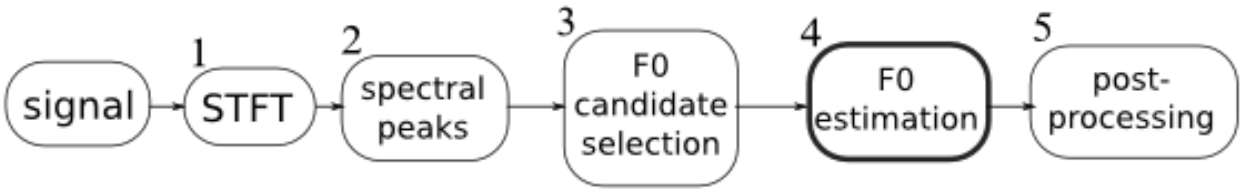
Multi-pitch algorithm overview

Reference

Zhiyao Duan, Bryan Pardo and Changshui Zhang, Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions, IEEE Trans. Audio Speech Language Process., vol. 18, no. 8, pp. 2121-2133, 2010.

Main processing blocks



For stages 1 and 2, existing Essentia processing modules can be re-used. Stages 3-5 need to be ported from the Matlab implementation.

F0-candidate selection

F0 candidates are estimated from the detected spectral peaks in each frame. As a simplification, only the lowest 5, the highest 5 and the strongest 5 peaks are kept, leading to a maximum of 15 peaks per frame. Each peak is associated with 13 F0 candidates: the detected peak, and surrounding frequencies: at -6%, -5%, … +1%, +2%, … +6% of the peak frequency —> max. 13*15=195 F0 candidates per frame.

F0 estimation

A number of F0s is estimated from the candidate set by maximizing the following likelihood function:

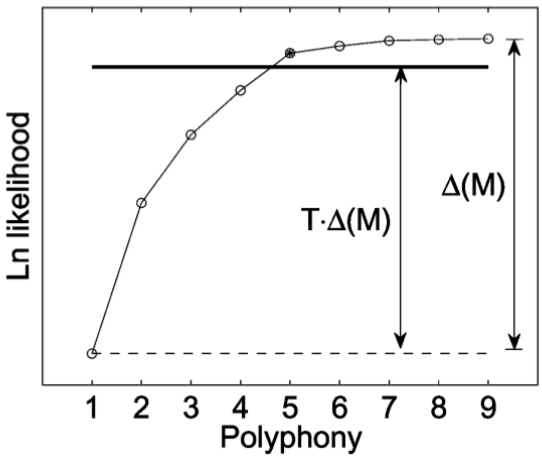
$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}\left(\mathbf{O}|\boldsymbol{\theta}\right)$$

It describes the likelihood of the spectrum (including its peaks and non-peak areas) being observed given a set of F0s. This likelihood function can be split into two cases: The likelihood of the detected peaks occurring and the likelihood of no peaks occurring in the non-peak regions.

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_{\text{peak region}}(\boldsymbol{\theta}) \cdot \mathcal{L}_{\text{non-peak region}}(\boldsymbol{\theta}).$$

The parameters of both functions have been estimated from training data and are provided in a malab matrix. In L(peakRegion), both spurious and non-spurious peaks are considered. The functions for the observed peaks depend on the distance to the closest match as one of the harmonics of the F0, as well as the peak magnitude in relation to the harmonic number. The non-peak region likelihood refers to the probability of an harmonic not being detected. This case occurs when the energy of the harmonic is too low to be detected as a peak. The corresponding probability depends on the harmonic number as well as the F0 itself.

In the implementation, the likelihood is first maximized for a single F0, then iteratively, the number of F0s, corresponding to the number of sources, is increased. Given the chosen model, the overall likelihood will probably increase, as additional sources are added. This phenomenon corresponds to maximum likelihood overfitting: “The larger the polyphony, the higher the peak likelihood…”. The algorithm therefor estimates the number of sources from the increase of the likelihood when iteratively adding sources:



The threshold T is learned from data.

Post-processing

Similar to the case predominant melody extraction, a frame-wise multiple-F0 estimation is not reliable. Therefore, temporal outliers are detected and replaced with interpolated values: