The Multi-Lag-Window Method for Robust Extended-Range F_0 Determination

Edouard GEOFFROIS *
geoffroi@etca.fr

DGA/DRET/ETCA
16 bis av. Prieur de la Côte d'Or
94114 Arcueil cedex France
http://www.etca.fr/

ABSTRACT

This paper addresses the problem of the fundamental frequency (F_0) determination of a speech signal, and proposes four improvements to conventional frequency-domain methods. The major improvement is a multi-scale analysis which extends the range of F_0 that can be correctly processed. It builds on the lag-window method proposed by Sagayama (1978), hence the name "multi-lag-window". Secondly, a modification of the lag-window method itself improves its robustness to periodic noises (while loosing its gain-independence property). Thirdly, a rescaling is introduced to permit a full Dynamic Programming search for the optimal F_0 curve. Finally, a mathematically justified peak interpolation is proposed for replacing the conventional, inaccurate parabolic interpolation. These four improvements result in an accurate, robust, extended-range F_0 determination method, which was tested on spontaneous speech from 20 speakers, ranging from less than 50 Hz to more than 600 Hz.

1. INTRODUCTION

Fundamental frequency (F_0), and more generally prosody, is an important aspect of speech. It conveys information on the speaker's attitude, on the syntactic structure of the utterance, on word choice in tone languages, etc. High expectations about its usefulness for automatic speech recognition has motivated ongoing research activity for many years. However, its use in actual systems remains very limited.

One of the bottlenecks to its use is the robust F_0 estimation from a speech signal. It is indeed known to be a difficult task, and none of the many existing Pitch Determination Algorithms (PDAs) are completely satisfactory [3]. One of the main limitations is the restricted range of F_0 values correctly processed. Another is the lack of robustness to periodic noise, due to the gain-independence property of conventional PDAs. The post-processing part can also be improved. The usual way Dynamic Programming (DP) is applied to select an F_0 curve is sub-optimal, and the often used parabolic interpolation scheme is inaccurate. This paper addresses these limitations in turn, and proposes solutions to each of them.

*This research was done while the author was a PhD. student at LIMSI/CNRS

2. FOUR IMPROVEMENTS TO F_0 DETERMINATION

2.1. Extended-Range F_0 Determination using Multi-Scale Analysis

PDAs work within a given range of F_0 values, outside of which tend to occur the so-called "pitch halving" and "pitch doubling" errors. Pitch halving is tightly linked to the very nature of the task: if a signal is periodic, then twice the fundamental period is also a period. This problem is generally implicitly dealt with by windowing at some stage of the processing, but this increases the risk of pitch doubling. For short-term analysis PDAs, the range of correctly processed F_0 values is thus essentially determined by the size of the analysis window.

To extend this range, the basic idea is to combine the results of several analyses, each parameterized for a limited interval of F_0 values but altogether spanning a wider interval. For example, suppose that two different settings are used for male and female voices. For any given input, the appropriate setting is expected to yield a periodicity function[1] presenting a peak corresponding to the F_0 of the signal. Combining the periodicity functions obtained with the two settings can be done by simply summing them. The resulting function will present a peak wherever any of the initial functions presented one. This can be generalized to three settings (Figure 1) or more.

This idea assumes that no spurious peak occurs when the setting is not suited to the F_0 of the analyzed signal. This is not the case for most PDAs such as the modified autocorrelation (SIFT) or cepstrum methods. However, there exists a PDA for which this assumption holds, called the lag-window method [4]. In a nutshell, it consists in smoothing the power spectrum to get its envelope, and dividing the the spectrum by the envelope to get the harmonic fine structure. Taking the inverse Fourier transform then yields the periodicity function. The smoothing is actually performed in the lag domain by multiplication with a window, called the lag-window.

[1] There is no agreed upon generic term to designate the function presenting peaks corresponding to the periodicity of the analyzed signal and on which peak picking is performed to estimate F_0 and the degree of voicing, whatever the method used to compute it. We will refer to such a function as a "*periodicity function*".

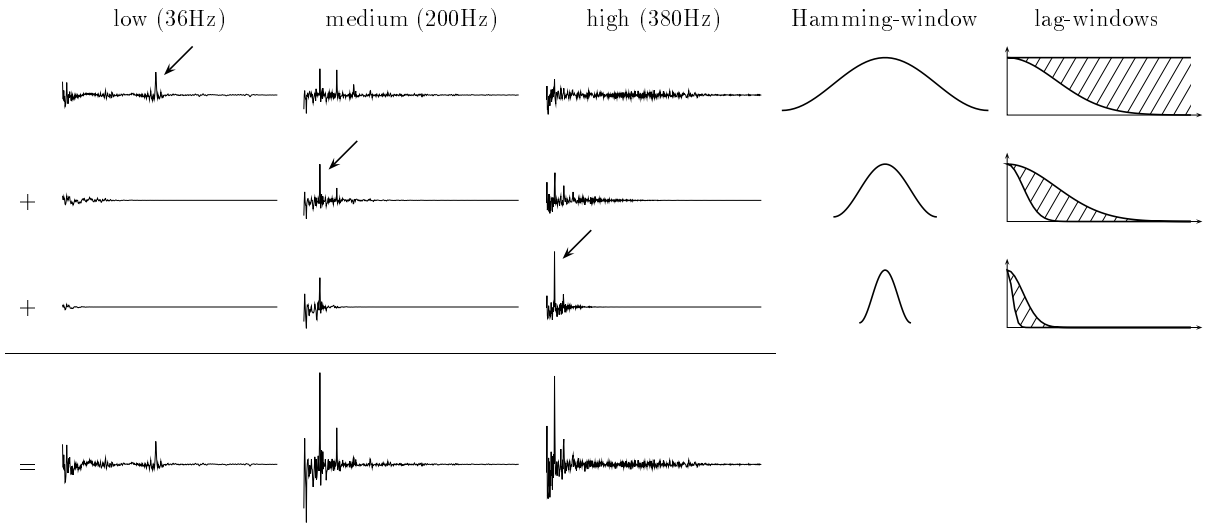


Figure 1: Multi-window F_0 analysis. Periodicity functions are computed with different window sizes (Hamming and lag-windows shown on the right). They present a peak if the scale factor of the windows is suitable for the value of F_0 (as pointed by the arrow), but only weak spurious peaks if not. Therefore, the summed periodicity functions (below) presents a correct peak in all cases.

The method involves two lag-windows[2] which can be adjusted to eliminate spurious peaks if F_0 is outside of the optimal range determined by the size of the analysis window. Viewing the spectrum as a signal and the lag-windows as frequency responses of filters may help explain why these windows can "filter out" frequencies outside the chosen range.

For a multi-scale analysis involving n different settings, the sizes of the n analysis Hamming windows were specified, and the n pairs of lag-windows were automatically chosen so that (1) each pair is centered on the lag corresponding to the optimal F_0 for this analysis, and (2) the upper lag-window of a pair was equal to the lower lag-window of the next pair, i.e., there are $n+1$ different lag-windows in total which form a partition of the lag-domain (excluding a small part for very short lags).

Each elementary lag-window PDA introduces a variable gain depending on F_0. It can be shown that this gain is [2]:

$$2 \times \frac{\text{upper-lag-window}(1/F_0) - \text{lower-lag-window}(1/F_0)}{1 - \text{lower-lag-window}^2(1/F_0)}.$$

The total gain for several settings introduces a distortion which cannot be neglected. It is therefore necessary to normalize the summed periodicity function by dividing it by the sum of the above expression for all settings.

To summarize, the basic algorithm can be applied at different scales and the results combined to yield the correct estimate on a wide range of F_0. In particular, both male and female voices could be processed with a single parameter set. Let us now see how the basic algorithm itself can be modified to improve its robustness to pollution by periodic signals.

[2]To our knowledge, the few publications in English giving explanations on the lag-window method mention only the version with a single lag-window, used to compute the spectral envelope [6]. In the original publication, a version involving a second lag-window, used for smoothing also the spectrum itself before dividing it by its envelope, was also presented, and argued to be especially useful for high pitch voices [4].

2.2. Robustness to Periodic Noise

The gain-independence of most PDAs becomes a problem in presence of a low-intensity but highly periodic signal such as a computer fan noise. Indeed, the PDA considers such a signal alone as strongly voiced, even though it may not be audible at normal level. The resulting high peak in the periodicity function can be particularly damaging when constraints are put on the F_0 curve (as in section 2.3). Even for speech segments, pollution by a periodic noise can be a problem if the speech power spectrum has a lower energy than the noise in some frequency band, as this perturbs the PDA.

Introducing a non-linearity for low-amplitude samples in the spectrum limits the perturbations caused by the noise. In the cepstrum method, this can be achieved by modifying the logarithm (for example by replacing $\log(x)$ by $\log(x + x_{\min})$). With the lag-window method, this can be obtained by thresholding the spectral envelope before the division. For high energy frequency bands, the result of the division is unchanged, and for low energy frequency bands it tends to zero rather than being dominated by the noise. In practice, the simple thresholding can be replaced by a smoother function (for example $y = \sqrt{x^2 + x_{\min}^2}$ instead of $y = \max(x, x_{\min})$) to ensure a more continuous transition between the limiting cases. The value of $x_{\min}$ should be approximately equal to the average of the spectrum of the lowest intensity audible sound.

This minor modification proved to be very effective against the periodic noise present in our database. Of course, it does not affect the processing of harmonics of normal amplitude. For further improving robustness, post-processing is necessary.

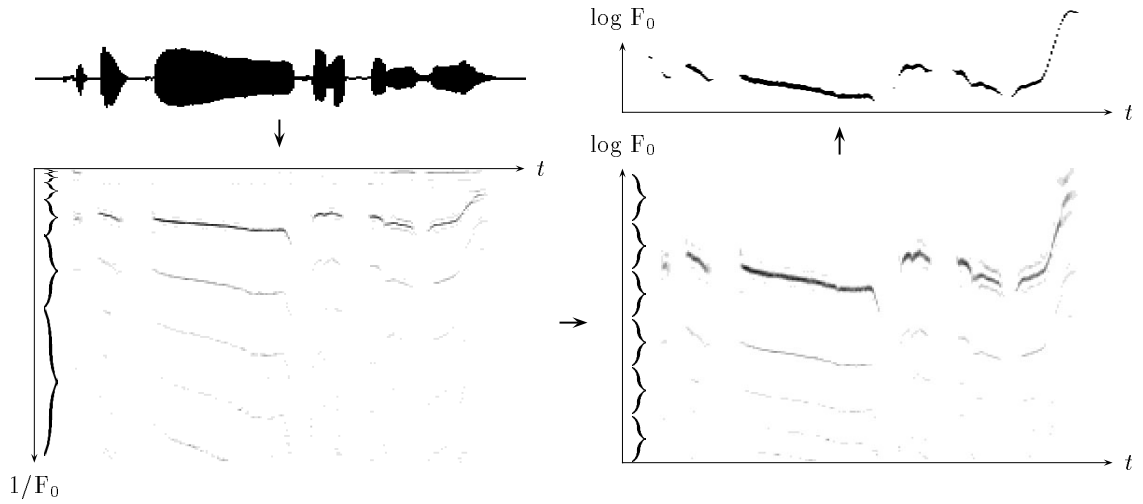


Figure 2: Rescaling of a periodicity diagram facilitates application of DP. On the natural scale of frequency-domain PDA outputs, constraints on the F_0 slope varies with the frequency band. Using a logarithmic scale renders these constraints uniform.

2.3. Optimal decoding of an F_0 contour

Post-processing is an important part of a PDA [3]. It can consist in correction or smoothing after a frame-by-frame extraction of F_0 values, or, preferably, in a DP search putting constraints on transitions between consecutive frames of the F_0 contour. Even when dynamic programming is used to search for an F_0 curve, however, a few peaks per frame are usually preselected, and $\log(F_0)$ is computed for each candidate peak [5]. The essential reason is that authorized transitions span a variable number of samples depending on the F_0 band. Furthermore, Voiced/Unvoiced (V/UV) decisions must be performed simultaneously with the DP, since for unvoiced parts the selected peaks are erratic.

A prior rescaling of the periodicity diagram[3] lag-axis into a log-F_0 scale (or any auditory scale) allows straightforward application of a conventional DP algorithm to the whole diagram (Figure 2). Indeed, after rescaling, the pattern of authorized transitions is the same on the whole range of F_0 values. Furthermore, penalties on authorized but large F_0 jumps can be very easily added into the DP process. Constraints and penalties on the second derivative were also implemented, but their usefulness was not evaluated separately.

The computation of $\log(F_0)$ for all lags need be done only once, for computing the new scale. Then the diagram can be rescaled efficiently using, for example, quadratic splines.

Processing all the data guarantees that the estimated F_0 contour is optimal in the sense of the scoring function. This function is defined as the sum of the "voicing measure" (the amplitude of the traversed sample of the periodicity function) along the contour, plus the penalties on the first and second derivatives.

[3]The "*periodicity diagram*" is the time-frequency representation of F_0 obtained by putting together periodicity function for successive frames. In other words, the periodicity diagram is to the periodicity function what the spectrogram is to the spectrum.

The DP algorithm does not have to commit itself to a V/UV decision. It automatically interpolates between voiced segments. For each frame, the voicing measure is output along with the estimated F_0 value. If necessary, a voicing threshold can be chosen later on to select voiced frames.

The voicing measure can also be used for displaying the F_0 contour. By using a line thickness proportional to the voicing measure, strongly voiced parts appear as bold lines while weakly voiced frames, where extraction errors are more frequent, immediately appear as less reliable, and V/UV transitions appear as lines fading away (Figure 2).

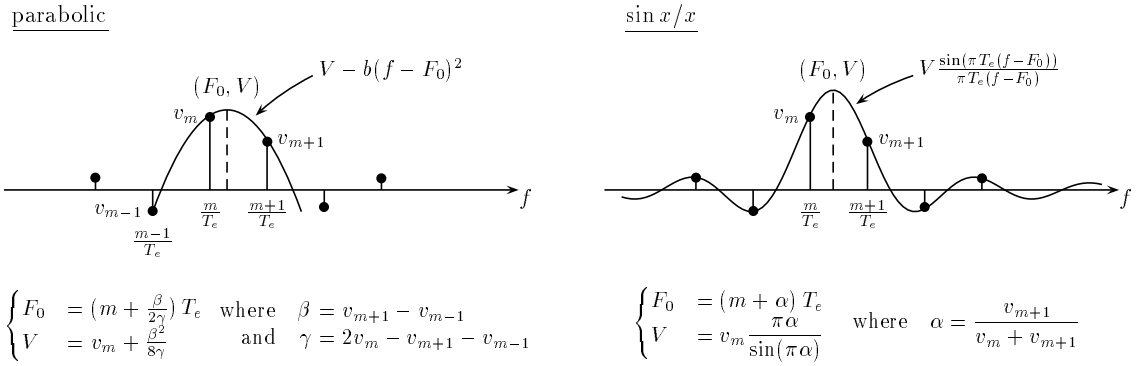
The first three improvements increase the robustness of the algorithm. But accuracy can also be improved, as discussed in the following section.

2.4. Accurate Peak Interpolation

In frequency-domain PDAs, the value of F_0 is estimated from the position of the peak in the periodicity function. The simplest method uses the position of the largest sample. As sampling can be sparse, especially for short lags i.e. for high pitch values, interpolating is often used to increase the accuracy. In the absence of further information, a three-point parabolic interpolation taking into account the two neighboring samples seems reasonable, and is the usual interpolation scheme.

However, we do have further information. We know that the periodicity function is the inverse Fourier transform of another function which is expected to be close to a sine function, the frequency of which we wish to measure. Therefore, the mathematically justified interpolation uses a $\sin x/x$ function [1].

As this $\sin x/x$ function seems more complex than a parabola, the equations could be expected to be difficult to solve, if at all tractable.



$$\begin{cases} F_0 &= (m + \frac{\beta}{2\gamma})\, T_e \\ V &= v_m + \frac{\beta^2}{8\gamma} \end{cases} \quad \text{where} \quad \beta = v_{m+1} - v_{m-1} \quad \text{and} \quad \gamma = 2v_m - v_{m+1} - v_{m-1}$$

$$\begin{cases} F_0 &= (m + \alpha)\, T_e \\ V &= v_m \dfrac{\pi\alpha}{\sin(\pi\alpha)} \end{cases} \quad \text{where} \quad \alpha = \frac{v_{m+1}}{v_m + v_{m+1}}$$

Figure 3: Accurate Peak Interpolation. The usual method is parabolic interpolation. Interpolating by a $\sin x/x$ function is mathematically justified, and also results in simple formulas. The figure illustrates that the parabolic interpolation tends to underestimate the peak height.

Surprisingly enough, the calculation proves to be straightforward [2], and leads to very simple formulas reproduced on Figure 3.

These formulas imply the prior selection of the maximum between v_{m+1} and v_{m-1}. This would not be harmful if the periodicity function was guaranteed to be of the form $\sin x/x$, as v_{m+1} and v_{m-1} are then equal to each other only if they are equal to zero. But in the general case, it is better to consider both values, and use them as corrective terms insofar as they are positive, as follows:

$$\begin{cases} F_0 &= (m + \alpha_1 + \alpha_{-1})\, T_e \\ V &= v_m \dfrac{\pi\alpha_1}{\sin(\pi\alpha_1)} \dfrac{\pi\alpha_{-1}}{\sin(\pi\alpha_{-1})} \end{cases}$$

$$\text{where} \quad \alpha_{\pm 1} = \begin{cases} \dfrac{v_{m\pm 1}}{v_m + v_{m\pm 1}} & \text{if } v_{m\pm 1} > 0 \\ 0 & \text{if not} \end{cases}$$

As can be seen from Figure 3, the accuracy improvement is more important for the voicing measure V than for the value of F_0.

3. EVALUATION

F_0 estimation was carried out on a 1-hour corpus of spontaneous speech recorded from 20 speakers (10 m/10 f) spanning a range of over 3 octaves (from less than 50 Hz to more than 600 Hz). F_0 was first estimated with the initial version of the PDA, i.e. with the lag-window method, using different settings for male and female speakers. F_0 contours were manually verified and regions where gross errors occured were marked. During manual verification it was noticed that the output was degraded for speakers recorded in the presence of a periodic noise. After each modification of the algorithm, the marked regions as well as regions where F_0 changed significantly from the previous stage were reverified to determine whether errors had been corrected or new errors were introduced.

The final algorithm allowed correct pitch extraction using a single setting (with three analysis window sizes) on the entire corpus, and with a drastically reduced sensitivity to periodic noise. The improvements were confirmed by the overall increases in performance of an automatic prosodic labeling system developed simultaneously [2].

4. CONCLUSION

The following four ideas to improve F_0 determination were proposed:

1. Combining analyses with different window sizes extends the range over which F_0 can be correctly processed.
2. Introducing a non-linearity for small samples in the spectrum increases the robustness to periodic noise.
3. Rescaling allows straightforward application of a conventional DP algorithm to the whole time-frequency representation and therefore optimal decoding of the F_0 contour.
4. Interpolation by $\sin x/x$ is mathematically justified, leads to simpler formulas, and results in smoother contours and better estimates of the degree of voicing.

These ideas were implemented, tested on spontaneous speech from 20 speakers, and demonstrated an ability to extract pitch accurately in a wide range of conditions.

5. REFERENCES

1. E. Oran Brigham. *The Fast Fourier Transform*, pages 102–105. Prentice Hall, 1974.
2. Edouard Geoffrois. *Extraction robuste de paramtres prosodiques pour la reconnaissance de la parole.* PhD thesis, Universit Paris XI, Orsay, 1995.
3. Wolfgang Hess. *Pitch Determination of Speech Signals — Algorithms and Devices*. Springer-Verlag, 1983.
4. Shigeki Sagayama and Sadaoki Furui. Pitch extraction using the lag-window method. In *IEICE Meeting*, pages 5–263, 1978. (in Japanese).
5. Bruce G. Secrest and George R. Doddington. An integrated pitch tracking algorithm for speech systems. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1352–1355, 1983.
6. Hiroshi Shimodaira and Mitsuru Nakai. Robust pitch detection by narrow band spectrum analysis. In *International Conference on Spoken Language Processing (ICSLP)*, pages 1597–1600, 1992.

Sound File References:

[a605s01.wav]