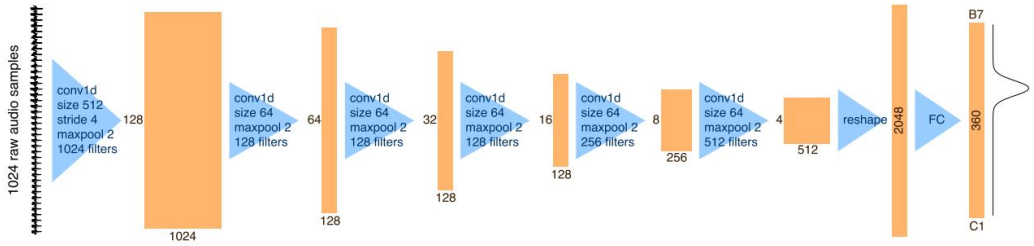

Figure 2: Model architecture [2]

(Mel-frequency cepstrum coefficients) etc. Output labels are mapped to these extracted features. Hence, a part of procedure is learnt from data while the remaining part (i.e. feature extraction) is independent of the data. [5] On the other hand, deep learning algorithms (like CREPE) do not necessarily need separate feature extraction step. It is nothing but end-to-end learning. CREPE algorithm operates directly on the time-domain (unlike time-frequency approach which is widely used in other models) of the audio signal, and hence does not need any separate feature extraction method.

4.2 Model description

The audio input is sampled at 16kHz frequency and a set of 1024 samples in given as input to the input layer. There are total 6 convolutional layers in the deep neural network, which reduces the input dimension from 1024 to 128 and then again expands to 512. In between the convolutional layers, the maxpooling and dropout (of 0.25) operations are carried out. The final output layer consists of a fully connected layer with sigmoid activations corresponding to 70-dimensional output vector. This output vector is used to estimate the pitch deterministically. Binary cross-entropy loss is minimised between the target vector and the predicted vector. The neural network is implemented in Keras library of Python.

Labels are provided in csv format containing pitch frequencies at a constant time interval (0.01 seconds). The network is trained to minimize the binary cross entropy between the target vector y and the predicted vector where both vectors are real numbers between 0 and 1. This loss function is optimized using the ADAM optimizer, with the learning rate 0.0002.

4.3 Accuracy evaluation

The output vector is not directly used for accuracy assessment, as it comes in the unit cents. Cent is a unit which represents musical intervals w.r.t. a reference frequency, this unit provides a logarithmic scale is for pitch estimation.

$$c(f) = 1200 \log_2 \frac{f}{f_{ref}}$$

If we plot the notes in terms of Hz, they are found unequal spaced because of their logarithmic scale. Here, the 12 semitones of an octave are located in a ascending order of ratios 1:1 to 2:1. These ratios are frequencies of notes divided by the reference frequency. Converting frequencies into scales makes the all the semitones equidistant.

Predicted pitch is said to be correct (or within acceptable limits) if it lies within half semitone above or below the ground truth semitone. This criteria is used to calculate the RPA (Raw Pitch Accuracy).

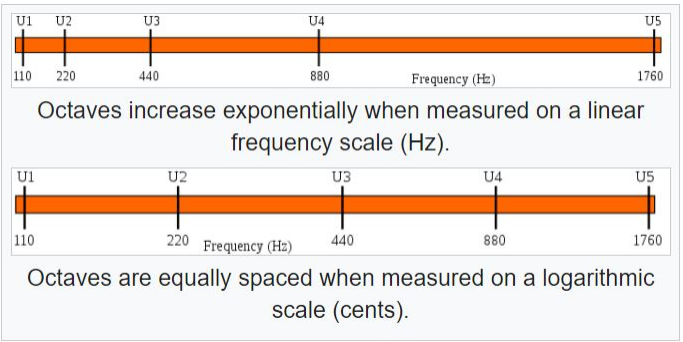

Figure 3: Comparison of linear and logarithmic scale of pitch

5 Results and discussion

5.1 Results

For a batch size of 16 and 5 epochs, the training accuracy is 0.7542 and the validation accuracy is 0.7733. The average RPA obtained for 10 test audio files from MIR-50K dataset is 33.25. The variation of RPA for different test data varies from 13.38 to 51.78.

The obtained results are on a lower side as compared to results presented in the original CREPE paper. The lesser training epochs, lesser training data, and different dataset can be the reasons attributed for such results.

5.2 Limitations of CREPE algorithm

CREPE algorithm generates an a deterministic output. On the other hand, in many music forms (e.g. Indian classical music), the notes played do not belong to exact location of the notes, but there are little perturbations (Vibrato / Murki), or continuous shift of one note to another note (slide). In such cases, deterministic identification of notes is not the best way to identify notes, but one can identify them probabilistic way. In this case, uncertainty or confidence interval associated with the prediction. The results of recent advances in the field of uncertainty estimation of machine learning are promising. Hence, the application of such state-of-art methodologies will also help in case of melody extraction problem. The algorithm from the paper 'simple and principled uncertainty estimation with deterministic deep learning via distance awareness' [6] is chosen for melody extraction with associated uncertainty as a further step.

6 Conclusion

In this report, the deterministic melody extraction using CREPE algorithm is demonstrated for MIR-1K dataset. Also, the limitations of the CREPE algorithm and future scope of uncertainty estimation of the extracted melody is discussed.

References

[1] Roberto Raieli. 6 - the current status of mir systems. In Roberto Raieli, editor, *Multimedia Information Retrieval*, Chandos Information Professional Series, pages 175–193. Chandos Publishing, 2013.

[2] J. W. Kim, J. Salamon, P. Li, and J. P. Bello. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165, 2018.

[3] M. Mauch and S. Dixon. Pyin: A fundamental frequency estimator using probabilistic threshold distributions. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663, 2014.

[4] Arturo Camacho and J. Harris. A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, 124 3:1638–52, 2008.

[5] Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark Sandler. A tutorial on deep learning for music information retrieval. *arXiv*, 2017.

[6] Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness, 2020.