Fundamental Frequency Estimation by Least-Squares Harmonic Model Fitting

András Bánhalmi, Kornél Kovács, András Kocsor, László Tóth

Research Group on Artificial Intelligence
of the Hungarian Academy of Sciences and University of Szeged, Hungary
{banhalmi,kkornel,kocsor,tothl}@inf.u-szeged.hu

10.21437/Interspeech.2005-170

Abstract

This paper proposes a pitch estimation algorithm that is based on optimal harmonic model fitting. The algorithm operates directly on the time-domain signal and has a relatively simple mathematical background. To increase its efficiency and accuracy, the algorithm is applied in combination with an autocorrelation-based initialization phase. For testing purposes we compare its performance on pitch-annotated corpora with several conventional time-domain pitch estimation algorithms, and also with a recently proposed one. The results show that even the autocorrelation-based first phase significantly outperforms the traditional methods, and also slightly the recently proposed `yin` algorithm. After applying the second phase – the harmonic approximation step – the amount of errors can be further reduced by about 20% relative to the error obtained in the first phase.

1. Introduction

Pitch estimation is one of the classic speech processing problems that still attracts research. The methods proposed during the decades can be roughly categorized into two main groups [3]. The group of mathematically inspired algorithms assume that the processed signal (a 30-50 ms speech segment) is quasi-periodic, and hence try to find the fundamental frequency of a periodic signal via signal processing techniques. These solutions are either time-domain – that is they work directly with the signal itself – or frequency-domain, which means that they look for periodicity information in a spectral representation of the signal. The other large group is that of the biologically inspired models, namely sophisticated auditory models that try to imitate human hearing and, as part of it, human pitch sensation. The method presented here is a mathematically inspired one. This is why we prefer using the term 'fundamental frequency' to 'pitch', although one may argue that in the case of speech signals the two terms are practically interchangeable. More precisely, our algorithm belongs to the category of time-domain methods, as it directly approximates the signal by a sum of harmonic sinusoids. For better results this approximation is preceded by a preprocessing step where a traditional autocorrelation-based estimate is calculated to initialize the harmonic approximation step. In the next section the harmonic approximation algorithm itself is explained first, followed in Section 3 by how it is combined with the autocorrelation-based method to get a better performance. In Section 4 the algorithm is compared with several conventional algorithms and one recently proposed algorithm by evaluating them on two pitch-annotated databases. Then we draw some conclusions about the efficacy of the proposed method in Section 5.

2. Least-Squares Harmonic Approximation

The well-known harmonics plus noise representation estimates the signal under analysis as a sum of time-varying sinusoids plus a filtered noise component [7]. In the special case when the modelled signal is speech, one may have the usual assumption that its short, 30-40 ms intervals can be regarded quasi-stationary [5]. Hence, when estimating the model parameters in such time steps, one can expect that the voiced parts of the signal can be very closely approximated by the sinusoidal components only, while the noise component will play an important role in modelling the voiceless segments. In the case of healthy speech one can have the further simplifying assumption that the sinusoidal components are harmonic, that is their frequencies are multiples of some fundamental frequency ω.

In the following we shall make an estimate of the fundamental frequency of a short speech excerpt; the signal will be given in the form of N signal samples $\mathbf{s} = (s_1, \ldots, s_N)^T$ taken at the discrete time instances $\mathbf{t} = (t_1, \ldots, t_N)^T$. We will approximate this signal by the sinusoidal model

$$h(t) = a_0 + \sum_{k=1}^{L} a_k \cos(f_k t + \psi_k). \tag{1}$$

In general case, fitting the model to the signal requires the estimation of the number of components L and also the frequencies f_k, the amplitudes a_k and the phases ψ_k for each component.

A reasonable way of obtaining proper parameters is by applying the least squares method, which minimizes the squared error between the original signal and its approximation:

$$\epsilon = \sum_{i=1}^{N} W_{ii}^2 (s_i - h(t_i))^2 \to \min, \tag{2}$$

where the diagonal matrix W is simply a weight matrix like a hamming window.

Without placing any restriction on the number of components or on their parameters, minimizing the error function of Eq. (2) poses a difficult global optimization problem. Just fixing the number of components L, the optimization still leads to a homogeneous function optimization problem that can be only handled by sophisticated optimization algorithms like the approaches proposed by Starer [10] or Kocsor et al. [4]. However, the task becomes significantly simpler with the further assumption that the components are harmonic – that is their frequencies are the multiples of fundamental frequency ω – so that $f_k = k\omega$. Then our approximation becomes harmonic:

$$h(t) = a_0 + \sum_{k=1}^{L} a_k \cos(k\omega t + \psi_k), \tag{3}$$

so there is only one frequency parameter (ω) to be estimated. Moreover, for a given ω the number of harmonics L is also given, as it can be simply calculated from ω and the sampling frequency of the signal. Most importantly, for a given ω the remaining two type of parameters, the amplitude and phase values can be determined by solving a simple linear equation system. The calculation of these parameters at a given ω value is as follows.

Making use of the trigonometric identity $\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$, it can be shown that Eq. (3) may be expressed in vector form. That is, $h(t)$ may be rewritten as

$$h(t) = \mathbf{b}_1^T(t)\mathbf{u} - \mathbf{b}_2^T(t)\mathbf{v},$$

where

$$\begin{aligned} \mathbf{b}_1^T(t) &= (1, \cos(1\omega t), \ldots, \cos(L\omega t)) \\ \mathbf{u}^T &= (a_0, a_1\cos\psi_1, \ldots, a_L\cos\psi_L) \\ \mathbf{b}_2^T(t) &= (\sin(1\omega t), \ldots, \sin(L\omega t)) \\ \mathbf{v}^T &= (a_1\sin\psi_1, \ldots, a_L\sin\psi_L) \end{aligned}$$

With this notation the error function of Eq. (2) takes the form

$$\epsilon = \|W(\mathbf{s} - B\mathbf{f})\|_2^2, \tag{4}$$

where $\mathbf{f}^T = (\mathbf{u}^T\ \mathbf{v}^T)$ and $B = (B_1\ B_2)$, its components being:

$$B_i^T = (\mathbf{b}_i(t_1), \ldots, \mathbf{b}_i(t_N)) \qquad i = 1, 2.$$

According to Eq. (4), when the fundamental frequency ω is fixed, then the error function takes a quadratic form, and its global optimum can be found by solving the linear equation system defined by the normal equation:

$$(B^T W^T W B)\,\mathbf{f} = B^T W^T W\,\mathbf{s},$$

or, equivalently, by calculating $\mathbf{f} = (WB)^+ W\mathbf{s}$, where $()^+$ denotes the Moore & Penrose pseudo inverse.

Having obtained $\mathbf{f}$, the amplitude and phase parameters of each harmonic component can be calculated via the formulas

$$\psi_k = \arctan\frac{u_{k+1}}{v_k}, \qquad a_k = \frac{v_k}{\sin\psi_k}.$$

3. Harmonic Approximation for Pitch Estimation

As was shown in the previous section, for a given fundamental frequency ω the rest of the parameters can be easily calculated. The optimal approximation for ω itself can be found by a search, that is by evaluating the error function at several different ω values. This does not restrict the practical applicability of the method, as in practice we can limit the possible value of the fundamental frequency to a reasonably small interval, and we usually need it only to within a certain resolution. The simplest solution is to partition the frequency interval assumed to contain the pitch into tiny intervals depending on the required resolution, and evaluate the error function at each of these. An evaluation of the error at one ω value is reasonably fast, hence this process will not cause an unmanageable computational burden. Still, there are a couple of useful observations that can be exploited to speed up the computation. For this we should rearrange the formula for the error function like so

$$\epsilon = \|W\left(\mathbf{s} - B\mathbf{f}\right)\|_2^2 = \|\left(W - (WB)(WB)^+W\right)\mathbf{s}\|_2^2,$$

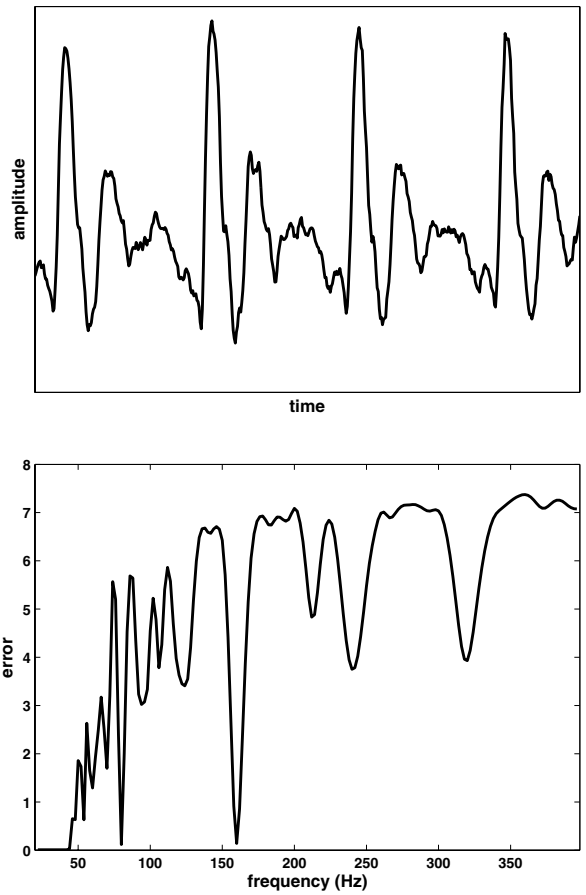

Figure 1: *Upper: a voiced speech excerpt from a male voice sampled at 22050Hz. Lower: the approximation error as a function of the assumed fundamental frequency.*

where we exploited the fact that $\mathbf{f} = (WB)^+W\mathbf{s}$. As one can see, the error at a given frequency can be calculated as a product of the signal vector s and the matrix $Z = \left(W - (WB)(WB)^+W\right)$. The first consequence of this is that, during the search for the optimal fundamental frequency estimate, the amplitude and phase parameters of the components do not have to be computed directly. The second is that when we have to compute the estimate for many speech frames, the Z matrix can be pre-computed and stored, thus speeding up the processing of the signal.

Having calculated the error at various ω values to a certain resolution, we obtain a curve like that shown on the lower part of Fig. 1. Theoretically, the point where this curve takes its minimum can be returned as our fundamental frequency estimate. However, in practice we observe that at low enough frequencies – that is below 50-60 Hz – the error goes down to practically zero, thus fooling a process that would simply pick the minimum value. This shows that the least-squares error minimization procedure has to be extended by properly chosen pre- and/or post-processing steps. Many possible methods could be proposed for this, and we chose the following simple solution: we took the output of a fast and simple pitch estimation algorithm as a first estimate, and evaluated our least squares algorithm only in a given small interval around this estimate. Besides speed improvements, this resulted in a refined, more pre-

cise value compared to the original one.

According to this, the method we developed consists of two phases. The estimator of the first phase must be fast enough and should commit octave errors relatively rarely. With this in mind, we applied a modified version of the well-known autocorrelation function (`acf`) method [9]. As the first step, we calculate the normalized autocorrelation coefficients of signal x at time t, the size of the time window being w:

$$r_t^{norm}(\tau) = \frac{\sum_{j=t}^{t+w-1} x_j x_{j+\tau}}{\left(\sum_{j=t}^{t+w-1} x_j x_j\right)\left(\sum_{j=t}^{t+w-1} x_{j+\tau} x_{j+\tau}\right)}. \quad (5)$$

Then we find that local minimum of the correlation function where the corresponding τ is minimal and also the corresponding correlation coefficient value is above some predefined constant. If the function has no such point, then we use the estimate of the previous frame as the local estimate. In the following this relatively fast and simple method for fundamental frequency estimation will be referred to as `cwt` (correlation with threshold).

In the second phase of processing, we refine the estimates of the `cwt` algorithm by running the least-squares harmonic approximation algorithm, evaluating it only in a certain interval and resolution around the estimate obtained from the `cwt` step. The algorithm that includes both the `cwt` and the subsequent harmonic approximation step will be referred to as `cwt-hap` later on. The `cwt` and `cwt-hap` algorithms were applied with the following parameters during the tests:

- **`cwt:`** The threshold on the correlation strength was set to 0.63 and the window size was 512 samples. The resulting estimates were median-smoothed with their left and right neighbor.
- **`cwt-hap:`** The correlation threshold was set to 0.63 and the windows size was set to 512. As for the parameters of the harmonic approximation step, the weighting window was a rectangular one. The frequency resolution was 2 Hertz, and the neighborhood in which we searched for a refinement over the `cwt` estimate was [-20Hz; 20Hz]. The resulting estimates were again median-smoothed with their left and right neighbor. Furthermore, to reduce the run time, the number of sinusoidal components were not calculated from the sampling rate and the fundamental frequency ω, as suggested earlier, but we limited the highest possible frequency to 5000 Hz (instead of half the sampling rate).

4. Experiments and Results

4.1. Details of Evaluation

The best way to compare pitch estimation algorithms is to evaluate them over a dedicated database that contains precise and verified pitch estimates. Fortunately, there are such databases available. The usual way to create reliable reference pitch estimates for them is to record a laryngograph signal in parallel with the speech signal, because pitch estimation is much easier than that for the former. In the databases we used, these estimates were further corrected manually, and so-called 'mask' data was also given, specifying where the signal is voiced – the pitch estimators were only tested on these signal segments.

The evaluation itself was as follows. When evaluating the methods, values that differed by more than 20% from laryngograph-derived estimates were counted as 'gross errors'. This criterion is used in many studies, and an error of this magnitude can be expected to be significantly reduced by a refinement algorithm such as our harmonic approximation method. Naturally, we took care to run all the algorithms under similar conditions: for example, the search interval for the fundamental frequency was always the same. The exact parameters for each algorithm will be specified later on.

4.2. Databases Used

For testing we applied the following, freely accessible pitch databases:

The Keele Pitch Database. It contains the recordings of 10 speakers, five male and five female, reading a phonetically balanced text, the 'North Wind story' for a total of 0.15 hours of speech [6]. For further details see the URL "http://www.liv.ac.uk/Psychology/HMP/projects/pitch.html".

The Edinburgh fda Evaluation Database. It contains the recordings of one male and one female speaker, each speaking 50 English sentences for a total of 0.12 hours of speech, for the purpose of evaluating F_0-estimation algorithms [1]. The database can be downloaded from the URL "http://www.cstr.ed.ac.uk/ pcb/fda_eval.tar.gz".

4.3. Reference Methods

During the tests we compared our method with the following algorithms, which are mostly traditional and well-known. For each algorithm we will give a brief description and the parameter settings that were used. In each case the window size was 512 samples, the windows shift was 100 samples, and the interval in which we searched for the fundamental frequency estimate was [60; 400] Hz (corresponding to a $[\tau_{min}; \tau_{max}]$ limit on the period, depending on the sampling rate). All the codes applied were our own implementation, apart from the `yin` algorithm that was downloaded from the web.

- **`acf(1):`** It will denote the standard autocorrelation coefficient based method that computes the coefficients

$$r_t(\tau) = \sum_{j=t}^{t+w-1} x_j x_{j+\tau} \quad (6)$$

for a given signal x and time index t, then converts these values to the weighted coefficients [9]

$$r'_t(\tau) = \begin{cases} r_t(\tau)(1 - \tau/\tau_{\max}), \text{if} \tau \leq \tau_{\max} \\ 0, \text{otherwise} \end{cases}, \quad (7)$$

the size of the time window being w. The resulting pitch estimate is the frequency value corresponding to that τ value where the function $r'_t(\tau)$ takes its global maximum.

- **`acf(2):`** This denotes a modified version of the previous method that calculates the following coefficients [9]:

$$r''_t(\tau) = \sum_{j=t}^{t+w-\tau-1} x_j x_{j+\tau}. \quad (8)$$

- **`amdf:`** The amdf estimator evaluates the following function for each possible τ:

$$\mathrm{d_t}(\tau) = \sum_{j=t}^{t+w-1} |x_j - x_{j+\tau}|, \quad (9)$$

then the pitch estimate obtained as the frequency value corresponding to that τ value where the function $d_t(\tau)$ takes its global minimum [8].

- **`nacf:`** This algorithm calculates a set of normalized autocorrelation coefficients, according to the formula

$$n_t(\tau) = r_t(0)[1 - r_t(\tau)^2 / (r_t(0) r_{t+\tau}(0))]. \quad (10)$$

The pitch estimate is again obtained as the frequency value corresponding to that τ period where the function $n_t(\tau)$ takes its global minimum.

- **`yin:`** this is a fundamental frequency estimator algorithm that consists of several processing phases [2]. We applied the original implementation of the inventor during the tests. The corresponding Matlab code can be downloaded from "http://www.ircam.fr/pcm/cheveign/sw/yin.zip".

4.4. Results and Discussion

The results are summarized in Table 1. They show that the conventional algorithms make gross errors relatively frequently, and our experience indicates that these are mostly octave errors. `yin` and the two algorithms proposed here can significantly reduce the number of errors; we observed that the remaining ones are mostly committed at the boundaries of voiced and voiceless sections. We were somewhat surprised to see that our `cwt` algorithm that was introduced only to initialize the harmonic approximation algorithm already outperformed the recently proposed `yin` method. Its error rate was decreased further by the harmonic approximation step, resulting in a further 20% relative reduction in the error rate.

method	DB1	DB2
acf(1)	9.0%	14.1%
acf(2)	4.8%	11.1%
amdf	7.0%	8.6%
nacf	7.6%	10.5%
yin	3.2%	4.4%
cwt	2.9%	4.0%
cwt-hap	2.3%	3.2%

Table 1: *Gross errors of the different methods on databases DB1 and DB2.*

5. Conclusions and Future Work

This paper proposed a pitch estimation algorithm that is based on optimal harmonic model fitting. The algorithm operates directly on the time-domain signal and has a relatively simple mathematical background. To increase its efficiency and accuracy, the algorithms was combined with an autocorrelation-based initialization phase. The results indicate that even the first phase outperforms the conventional methods and is slightly better than the results for the `yin` algorithm. The harmonic approximation (second phase) brings about a further improvement as well.

Looking at the behavior of the algorithm more closely, we can say that the significance of the first phase turned out to be much higher than we had initially expected. Besides reducing the processing time, which was the main motivation for its introduction, it actually attained such a good performance that the harmonic approximation phase improved its results at a lesser rate than we had originally hoped. The other reason for this is that just evaluating the error of the harmonic model in a small interval around the result of the first phase can correct only certain kind of errors. We think that a significant further improvement might be obtained by refining the strategy of how we calculate the pitch estimates from the error function values. For example, wider intervals could be examined if we had proper heuristics to exclude false optima. This is what we intend to look into in the near future.

6. References

[1] Bagshaw, P. C., Hiller, S. M. and Jack, M. A., Enchanced pitch tracking and the processing of F0 contours for computer and intonation teaching, Proc. Eurospeech, pp. 1003-1006, 1993.

[2] de Cheveigne, A., Kawahara, H., YIN a fundamental frequency estimator for speech and music, Acoustical Society of America:697–708, 2002.

[3] Hess, W., Pitch Determination of Speech Signals, Springer, 1983.

[4] Kocsor, A., Tóth, L., Bálint I., On the Optimal Parameters of a Sinusoidal Representation of Signals, Acta Cybernetica 14, pp. 315-330, 1999.

[5] Oppenheim, A. V. and Shafer, R. W., Discrete-Time Signal Processing, Prentice-Hall, Englewood Cliffs, New York, 1989.

[6] Plante, F., Meyer, G. F. and Ainsworth, W. A., A pitch extraction reference database, Proc. Eurospeech, pp. 837-840, 1995.

[7] Quatieri, T. F. and McAulay, R. J., "Audio Signal Processing Based on Sinusodial Analysis/Synthesis", in: Applications of Digital Signal Processing to Audio and Acoustics (ed. Kahrs, M. and Brandenburg, K.), Kluwer, 1998

[8] Ross, M. J., Shaffer, H. L., Cohen, A., Freudberg, R., Manley, H. J., Average magnitude difference function pitch extractor, IEEE Trans. Acoust., Speech, Signal Process. 22, 353-362, 1974.

[9] Rabiner, L. R., Schafer, R. W., Digital Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, 1978.

[10] Starer, D., Nehorai, A., Newton Algorithms for Conditional and Unconditional Maximum Likelihood Estimation of the Parameters of Exponential Signals in Noise, IEEE Trans. Sig. Proc., Vol. 40, No. 6, 1992.