Teacher-Student Networks for Melody Estimation

Prof. Vipul Arora, Bhavesh Jain, Ramyata Pate, Udhav Gupta
Department of Electrical Engineering
Indian Institute of Technology, Kanpur

May 18, 2021

1 Introduction

The task offered by problem of melody extraction is to obtain the frequency values corresponding to the dominant audio signal in the input monophonic or polyphonic audio data. This problem comprises of many sub-tasks, including, selection of frequency values present as fundamental frequency, figuring out the frequency corresponding to the most dominant voice.
The melody extraction problem from audio signals gets complicated when we start dealing with polyphonic audio data. This is because in generalised audio signals, the sounds are highly correlated over both frequency and time domains. This complex overlap of many sounds, makes identification of predominant frequency challenging.
In advancement to this, we can deal with filtering background noise in the data, distinguish the components of audio data as voiced and non-voiced, perform instrument recognition. Some mainstream applications are query-by-humming, cover song identification, genre classification, automatic generation of karaoke accompaniment and singer characterization.

2 Targeted Problem

Melody estimation or melody extraction refers to the extraction of the primary or fundamental dominant frequency in a melody. This sequence of frequencies obtained represents the pitch of the dominant melodic line from recorded music audio signals. The music signal may be monophonic or polyphonic.
Monophonic Signal: Signal is generated from a single source.
Polyphonic Signal: Signal is a combination of signals generated by multiple sources simultaneously.

Melody estimation in a monophonic audio is easier as compared to polyphonic audio. In this paper we have targeted melody estimation for polyphonic audio.

We have compared 2 baseline models based on Signal Processing and patch-based Convolutional Neural Networks in this paper. We will be implementing Student-Teacher over the 2nd model for learning from less data using a semi-supervised approach with reference to the following paper-

`https://arxiv.org/pdf/2008.06358.pdf`

3 Dataset Used

We have used the dataset MIR-1K for training the model. It contains 1000 audio clips with pitch labeling for window size 40ms and hop size 20ms. Most audio clips are extracted from Karaokes performed mostly by amateurs. Data comprises of male and female audio sounds alike. Unvoiced sounds like friction sounds, inhaling sound, sound of instruments, are also added to the dataset. The datasets used for evaluating the models are MIREX-05 and adc2004.

4 Methods

4.1 Signal Processing

In this approach we have leveraged the basics of signal processing by computing the auto-correlation of the STFT representation of the input signal. Fundamental frequency is the number of samples in one time period. This time period can be taken to be the distance between the first 2 peaks in the auto-correlation.

4.2 Patch based CNN

We take a novel representation in time-frequency domain i.e CFP. This spectrogram is split into patches of size 25×25. These patches are used as an input to CNN. The CNN comprises of two convolutional layers, followed by three fully connected layers. The convolutional layers have $8, 5 \times 5$ filters and $16, 3 \times 3$ filters. The number of units in fully connected layers are 128, 64, 2 respectively. Now based on the highest frequency present in a patch, it is classified as either vocal melody(labelled 1) or not (labelled 0). As only small portion of our input data contains vocal melody, data imbalance occurs. Hence, 10% of non-vocal peaks are randomly selected into training data.
.The model is trained using the binary cross entropy loss between the CNN output and the ground truth is minimized using batch gradient descent with the Adam optimizer.
The CNN outputs the probability of the patch being a Vocal Melody. If the output probability$>$0.5, then it is considered as vocal melody. Hence, we have the required spectrogram, with vocal melodies set as 1 and others as 0.

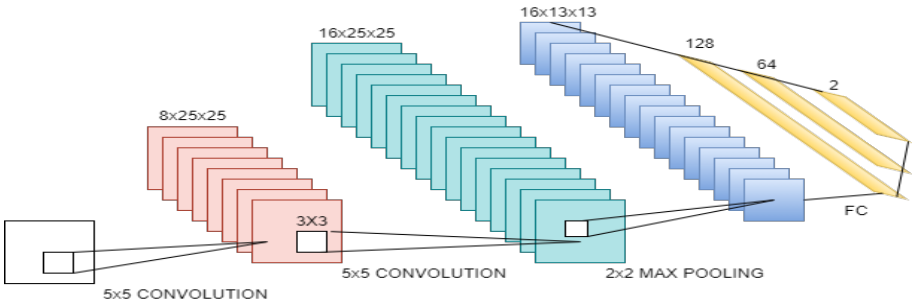

Figure 1: CNN

4.3 Model Architecture

We propose convolutional recurrent neural network (CRNN) as the baseline architecture. The CRNN architecture consists of 2 ResNet blocks and a bi-directional long short-term memory layer. We first merge the audio waveforms into a mono channel and downsample them to 8 kHz. We then calculate the logarithmic-magnitude spectrogram using short-time Fourier transform with a 1024-point Hann window and an 80-point hop size.The CRNN architecture takes 31 consecutive frames of the spectrogram as input and predicts a pitch label quantized with a resolution of 1/8 semitone. The size of the output layer is 442.

4.4 Proposed Teacher-Student Models

Initially we train out Teacher network in a supervised manner i.e with true labels on a dataset $\mathcal{D}$ using the cross entropy loss function.
After this we train the Student network in a semi-supervised manner i.e. on a different dataset $\mathcal{U}$, we generate the pseudo labels from our Teacher network and use them as true labels for training the Student network with the following cross-entropy function.

$$\mathcal{L}_b = \frac{1}{M}\sum_{u=1}^{M}(H(\dagger_u, p(y|x_u;\Theta_s)) + H(\dagger_u, \dagger_t))$$

Where $\dagger_u$ is the pseudo labels generated by Teacher Network on $\mathcal{U}$ and $p(y|x_u;\Theta_s)$ are the predictions made by Student network. And $\dagger_t$ are true labels.

5 Feature Extraction for patch based CNN

A patch captures events localised in time as well as frequency (or pitch). To effectively localize a pitch event in frequency domain without interference from the harmonics, we use the Combined Frequency and Periodicity representation (CFP). The CNN model proposed in the baseline is then applied on patches selected from this representation.

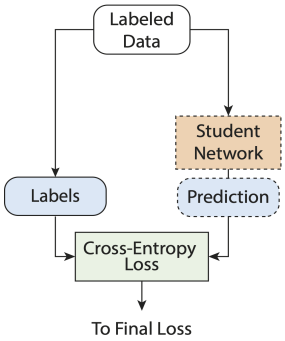

Figure 2: Supervised Loss

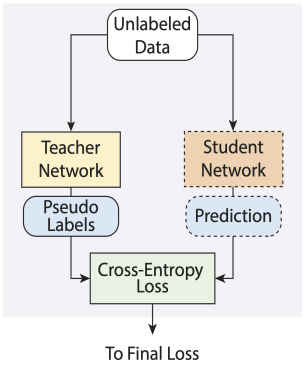

Figure 3: Basic Teacher-Student Loss

5.1 CFP representation:

The path-based CNN uses the Combined Frequency and Periodicity (CFP) approach to represent data. This CFP representation is a combination of the Generalized Cepsrtum (GC) and Generalized Cepstrun of Spectogram (GCoS). The GC is obtained by removing the slow-varying portions in the time domain, resulting in sub-harmonics in the lower frequency range. On the contrary, the GCoS majorly represents the harmonics (higher frequency) as the slow varying-portions in the frequency domain are removed. $\mathbf{X}$:=STFT of input signal amplitude
n, k:=Time and frequency indices respectively
q:=Quefrency
$\mathbf{W}_f$, $\mathbf{W}_t$:=High Pass Filters
$\mathbf{F}$:=N-point Discrete Fourier Transform Matrix
σ_i:=Activation function

$$Spectogram : \mathbf{Z}_0[k,n] = \sigma_0(\mathbf{W}_f\mathbf{X})$$
$$GC : \mathbf{Z}_1[q,n] = \sigma_1(\mathbf{W}_t\mathbf{F}^{-1}\mathbf{Z}_0)$$
$$GCoS : \mathbf{Z}_2[k,n] = \sigma_2(\mathbf{W}_f\mathbf{F}\mathbf{Z}_1)$$
$$CFP : \mathbf{Y}[p,n] = \tilde{\mathbf{Z}}_1[p,n]\tilde{\mathbf{Z}}_2[p,n]$$

Before calculating the final CFP representation, we map GC and GCoS to log-frequency domain (similar to pitch). The CFP representation mainly consists of the fundamental frequency as the harmonics and sub-harmonics get suppressed on combining GC and GCoS.

5.2 Patch Selection

We assume for every frame in Y, the peak is a vocal melody. We select a patch around each such peak of size 25×25. These patches are the input for the CNN

model we have used above.

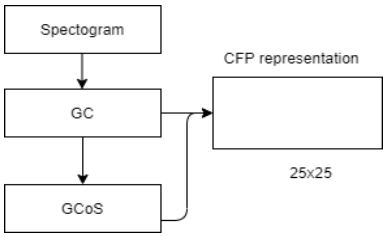

Figure 4: Data Representation

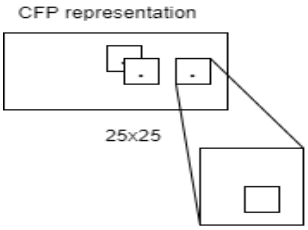

Figure 5: Patch Selection

6 Results and Conclusions

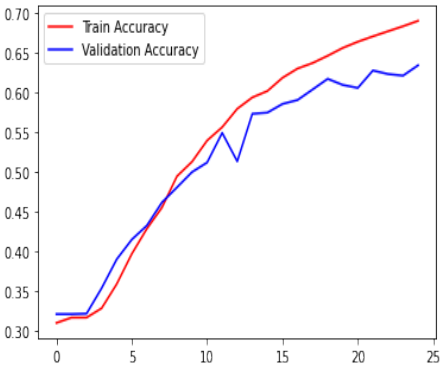

Figure 6: Accuracy for CRNN model

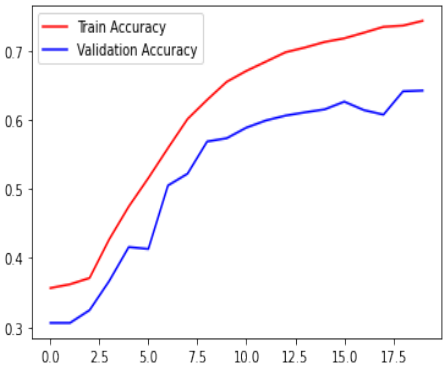

Figure 7: Accuracy for Teacher Student model

Method	ADC2004	MIREX05
CRNN	20.29/28.14	36.87/45.54
Teacher Student	33.69/41.89	65.67/69.92
ResNet NS	34.44/39.05	76.84/79.07
Patch CNN	-/-	62.86/64.54
SP	5.01/22.18	14.86/34.86

For 2 datasets ADC2004 and MIREX05, from different methods we have evaluated RPA/RCA values. ResNet RS is the model mentioned in the paper, and hence is used by us for comparison.

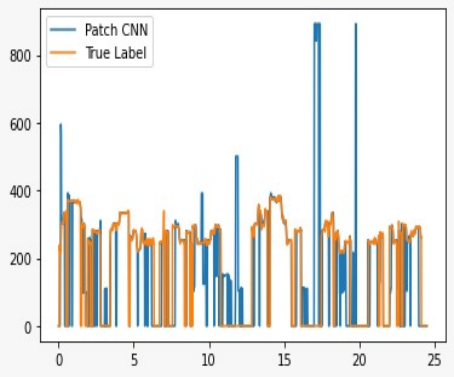

Figure 8: result1

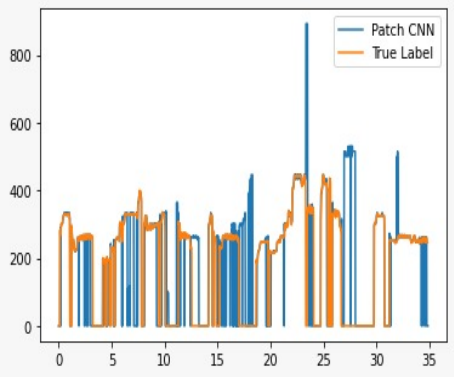

Figure 9: result2

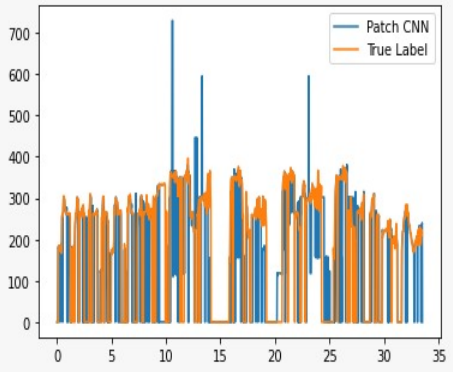

Figure 10: result3

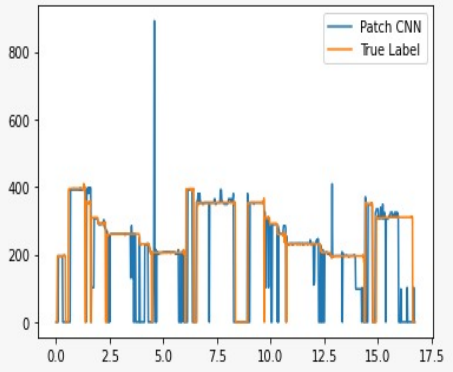

Figure 11: result4

7 Acknowledgement

We thank Prof. Vipul Arora and all the Teaching Assistants of the course EE698R, 2021 for their constant guidance.

References

[1] *http* : *//mirlab.org/dataset/public/*
[2] *https* : *//ieeexplore.ieee.org/document/*8462420
[3] *https* : *//wp.nyu.edu/ismir*2016*/wp−content/uploads/sites/*2294*/*2016*/*07*/*072$_{P}$*aper.pdf*
[4] *https* : *//ieeexplore.ieee.org/document/*8462534
[5] *https* : *//ieeexplore.ieee.org/document/*6739213
[6] *https* : *//arxiv.org/pdf/*2008.06358*.pdf*