

YIN, a fundamental frequency estimator for speech and music, by Cheveigne & Kawahara

Fundamental frequency of a periodic signal, F_0 , is the inverse of its period, and it can be used in many applications in different areas, such as: speech recognition, automatic score transcription, and real-time interactive systems as music applications, or as metadata for multimedia content indexing. Of course, all of these are with the assumption that correct estimation is possible. F_0 's definition may vary depending the applications, e.g. for voiced speech it is the rate of vibration of the vocal folds.

This paper presents a relatively simple algorithm to estimate fundamental frequency of speech and music signals in which the error rate is almost third of the well-known methods. The algorithm can be implemented with a low latency and few parameters. It involves autocorrelation and cancellation. It may be extended in several ways to handle several forms of aperiodicity that happen in some applications.

The methods include six consecutive steps, in which they try to reduce the error in each step. The steps are: autocorrelation, difference function, cumulative mean normalized difference function, absolute threshold, parabolic interpolation, and best local estimate. For the first few steps we assume that the period is a multiple of the sampling rate.

Autocorrelation method is one of the methods used in which the signal is compared to a shifted version of the same signal and it has too many errors for many applications.

Difference function is the second step applied to the signal that applying it to the signals causes the error rate drop to 1.95% from 10% for unbiased autocorrelation. Since amplitude changes causes the period-to-period dissimilarities to increase with a lag, the difference function won't suffer from making "too low" errors.

The third step, *cumulative mean normalized difference function*, tries to deal with "too high" errors. Because of the imperfect periodicity, the difference function of the example speech signal is zero at zero lag and non zero at the period. To improve the system the authors propose to replace the difference function in step 2 with the cumulative mean normalized difference function. To do so, they divide each value of the old function by its average over shorter-lag values. This division reduces the rate of "too high" errors to 1.69% (from 1.95%), and sets the upper frequency limit of the search range. It also normalizes the function for the next error-reduction step.

The fourth step, absolute threshold, is supposed to deal with "too low" errors. A common problem is that one of the higher-order dips of the difference function is deeper than the period dip and if it falls within the search range, result will be sub-harmonic. In this step an absolute threshold is set and the smallest value that takes the minimum difference deeper than the threshold. Choosing 0.1 as threshold helps the error rate to get reduced to 0.78%.

The fifth step is parabolic interpolation. Until the last step it was assumed that the period is a multiple of the sampling rate but if it is not, estimate can be off up to a half of the sampling period or cause a gross error. Parabolic interpolation is proposed to solve this problem. It had a low effect on the gross error rates (just 1% reduction) but it had a noticeable effect on fine errors at all F_0 s and removed gross errors at high F_0 while testing with synthetic stimuli.

Engineering Approaches to Music Perception & Cognition

Baharak Zali
8679928115

Homework # 5
Feb. 17, 2005

Review #1

Comparing the gross error rates for fundamental frequency estimation of this method and six other algorithms existing on the internet, as well as with four other models that were implemented locally (total of ten methods), all on four different databases showed that YIN has a significant advantage over the other 10 methods. The average error rate of YIN over the four databases was 1.03, while the closest error rate was 3.1 and belonged to additive algorithms and the highest average rate was 16.8. They also tested the effects of the window size, threshold, and low pass pre-filter cutoff frequency on the error rates separately and did not observe any critical dependency on these parameters, at least on the used sample databases.

The authors tried some error reduction methods to reduce the estimation errors which involve modifying the model to make less matches of the signal with unexpected set of parameters. The extensions were concerned about variable amplitude, variable fundamental frequency, and additive noise in four different forms: slowly varying DC, periodic, different spectrum from target, and same spectrum as the target. None of these extensions improved error rates, probably because the periodic model that YIN used was accurate enough for this task.

Most of the formal experiments are done on voice and speech databases and no formal evaluation is done on the performance of YIN on music databases but this technique seems to be appropriate for music also. There are several reasons for lack of this evaluation and they include lack of a well-labeled database, wide-range of music styles and instruments and etc.

The significant portion of the idea and work was suggested even decades ago but integrating those in a system, adding the last step, the analysis of why the system works, and the formal evaluation were the new sections.