F_0 estimation of one or several voices

Alain de Cheveigné, Alexis Baskind

Ircam - CNRS, 1
place Igor Stravinsky, 75004, Paris, France.
{cheveign,baskind}@ircam.fr

Abstract

A methodology is presented for fundamental frequency estimation of one or more voices. The signal is modeled as the sum of one or more periodic signals, and the parameters estimated by search with interpolation. Accurate, reliable estimates are obtained for each frame without tracking or continuity constraints, and without the use of specific instrument models (although their use might further boost performance). In formal evaluation over a large database of speech, the single-voice algorithm outperformed the best competing methods by a factor of three.

1. Introduction

The fundamental frequency (F_0) of a musical sound predicts the pitch that it evokes, and for speech it carries prosodic (and in some cases segmental) information. It is of direct use in applications such as speech processing, music transcription and multimedia information retrieval, and serves also indirectly in algorithms for spectral estimation or coding. However, the limited reliability of common estimation algorithms is a severe obstacle for many applications.

Some applications involve several simultaneous voices (concurrent speakers or polyphony). Estimating several F_0s from a single signal is hard. Even if a single estimate is required, the competing voices interfere with estimation. An isolated voice (homophony) may pose similar problems in the presence of reverberation, as each new note is accompanied by the decay of the previous note. The monumental review of Hess [5] lists hundreds of single-voice algorithms, and many more have been proposed since. Comparatively fewer algorithms have been proposed for multiple voices, but their numbers are already considerable (e.g. [7,8], for reviews see [1,2]).

This paper presents methods for single voice F_0 estimation (YIN), and multiple voice F_0 estimation (MMM), based on a common methodology. The ideas are partly new, partly the systematization of older ideas (in particular autocorrelation). Formal evaluation shows that the single-voice algorithm YIN performs much better than competing methods for single voice estimation. The multiple-voice version awaits formal evaluation.

2. Single voice: YIN

2.1. F_0 estimation

The signal x_t is modeled as a periodic function with period T, by definition invariant for a time shift of T: for all t, $x_t - x_{t-T} = 0$. The period is the smallest positive number T for which this holds true. Conversely, if T is unknown it may be found by forming the difference function:

$$d_t(\tau) = (1/W)\sum_{j=1}^{W}(x_j - x_{j-\tau})^2 \quad (1)$$

and taking the smallest positive value of τ for which the function is zero. Thus described, the method is similar to the AMDF method [5,6] that takes the absolute value rather than the square. Developing the squared difference gives:

$$d_t(\tau) = r_t(0) + r_{t-\tau}(0) - 2r_t(\tau) \quad (2)$$

where $r_t(\tau)$ is the running autocorrelation (AC) function:

$$r_t(\tau) = (1/W)\sum_{j=1}^{W} x_j x_{j-\tau} \quad (3)$$

The first two terms in Eq. 2 are short-term power estimates sampled at times t and $t-\tau$. To the degree that they are constant w.r.t. τ, functions d and r are opposites one of the other: to a zero of one corresponds a peak of the other. Note that the definition of Eq. 3 differs with that of the familiar short-term AC function that uses $W-j$ instead of W in the summation limit. The method thus seems similar to the autocorrelation method, of which there are many variants. It differs however in several details, enumerated here.

First, the second power term in Eq. 2 is *not* constant (especially for small W), so the behaviors of d and r are not quite equivalent. Of the two, d is to be preferred as it fits the periodic signal model. Second, d supports the definition of a "cumulative mean-normalized" difference function:

$$d'_t(\tau) = \begin{cases} 1 & \text{if } \tau = 0 \\ d_t(\tau)/[(1/\tau)\sum_{j=1}^{\tau} d_t(j)] & \text{otherwise} \end{cases} \quad (4)$$

This new function is a better substrate than d for period estimation. At zero lag it equals 1 rather than 0, and doesn't approach 0 before the period. This avoids "too high" F_0 estimation errors that arise if the algorithm incorrectly chooses the zero-lag dip, and removes the need to set an upper limit on F_0 estimates. It also normalizes the function w.r.t. amplitude, an important prerequisite for the next step. In this third step, a threshold θ is fixed, and among the set of minima of $d'(\tau)$ that fall below it, the one with the smallest τ is taken as the period estimate.

The threshold θ is the main parameter of the algorithm. It sets a limit on the ratio of aperiodic to total power allowable within a "periodic" signal. Its value is not too critical ($\theta = 0.1$ works well in most cases). Parabolic interpolation is used to counter the effects of limited sampling resolution.

2.2. Periodic/aperiodic partition

Suppose a signal x_t and T an estimate of its period. The signal can be expressed as a sum of two signals $a_t = (x_t - x_{t-T})/2$ and $b_t = (x_t + x_{t-T})/2$. We have $b = x$ if x is purely periodic, whereas a is zero unless x is *not* perfectly periodic. In this

	gross error (%)	
method	gross error	(low / high)
pda	**16.8**	(14.2 / 2.6)
fxac	**15.2**	(14.2 / 1.0)
fxcep	**6.0**	(5.0 / 1.0)
ac	**5.1**	(4.1 / 1.0)
cc	**4.5**	(3.4 / 1.1)
shs	**8.7**	(8.6 / 0.18)
acf	**5.0**	(0.23 / 4.8)
nacf	**4.8**	(0.16 / 4.7)
additive	**3.1**	(2.5 / 0.55)
TEMPO	**3.4**	(0.53 / 2.9)
YIN	**1.03**	(0.37 / 0.66)

Table 1: Gross error rates for YIN (bottom line) and several other methods. Note that the new method produces much fewer errors than competitors. See [4] for further details.

sense a and b are "periodic" and "aperiodic" parts, respectively. Denoting:

$$||x_t||^2 = (1/W) \sum_{j=t+1}^{t+W} x_j^2 \tag{5}$$

it is easy to verify that:

$$(||x_t||^2 + ||x_{t-T}||^2)/2 = ||a_t||^2 + ||b_t||^2 \tag{6}$$

The left hand is a reasonable estimate of the signal power, while terms on the right are powers of the aperiodic and periodic parts. The partition is thus also a partition of signal *power*. An overall "measure of aperiodicity" (relative to T) can be defined as:

$$A_T = ||a_t||^2/(||a_t||^2 + ||b_t||^2) \tag{7}$$

Its value is 0 for a periodic signal, 0.5 for white noise, and 1 for an "antiperiodic signal" (a signal such as $x_t = -x_{t-T}$). This measure can be used as an ingredient of a "voicing detection" algorithm (a task that is not addressed here).

Parseval's relation implies a similar partition of *power spectra*: the power spectrum of x is the sum of the power spectra of a and b. Thus, the decomposition gives a convenient access to the "periodic" and "aperiodic" spectra (Fig. 1, top). [Note that other definitions of periodic and aperiodic parts may be preferable for certain applications. This definition offers optimal temporal resolution, and generalizes well to several periods as explained below.]

2.3. Evaluation

The YIN method was evaluated over a large database of speech (about 1.9 hours of speech of 48 speakers in four languages). Results are summarized in Table 1. The method was evaluated also with singing voice [6]. A feature useful for music is that the F_0 range can be made as wide as desired without degrading performance. Practical limits are set by the sampling rate (estimation is unreliable beyond $f_s/4$) and computation cost (inverse to the square of the lower limit). Frequency resolution is unlimited (thanks to parabolic interpolation) and time resolution is optimal (one can't go below $2T$). No tracking is involved, so estimates are purely local in time. The algorithm is simple and may be implemented efficiently.

3. Two voices: MMM

3.1. F_0 estimation

The waveform z_t is modeled as the sum of two periodic signals: x_t with period T, and y_t with period U. The sum is not necessarily periodic, but we have for all t:

$$z_t - z_{t-T} - z_{t-U} + z_{t-T-U} = 0 \tag{8}$$

Conversely, if the periods T and U are unknown they may be found by forming the difference function:

$$d_t(\tau, \upsilon) = \sum_{j=t+1}^{t+W} (z_t - z_{t-\tau} - z_{t-\upsilon} + z_{t-\tau-\upsilon})^2 \tag{9}$$

and taking the smallest values of (τ, υ) for which the function is zero. If x and y are periodic the algorithm is *guaranteed* to succeed, except in a number of pathological cases where the problem is undetermined. This occurs when all components of the compound are multiples of the same fundamental within the search range (as when one period is multiple of the other).

A concern is computation cost, as Eq. 9 must be evaluated for all (τ, υ). To save computation, the square may be expanded and expressed as a function of autocorrelation terms:

$$\begin{aligned} d_t(\tau,\upsilon) \quad = \quad & r_t(0) + r_{t-\tau}(0) + r_{t-\nu}(0) + r_{t-\tau-\upsilon}(0) \\ & -2r_t(\tau) - 2r_t(\upsilon) + 2r_t(\tau+\upsilon) \\ & +2r_{t-\tau}(\upsilon-\tau) - 2r_{t-\tau}(\upsilon) - 2r_{t-\upsilon}(\tau) \end{aligned}$$

Supposing that all useful values of $r_t(\tau)$ have been precalculated, evaluation of each step entails only 5 sums (if terms are rearranged) instead of about $4W$ sums and W squares with Eq. 9. A similar formula can be expressed using difference functions rather than autocorrelation functions.

Another concern is sampling resolution. Limited sampling resolution implies imperfect cancellation, and thus poor resolution. This is addressed by quadratic interpolation in the vicinity of each local minimum of the $d(\tau, \upsilon)$ surface.

Zeros of Eq. 9 occur along both axes $\tau = 0$ and $\upsilon = 0$. To exclude them (without setting an upper limit on the F_0 search limit), the function is modified in two steps:

$$d'_t(\tau,\upsilon) = \begin{cases} 1 & \text{if } \tau = 0 \text{ or } \upsilon = 0 \\ d_t(\tau,\upsilon)/[(1/\tau)\sum_{j=1}^{\tau} d_t(j,\upsilon)] & \text{otherwise} \end{cases}$$

$$d''_t(\tau,\upsilon) = \begin{cases} 1 & \text{if } \tau = 0 \text{ or } \upsilon = 0 \\ d'_t(\tau,\upsilon)/[(1/\upsilon)\sum_{j=1}^{\tau} d'_t(\tau,j)] & \text{otherwise} \end{cases}$$

This new function is equal to *one* rather than zero along the axes. It is however zero at (U,V) *and all multiples* (kU,jV). In order to find (U,V) while avoiding multiples, a threshold θ is set, and among the minima that fall below it, the one with the smallest coordinates is chosen to give the period estimates. As discribed so far, the algorithm produces two estimates for each frame whatever the number of sources.

3.2. Estimating the number of sources

This section explains how to estimate the number of sources. Consider a signal z_t, and U and V two period estimates. z_t can be expressed as the sum of four signals:

$$\begin{aligned} c_t &= (z_t - z_{t-U} - z_{t-V} + z_{t-V-U})/4 \\ d_t &= (z_t + z_{t-U} - z_{t-V} - z_{t-V-U})/4 \\ e_t &= (z_t - z_{t-U} + z_{t-V} - z_{t-V-U})/4 \\ f_t &= (z_t + z_{t-U} + z_{t-V} + z_{t-V-U})/4 \end{aligned} \tag{10}$$

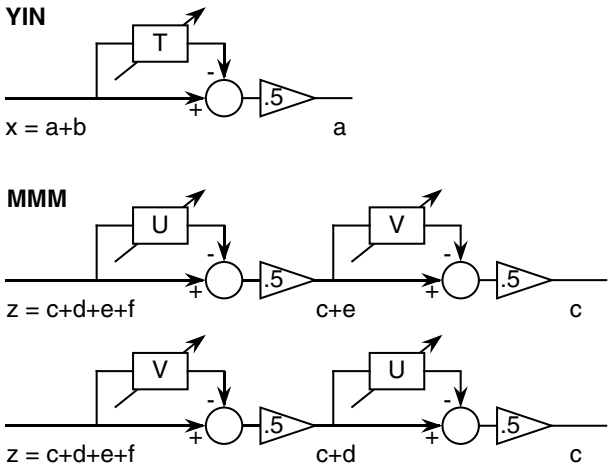

Figure 1: Top: principle of the periodic/aperiodic partition used by YIN. The signal x can be seen as the sum of a *periodic* part (b) and an *aperiodic part* (a). Their powers sum up to the power of x. On the basis of the ratio of powers of a to x, one can decide if the signal is periodic or not. Bottom: similar partition generalized to two periods (MMM). The signal z can be seen as the sum of four signals. Their powers sum up to that of z. On the basis of the reduction in power obtained by filtering out period U, then period V (or V, then U), it is possible to estimate the number of periods contained in the composite signal.

It is easy to verify that they sum up to z_t, and that their powers sum up to the following estimate of total signal power:

$$(||z_t||^2 + ||z_{t-U}||^2 + ||z_{t-V}||^2 + ||z_{t-U-V}||^2)/4 \qquad (11)$$

On their basis, we can define several "partial periodicity measures":

$$\begin{aligned} A_{UV} &= ||c_t||^2/(||c_t||^2 + ||d_t||^2 + ||e_t||^2 + ||f_t||) \\ A_U^V &= ||c_t||^2/(||c_t||^2 + ||d_t||^2) \\ A_V^U &= ||c_t||^2/(||c_t||^2 + ||e_t||^2) \end{aligned} \qquad (12)$$

All measures share the same numerator: the amount of power left over after canceling both U and V. The denominator of the first is total power (as measured by Eq. 11). A_{UV} thus reflects the relative "error" that remains after modeling the signal as a sum of two periodic sounds. The denominator of the second is the power that remains after canceling only U. A_U^V thus measures the *improvement* to the fit when the second period V is added to the single-period (U-only) model. A_V^U likewise measures the improvement when U is added to the V-only model. All measures vary between 0 and 1.

On the basis of these measures we can propose the following algorithm for determining the number of periods. It involves three threshold parameters, θ_1, θ_2, θ_3:

```
Treat x as single-period, estimate T and A_T.
Treat as two-period, estimate U, V, A_U^V, A_V^U, A_UV.
If A_T < θ1
   if T ≠ U and T ≠ V
      --> one period:  T
   else if V = T
        if A_V^U < θ2
           --> two periods:  (T, U)
        else
           --> one period:  T
   else if U = T
        if A_U^V < θ2
           --> two periods:  (T, V)
        else
           --> one period:  T
else
   if A_UV < θ3
      --> two periods:  (U, V)
   else
      --> more than 2 periods (or aperiodic).
```

To paraphrase, T is reported if the single-period model gives a good fit ($A_T < \theta_1$). In that case a second period (e.g. U) may also be reported if the two-period model gives a *better* fit (e.g. $A_V^U < \theta_2$), but only if T is among the solutions of this two-period model (e.g. $T = V$). Otherwise U and V are assumed to be harmonics of T and the two-period model is rejected. If the single-period model gives a *bad* fit, and the two-period model a *good* fit ($A_{UV} < \theta_3$), the algorithm reports (U,V). If neither model gives a good fit, the signal is assumed to contain more than two periodic signals.

3.3. Dealing with amplitude variation

The algorithm works by joint cancellation of both signals. If either is imperfectly periodic, cancellation is imperfect and both estimates may be compromised. One common form of aperiodicity, amplitude variation (as in piano or guitar notes), can be handled by modeling the observed waveform z_t as the sum of two signals x_t and y_t that are "periodic with variable amplitude", i.e. such that for all t, $x_t - \alpha x_{t-T} = 0$ and $y_t - \beta y_{t-U} = 0$. The sum z_t then obeys:

$$z_t - \alpha z_{t-T} - \beta z_{t-U} + \alpha\beta z_{t-T-U} = 0, \qquad \forall t \qquad (13)$$

If the periods T and U are unknown they may be found (given α and β) by forming the difference function:

$$d_t(\tau, \upsilon) = \sum_{j=t+1}^{t+W} (z_t - \alpha z_{t-\tau} - \beta z_{t-\upsilon} + \alpha\beta z_{t-\tau-\upsilon})^2 \qquad (14)$$

and choosing the smallest values of (τ,υ) for which it is zero. As before, it is computationally useful to expand Eq. 14 as a function of autocorrelation terms. Conversely, if (T, U) are known, (α, β) can be found by setting derivatives of $d_t(\tau, \upsilon)$ with respect to these parameters to zero:

$$\begin{aligned} \alpha &= [r_t(\tau) - \beta r_t(\tau + \upsilon) + \beta^2 r_{t-\upsilon}(\tau)]/[r_{t-\tau}(0) \\ &\quad + \beta^2 r_{t-\tau-\upsilon}(0) - 2\beta r_{t-\tau}(\upsilon)] \\ \beta &= [r_t(\upsilon) - \alpha r_t(\tau + \upsilon) + \alpha^2 r_{t-\tau}(\upsilon)]/[r_{t-\upsilon}(0) \\ &\quad + \alpha^2 r_{t-\tau-\upsilon}(0) - 2\alpha r_{t-\upsilon}(\tau)] \end{aligned} \qquad (15)$$

Estimation of α requires knowledge of β and vice-versa, but both can be found by applying Eqs. 15 iteratively. Each step reduces d which is bounded from below by 0 so this procedure must converge. If none of (T, U, α, β) are known a priori, the algorithm can alternate between estimating (T, U) and (α, β).

3.4. Evaluation

Informal evaluation shows good performance on mixtures of relatively "clean" speech or musical instrument sounds, even when F_0s are close, spectral envelopes are similar, and amplitudes are mismatched. As for most methods, performance degrades rapidly in the presence of noise or aperiodicity (one important case was treated in the previous section). Formal evaluation is being performed and will be reported soon.

4. More than two sources

By construction, MMM may be extended to N periodic sources, failing only if all partials happen to be multiples of $N-1$ or fewer frequencies within the search range. For $N > 2$ the method becomes more sensitive to aperiodicity and noise, and is also computationally expensive.

5. Conclusion

This paper presented methods to estimate one period (YIN) or more (MMM) from music or speech sounds. YIN demonstrates performance well beyond that of the best competing methods, showing that the basic approach is sound and competitive with spectral methods. MMM has not yet been formally evaluated, but informally it appears to work well. Both methods are derived directly from periodic (or sum of periodic) signal models, and are thus well grounded in principle. Both operate frame-by-frame, without using continuity constraints or voice or instrument models. For single voices such constraints do not seem to be necessary (performance is excellent), but to handle two voices or more it is likely that they will be important for the best performance (e.g. [8]). We believe that our time-domain methodology offers a good alternative to the usual spectral domain approach for applying such constraints.

6. References

[1] de Cheveigné, A. (1993). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. 93, 3271-3290.

[2] de Cheveigné, A., and Kawahara, H. (1999). "Multiple period estimation and pitch perception model," Speech Communication 27, 175-185.

[3] de Cheveigné, A., and Henrich, N. (2002). "Fundamental frequency estimation of musical sounds," J. Acoust. Soc. Am. 111, 2416 (Abstract).

[4] de Cheveigné, A., and Kawahara, H. (2002). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. 111, 1917-1930.

[5] Hess, W. (1983). "Pitch determination of speech signals," Berlin, Springer-Verlag.

[6] Ross, M. J., Shaffer, H. L., Cohen, A., Freudberg, R., and Manley, H. J. (1974). "Average magnitude difference function pitch extractor," IEEE Trans. ASSP 22, 353-362.

[7] Wu, M., Wang, D., and Brown, G. J. (2002). "A multi-pitch tracking algorithm for noisy speech," Proc. IEEE ICASSP, 369-372.

[8] Walmsley, P. J., Godsill, S. J., and Rayner, P. J. W. (1999). "Polyphonic pitch tracking using joint bayesian estimation of multiple frame parameters," Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.