EE798P: Audio Representation Learning
P4: Melody Estimation

We have not yet finalised a particular paper or method to follow for melody estimation but here’s a general compilation method

Intro duction

Melody extraction is the complex task of automatically isolating the dominant melodic line within a polyphonic music signal, where multiple instruments or notes can sound concurrently. It involves representing the melody as a sequence of fundamental frequency (F0) values corresponding to the perceived pitch of the dominant instrument. The challenge lies in dealing with the superimposed sounds of various instruments and determining the correct sequence of pitch values. This task involves identifying melody regions, ensuring F0 values are within the correct octave range, and selecting the right melody pitch when multiple notes are present simultaneously.

Methods and Motivation Behind them

1. Feature Representation: We are gonna use the following as feature

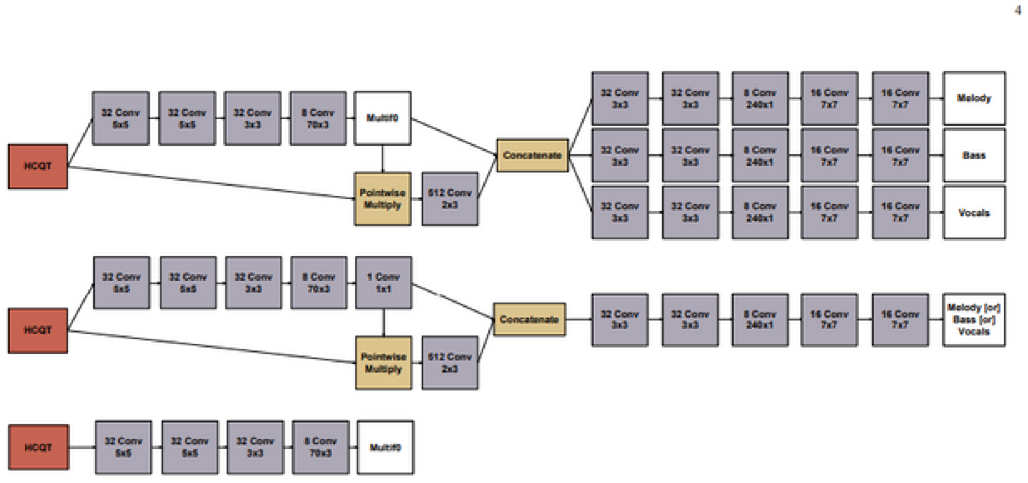


Fig. 3. (Top) Multitask architecture producing multiple-f_0, melody, bass and vocal salience outputs. (Middle) Single-task architecture producing Melody (or Bass or Vocal) salience output. (Bottom) Single-task architecture producing multiple-f_0 output. Red boxes represent HCQT input features, grey boxes denote convolution operations, yellow boxes denote other mathematical operations, and white boxes are used for output layers from which loss is measured.

Figure 1: Multitask Learning for Fundamental Frequency Estimation in Music

We are using HCQT and there concatenated layer as one of our feature maps. Right now we are still exploring we are also working on CFP(combined frequency and periodicity), NMF(negative matrix factorisation), MFCC and some others.

2. Approaches: 1st: method/paper that we are referring are CNN based approach.

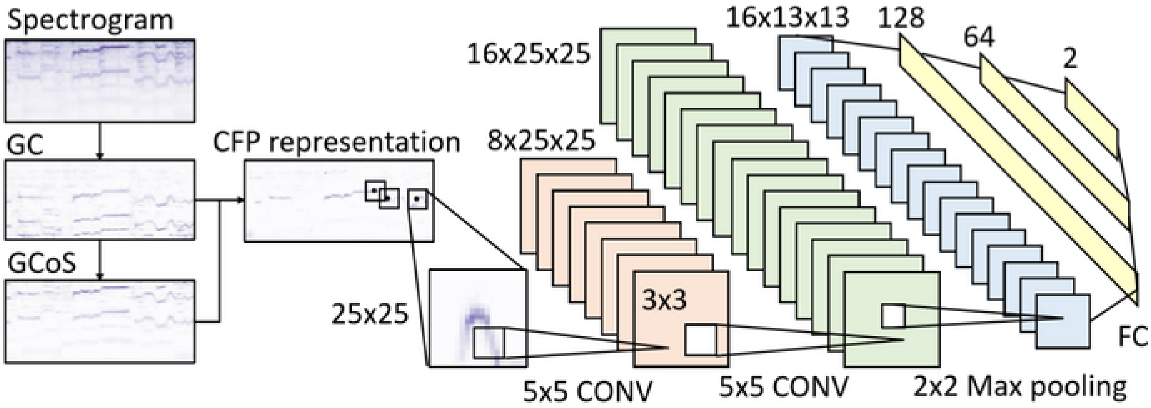
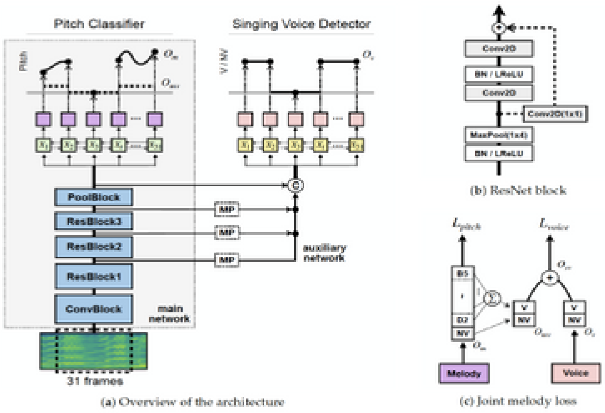


Fig. 1. The proposed system.

Figure 2: VOCAL MELODY EXTRACTION USING PATCH-BASED CNN

2nd: is based on attention for this we are using models based on LSTM and GRU.



(a) Overview of the architecture

(b) ResNet block

(c) Joint melody loss

Figure 13: Illustration of the overall architecture, the ResNet block and the joint loss estimation of the joint detection and classification approach [33].

Figure 3: VOCAL MELODY EXTRACTION USING PATCH-BASED CNN

By using 1st and 2nd we are exploring what are the capabilities of both when used individually. 3rd we are exploring and combined approach.

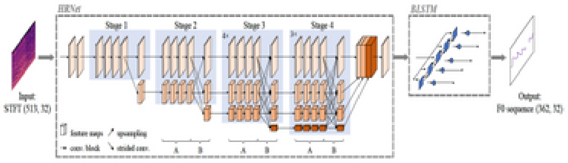


Figure 18: Illustration of the HRNet-BLSTM architecture melody extraction [16].

Figure 4: HRNet-BLSTM architecture melody extraction

3. Separation based Melody Extraction Models : We are also using separation-based melody and exploring the algorithm used there.

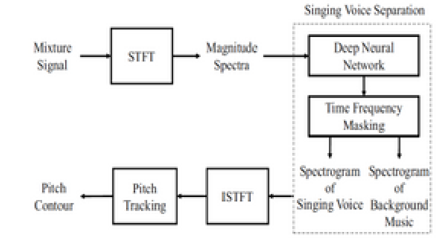


Figure 28: A two stage singing voice separation and dynamic programming based melody extraction from the separated vocals [32].

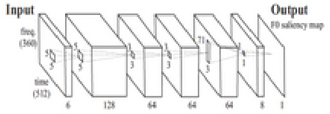


Figure 29: F0 saliency estimation network [30].

Figure 5: 1st: Based on inverse STFT

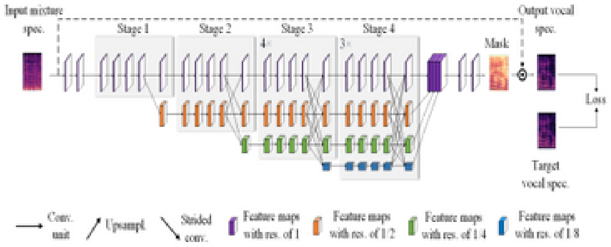


Figure 32: High resolution HRNeT for singing voice separation [29].

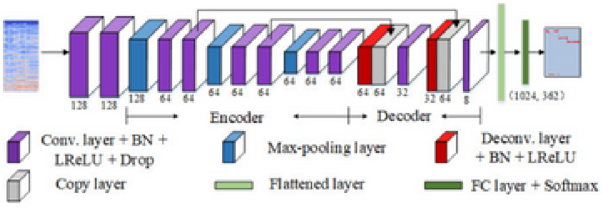


Figure 33: Encoder-decoder based melody extraction model [29].

Figure 6: 2nd: Encoder-decoder based melody extraction model