A fundamental frequency estimator by crosscorrelation of adjacent spectra

Philippe Martin

EA333, ARP, UFR Linguistique
University Paris-Diderot, Paris, France
philippe.martin@linguist.jussieu.fr

Abstract

All fundamental frequency estimators based on spectral analysis rely heavily on a proper harmonic selection of the voice analyzed. Since in practice other spectral peaks pertaining to different sources are present in the signal, various schemes have been design to ensure a satisfactory elimination of non pertinent harmonics.

This paper introduces a new harmonic selection algorithm based on harmonic alignment of adjacent frames. This new approach ensures better pitch tracking, and gives satisfactory results even in the presence of another speech source. The algorithm has been implemented and incorporated in the speech analysis program WinPitch, and tested on various examples provided in the Speech Separation Challenge project.

1. Introduction

Despite their somewhat limited resolution in time, due to the inherent window duration requirements of the Fourier transform to obtain a reasonable frequency resolution to separate harmonics, speech fundamental frequency estimator based on spectral analysis gives generally more satisfactory results than time domain approaches, thanks to the use of most of the frequency information provided by the voice harmonics. The spectral comb, introduced in 1981 [1], [2], is one of these methods.

Although very robust in principle, as it ensures the correct detection of F0 even if some harmonics are missing in the spectrum (it actually only needs two consecutive harmonics to succeed), its implementation requires care in order to obtain satisfactory results. In particular, the frequency resolution of the spectrum evaluated for each time frame should be appropriate (which involves a rather large time window for low pitch voices), and a sufficient number of pertinent harmonics (i.e. related to the speech source analyzed) should be carefully selected, using for instance an intensity threshold to eliminate the presence of unwanted spectral noise components.

Although a human eye can readily separate harmonics of multiple sources signals by merely inspecting a narrow band spectrogram, it is remarkable that only of handful of spectral based Fo analysis methods do effectively use the property of continuity of harmonics from frame to frame [3] [4], except perhaps in a limited way by smoothing the raw Fo curve [5]. However, the commonly used dynamic programming approach to track harmonic continuity is very complex, as specific harmonics can appear and disappear from one frame to the next, depending on the signal to noise ratio retained in the process. In this paper, I will introduce another method to enhance the selection of pertinent harmonics used in the spectral comb method, based on the crosscorrelation of spectral components of sequential speech time frames.

2. Principles of operations

In theory, speech harmonics exhibits continuity in time: harmonics found in one given frame should be retrieved in surrounding frames belonging to the same voiced segment. In practice, as it can been easily seen on spectrograms, the intensity threshold used to define the noise level in the harmonic selection process may cause successive time spectra to contain reappearing or disappearing harmonics, depending on the formant pattern and the cut-off intensity level selected for a given frame. This may create difficult problems when a dynamic algorithm approach is used to restore the continuity of each harmonic, as the disappearance of a specific spectral component may involve a large number of frames.

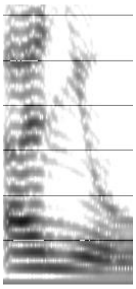


Figure 1: *Whimsical harmonics appearing and disappearing along the time scale*.

A totally different scheme is proposed here, using the obvious property of harmonics to be integer multiples of the fundamental. Excluding the case of fundamental frequencies of multiple speech sources closely linked as in choir signing for example, harmonics of different sources, including noise sources, evolve together proportionally from one frame to the next independently from the other sources. This property can be used to separate harmonics from multiple sources. The separation can be effectively done by aligning pertinent harmonics from one spectrum to the next by linearly changing the frequency scale of one of the spectra, and adding the intensity of aligned harmonics. The optimal shifting value is then obtained from the maximum value of the crosscorrelation between a reference spectrum and the frequency scale shifted spectrum of an adjacent frame. In this process, non pertinent harmonics don’t get aligned on the modified frequency scale, and the sum of their peaks will not be enhanced, whereas aligned pertinent harmonics will get reinforced by summing their peaks.

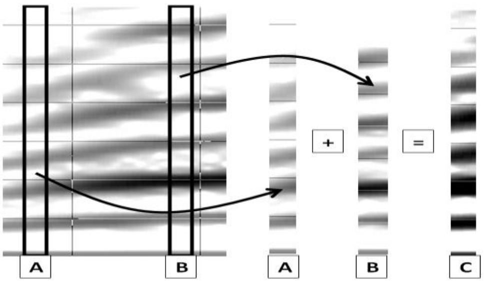


Figure 2: *Reinforcement of aligned harmonics: spectrum B has its frequency scale shifted to align its peaks with peaks of spectrum A.*

Detecting the maximum of a correlation function involving a given spectrum and a frequency scale shifted spectrum of a consecutive time frame appears to be a simple and elegant way to align harmonics of consecutive time frames. Indeed it is easy to see that the maximum of such a function will be reached when the harmonics of the dominant source are aligned, as the sum of the aligned spectral peaks will then be at a maximum. Values of the frequency scale shifts aligning unwanted harmonics will give a lower crosscorrelation value, as long as the overall sum of those harmonics is inferior to the expected dominant source harmonics.

Formally, a maximum of the crosscorrelation function is reached when harmonics of the dominant spectrum will be aligned, i.e. for a value of α giving a maximum of B(α) for two consecutive spectra F_t and F_{t+1} at time t and t+1:

$$B(\alpha) = \sum_{f=0}^{\alpha F_{Max}} Ft(f) F_{t+1}(\alpha f)$$

It is easy to notice that this process is insensitive to the appearance or disappearance of harmonics that would be caused by a change in the formant pattern from one frame to the next.

3. Implementation

3.1. Total alignment is NP complete

Extended to a complete voiced speech segment, alignment of the dominant source harmonics appears to be a NP complete problem. Assuming time frames are spaced by t = 20 ms, a voiced segment of duration d is divided into n = d/0.02 frames. Since the maximum rate of change of fundamental frequency in speech has been estimated at 1% per ms [6], a time spacing of 20 ms involves a maximum change of 20 % between consecutive frames. If each step used in the crosscorrelation function is chosen as 0.5 %, 40 steps are necessary to cover the -20% - +20% Fo range to evaluate the maximum. Therefore the brute force complete alignment process of the n frames requires s exp (n-1) evaluations of the correlation function, which, taking the values above would require 40 exp (50-1) operations for a 1 second voiced frame!

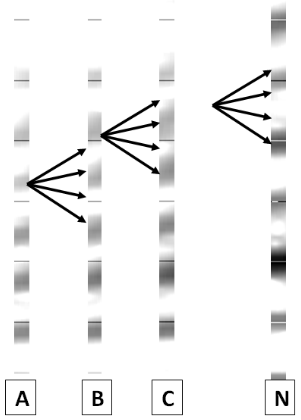


Figure 3: *Alignment of voiced segment harmonics is NP complete. The best alignment of N spectra has to be found among s exp (N-1) frequency scale shifts combinations (s is the number of shit values for each alignment)*

3.2. Suboptimal solutions

Possible suboptimal solutions do exist. A specific frame can be selected as reference (for instance a frame located in the middle of the voiced segment) and spectral alignment of adjacent frames aligned with the reference, operating sequentially from right to left until the boundaries of the voiced segment are reached. After each step, peaks of aligned harmonics are added, and non aligned harmonics peaks removed. After harmonic enhancement, the classical spectral comb method is applied to the resulting spectrum obtained for each frame.

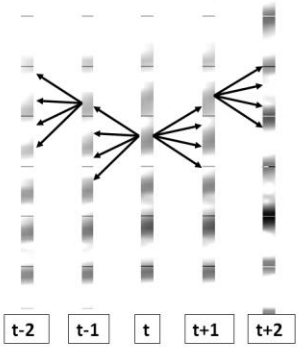
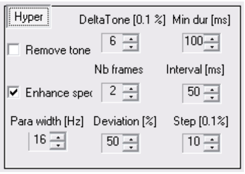


Figure 4: *Suboptimal alignment of alternate adjacent frames spectra: only 2 consecutive spectra before and after reinforced spectrum are considered, allowing a feasible solution.*

3.3. Implementation in WinPitch

The implementation done in the WinPitch software program [7] uses only a maximum of 5 spectrally aligned frames. Each frame has its pertinent harmonics enhanced after alignment of four equally time spaced spectra before and after it (Figure 4).

The time spacing, the number of shifts steps and the range of frequency scale shifts are programmable (Fig. 5), to allow an easy selection of the best set of shifting parameters for a given speech sample.


Hyper

DeltaTone [0.1 %]: 6

Min dur [ms]: 100

□ Remove tone

Nb frames: 2

Interval [ms]: 50

☑ Enhance spec

Para width [Hz]: 16

Deviation [%]: 50

Step [0.1%]: 10

Figure 5: *Adjustable parameters in WinPitch implementation of harmonic reinforcement by spectral alignment.*

4. Some experimental results

The following figures give some indications about the effectiveness of the harmonic reinforcement process. Test examples have been chosen among files available in the “Speech Separation Challenge” [8], with various intensity level differences between a two sources speech signal. “Enhanced harmonics spectral comb pitch tracking” refers to the spectral comb process operating on four aligned spectra.

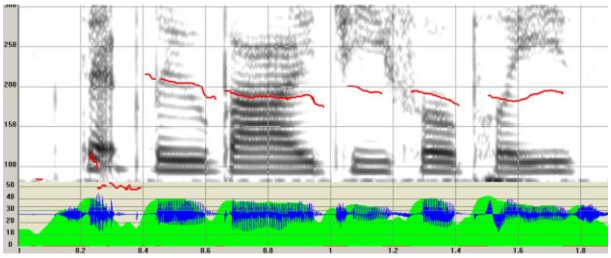


Figure 6: *Standard spectral comb pitch tracking (clean recording S11_SBBG4P.wav). The Fo curve is displayed in red.*

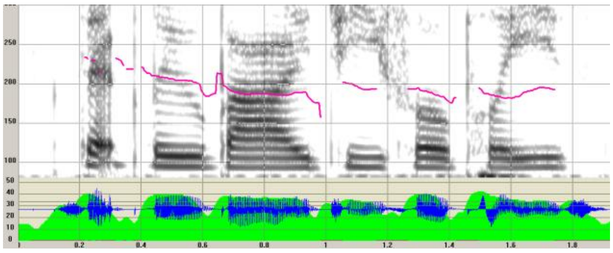


Figure 7: *Enhanced harmonics spectral comb pitch tracking (clean recording S11_SBBG4P.wav)*

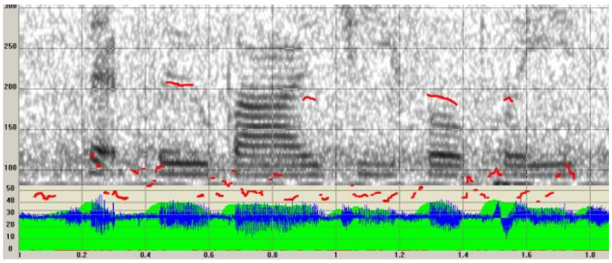


Figure 8: *Standard spectral comb pitch tracking (6 dB signal / noise S11_SBBG4P.wav)*

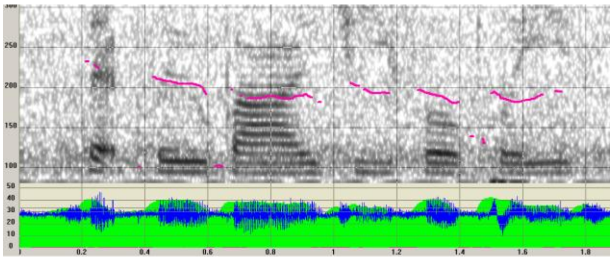


Figure 9: *Enhanced harmonics spectral comb pitch tracking, 2 frames, 50 ms interval (6 dB signal / noise S11_SBBG4P.wav)*

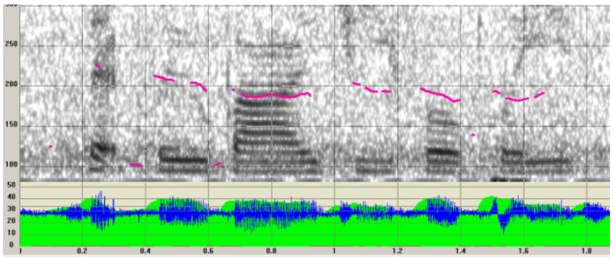


Figure 10: *Enhanced harmonics spectral comb pitch tracking, 4 frames, 50 ms interval (6 dB signal / noise S11_SBBG4P.wav)*

The comparison between fig. 9 and fig. 10 show now noticeable improvement between the use of 2 and 4 frames (positioned symmetrically around the target spectrum).

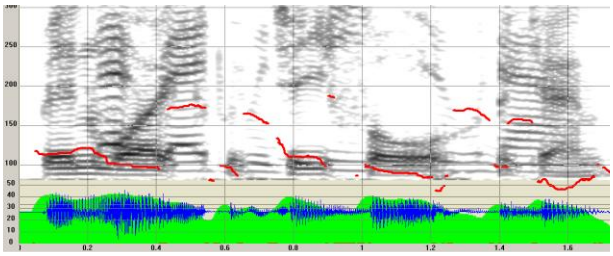


Figure 11: *Standard spectral comb pitch tracking (2 sources 6 dB difference T12_BWWG4A_M7.wav)*

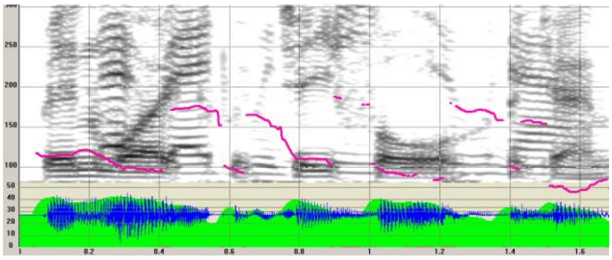


Figure 12: *Enhanced harmonics spectral comb pitch tracking, 4 frames, 50 ms interval (2 sources 6 dB difference T12_BWWG4A_M7.wav)*

Fig. 11 shows an example of multisource pitch tracking for a male and female voice with 6dB level difference. For each frame, the resulting pitch corresponds to the highest sum of peak value of the two spectra. The comparison of Fig. 11 and Fig. 12 shows a clear improvement of the overall pitch curve.

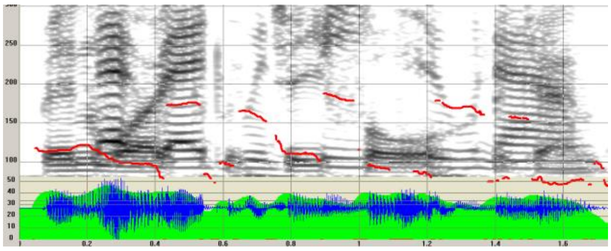


Figure 13: *Standard spectral comb pitch tracking, 4 frames, 50 ms interval (2 sources 0 dB difference T12_BWWG4A_M7.wav)*

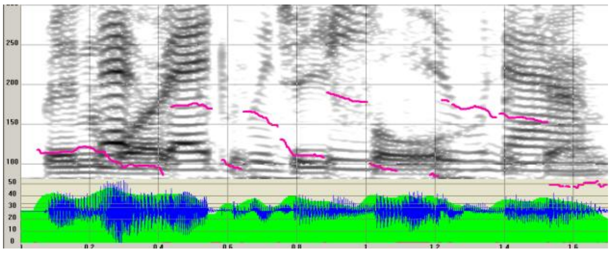


Figure 14: *Enhanced harmonics spectral comb pitch tracking, 4 frames, 50 ms interval (2 sources 0 dB difference T12_BWWG4A_M7.wav)*

Comparing Fig. 13 and Fig. 14, the pitch tracking improvement is only noticeable at the end of the sentence analyzed.

5. Conclusions

Enhancement of harmonics obviously give no improvement on pitch tracking in the case of clean recording devoted from any noise or other speech source. For a low single to noise ratio or in multisource cases, enhancing harmonics give a clear improvement in most cases. Further research is in development to implement a process which makes possible in practice a better use of the harmonic alignment for a whole voiced speech segment, by selecting an optimal set of alignment parameters.

6. References

[1] Martin, Ph. "Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne", 12e Journées d'Etude sur la Parole, SFA, Montréal, 1981.

[2] Martin, Ph. "Comparison of pitch detection by cepstrum and spectral comb analysis", IEEE-ICASSP, 180-183, 1982.

[3] Every, M.R. and Jackson, P.J.B., "Enhancement of harmonics content of speech based on a dynamic programming pitch tracking algorithm", ICLSP, 2006.

[4] Doval, B. and Rodet, X. “Fundamental Frequency Estimation and Tracking using Maximum Likehood Harmonic Matching and HMM”, Proc. IEEE-ICASSP 93, 221-224.

[5] Liénard, J-S., Signol, F. and Barras, C., “Speech fundamental frequency estimation using the Alternate Comb”, Interspeech 2007, Antwerp, FrC.O3-5.

[6] Kondoz, A.M., 1994, “Digital Speech”, Wiley, Chinchester.

[7] WinPitch, www.winpitch.com

[8] Cooke, M.P. and Lee, T.W. “Speech Separation Challenge”, http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm