Crosscorrelation of adjacent spectra enhances fundamental frequency tracking

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 sources external to the speech considered may be present in the signal, various schemes have been designed to ensure a satisfactory elimination of non pertinent harmonics.

This paper introduces a new harmonic selection algorithm based on the brush method, which ensures a very good selection of voice harmonics, 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.

Index Terms: speech analysis, fundamental frequency, speech separation

1. Introduction

Despite their somewhat limited resolution in time, due to the inherent properties of the Fourier transform to obtain a reasonable frequency resolution to separate harmonics, speech fundamental frequency estimation 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 [2], [3], is one of these methods, using a different approach than some better known algorithms as found in [1].

Although very robust in principle, as it ensures the correct detection of the fundamental frequency F0 even if some harmonics are missing in the signal, its implementation requires care in order to obtain satisfactory results. In particular, the frequency resolution of the spectrum evaluated for each time frame must be appropriate (which involves a rather large time window for low pitch voices), and a sufficient number of pertinent harmonics (i.e. effectively related to the speech source analyzed) should be carefully selected using for instance an intensity threshold to eliminate 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 F0 analysis methods do effectively use the property of continuity of harmonics from frame to frame [4] [5] [6], except perhaps in a limited way by smoothing the raw F0 curve at the last stage of the process. However, the commonly used dynamic programming and HMM to track harmonic continuity are very complex, as specific harmonics seem to appear and disappear from one frame to the next, depending on the clipping threshold above noise 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 exhibit continuity in time: harmonics found in one given frame should be retrieved in surrounding frames belonging to the same voiced segment (except of course at the beginning and end of voiced segments). In practice, as it can be easily seen on any spectrogram, 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 of their intensity for a given frame (Fig. 1). 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: Harmonics appearing and disappearing along the time scale.

In order to solve this difficulty, a totally different approach is proposed here, using the obvious property of harmonics to be integer multiples of the fundamental. Excluding the case where fundamental frequencies of multiple speech sources are closely linked as in choir singing for example, harmonics of different sources, including noise, evolve differently together proportionally from one frame to the next. This property can be used to separate dominant harmonics, whose sum of intensities is larger than the others. This can be effectively done by aligning pertinent harmonics from one spectrum to the next, modifying linearly the frequency scale of one of the spectra, and adding the intensity of aligned harmonics. In this process, non pertinent harmonics don’t get aligned on the frequency scale, and the sum of their peaks will not be enhanced, whereas aligned pertinent harmonics will get reinforced by summing their peaks (Fig. 2).
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 excepted dominant source harmonics.

A maximum of the crosscorrelation function is reached when harmonics of the dominant spectrum are aligned, i.e. for a value of $\alpha$ giving a maximum of a Brush function $B(\alpha)$ for two consecutive amplitude spectra $F_t$ and $F_{t+1}$ at time $t$ and $t+1$, $f$ being the frequency:

$$B(\alpha) = \sum_{f=0}^{f_{\text{max}}} F_t(f)F_{t+1}(\alpha f)$$

It is easy to notice that this process is insensitive to the appearance or disappearance of harmonics from one frame to the next (in particular those pushed outside the spectral frequency range due to a frequency shift value $\alpha > 1$). Indeed, missing harmonics will not modify the value of the maximum of $B(\alpha)$ as long as the sum of intensities of the aligned desired source harmonics is higher than the sum of intensities of the noise or other source components. When two speech sources are present in the signal, the process will switch from one speaker to the other when the above condition changes, for instance during an unvoiced segment occurring during a voiced segment of another source (see an example below).

### 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 $t = 50$ ms, a voiced segment of duration $d$ is divided into $n = d/0.05$ frames. Since the maximum rate of change of fundamental frequency in speech has been estimated at 1% per ms [7], a time spacing of 50 ms involves a maximum change of 50% between consecutive frames. If a 1% step is chosen to vary $\alpha$ in the evaluation of the crosscorrelation function, 100 steps are required to cover the -50% - +50% fundamental frequency (Fo) range to evaluate the maximum. Therefore the brute force complete alignment process of the $n$ frames requires $n$ evaluations of the correlation function, which, taking the values above would require $100^5$ correlations for a 300 ms voiced frame.

#### 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 alternating right and left frames 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 the harmonic enhancement process, the classical spectral comb method is applied to the resulting spectrum obtained for each frame.

### 3.3. Implementation in WinPitch

The implementation done in the WinPitch software program [8] uses only a maximum of 5 spectrally aligned frames. Each frame has its pertinent harmonics enhanced after alignment of 4 equally time spaced amplitude spectra before and after it (Fig. 4).
The time spacing, the number of shifts steps and the range of frequency scale shifts are programmable (Fig. 5).

![Adjustable parameters in WinPitch implementation of harmonic reinforcement by spectral alignment.](image)

As intensity values of aligned spectra are cumulatively added for each frame (the order of frames in the alignment process is alternate: next frame, preceding frame, second next, second preceding), the retained harmonics will correspond to the highest energy source in the segment.

### 3.4. A faster harmonic alignment

The maximum value of the crosscorrelation function is obtained when harmonics of the dominant source (i.e. whose sum of intensities is the largest) are aligned for a given value of the frequency scale shift coefficient $\alpha$. In practice, the maxima observed in the sequence of the $n/2$ amplitude values given by the Fourier transform of a properly windowed signal are obtained by parabolic interpolation, rejecting peaks with amplitude below a certain noise threshold. This ensures a better selectivity in the alignment process, given a typical width of the parabola reconstructed from the interpolated spectral peaks of 16 Hz at -6 dB.

Instead of shifting the frequency scale shift coefficient $\alpha$ by small steps of say 1%, a possibly faster equivalent result is reached if selected candidate harmonics of frame $t+1$ are directly aligned with a given harmonic of frame $t$. Considering the maximum shift 1%/ms of $F_0$ between two consecutive frames, harmonics candidate can be selected inside a frequency range given by $F_{0_{\text{min}}} = F_{0_{\text{ref}}} - \Delta f$, $F_{0_{\text{max}}} = F_{0_{\text{ref}}} + \Delta f$, $F_{0_{\text{ref}}}$ being the fundamental frequency of the reference harmonic, and $\Delta f$ the maximum excursion of $F_0$ between the two consecutive frames (Fig. 6).

![A more efficient way to obtain the best alignment operates by selecting all possible matching harmonics in the next frame $t+1$ in a frequency range determined by the maximal possible $F_0$ shift between the reference harmonic in frame $t$.](image)

For example, a reference harmonic peak of 150 Hz will be tentatively matched with harmonic peaks of a frame 50 ms away in a range of 150-(50) and 150+(50), i.e. 100 Hz and 200 Hz. The drawback of this process pertains to the possible lack of harmonics in a given range, which may impose the use of a sufficiently large set of spectral peaks to be considered in the reference spectrum, and a corresponding large range in the spectrum to align.

### 4. 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” [9], with various intensity level differences between a two sources speech signal.

![Pitch curve, clean recording, male voice, comb method, no smoothing (s14_bbae2a)](image)

The example of Fig. 7 and Fig. 8 (clean signal, no smoothing) show little difference between the comb and brush methods.

![Pitch curve, 6 dB SNR, male voice, comb method, no smoothing (s14_bbae2a)](image)

Fig. 9 and Fig. 10 show little improvement in harmonic continuity tracking by the brush algorithm over the comb method.
method, which operates frame by frame (recording with a 6 dB SNR, male voice). Indeed, random noise added in these examples is not eliminated or reduced by the harmonic based correlation process.

By contrast, examples mixing a male and a female voice given in Fig. 11 and Fig. 12 show a clear advantage of the brush method to select the dominant harmonics in speech separation.

---

**Fig. 11:** Pitch curve, male and female voices at 9 dB apart, comb method, no smoothing (s14_bwasw2a_m25lbwisa)

**Fig. 12:** Pitch curve, male and female voices at 9 dB apart, brush method, no smoothing (s14_bwasw2a_m25lbwisa)

**Fig. 13:** Comparison of pitch curves obtained from a spectrogram (top), with Praat (autocorrelation) (mid) and WinPitch (Brush, 5 frames of 50ms) (bottom) for a male and a female speaker with 0dB intensity interval (standard smoothing used in both cases). Sections circled show the most important differences between the two methods (recording t14_bwas2a_m25_lbwi4n.wav).

---

5. Conclusions

Enhancement of harmonics obviously give no improvement on pitch tracking for clean recordings devoted from any noise or other speech source. For a low single to noise ratio or in multisource cases however, enhancing harmonics through crosscorrelation of successive amplitude spectra results in a clear improvement in most cases. Compared to methods (e.g. autocorrelation) used in popular speech analysis software such as Praat [10], the brush process gives better results in multiple speech source cases, even when a suboptimal process (5 frames) is used. Readers can readily experiment the effectiveness of the algorithm using the WinPitch pitch analyzer [8].

A faster harmonic step alignment described above is currently being implemented in order to find the optimal harmonic alignment for a lengthy voiced speech segment.

6. References


