On the relationship between instantaneous frequency and pitch in speech signals

Zied Mnasri
Electrical Engineering Department, Ecole Nationale d'Ingénieurs de Tunis, University Tunis El Manar,
zied.mnasri@enit.utm.tn

1. Introduction

In this paper, a novel relationship between instantaneous frequency (IF) and fundamental frequency (F0) in voiced parts of speech signals is presented. F0 detection from IF was already conducted by Qiu & Al, Kobayashi & Al. and Huang & Al., mainly using HHT (Hilbert Huang Transform). However, most of these techniques rely on empirical processing. In the proposed method, IF is calculated as the time-derivative of the phase of the analytic signal, yielding from Hilbert transform. Whereas F0 contour can be extracted using any classical pitch tracking technique (autocorrelation, cepstrum, SHR …etc.), to establish a direct relationship between IF and F0. This relationship has been verified independently of the tool used to extract F0. This relationship states that the envelope of the residual of the instantaneous frequency, defined as the difference between IF and the maximum of harmonics tends to F0. Such a direct relationship may be useful for further developments of F0 extraction directly from the speech signal, avoiding the approximation that exists in most pitch extraction techniques.

2. Established relationship between pitch and instantaneous frequency

Starting from the assumption that IF carries F0 and its harmonics, some novel notations are proposed in the following.

Instantaneous pitch: It can be defined as the smallest possible F0 value for which IF is the closest to its highest multiple (or to its highest harmonic).

Instantaneous harmonic: It is the multiple of the instantaneous pitch which ids the closed to the corresponding IF. Consequently, the instantaneous harmonic order is defined as the floor of IF divided by F0, as in (1):

$$N_h(k) \quad = \quad floor\left(\frac{f_i(k)}{f_0(k)}\right) \quad (1)$$

Instantaneous residual frequency: It is defined as the difference between IF and the largest harmonic at each instant, as in (2)

$$f_{ir}(k) \quad = \quad f_i(k) \quad - \quad N_h(k)f_0(k) \quad (2)$$

Finally, F0 contour is obtained from the maximum value of the instantaneous residual frequency. These maxima are calculated on overlapping frames of small duration (less than 40ms), as in (10).

$$\begin{matrix} f_{0est}(n_k) & = & max(f_{ir}(n_k)) \\ (k-1)shift \;\; \leq & n_k & \leq \;\; (k-1)shift + frame\_length \end{matrix} \quad (3)$$

This relationship between IF and F0, as given in (2) and (3), was verified and validated on a large set of signals. Actually, F0 used in (1) and (2) are extracted by any conventional technique of pitch tracking. In the case of this study, SHR algorithm was used with 20-ms frame duration and 5-ms shift, and with activating the voicing check option, that sets F0 values to zero in unvoiced parts of speech.

3. Proposed algorithm

After checking the accuracy of the proposed relationship between IF and F0, a practical method is developed to estimate F0. Actually, SHR-extracted F0 values were used only to check the proposed relationship.

A practical method to estimate F0 starting from IF and using dynamic programming will be carried out as follows:

For each time index:

1 Sweep the possible values of F0, i.e. the frequency range of [50 Hz, 300Hz] in case of speech signal.

2 Calculate the residual IF as in (9) for each candidate F0 value.

3 Keep the f0 value which minimizes the difference between the areas swept by previous f0 kept values and the envelope of the residual IF calculated so far.

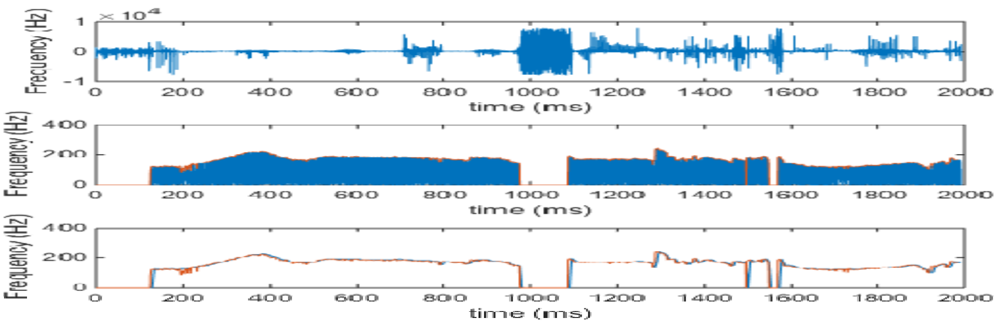

Figure1. (1) IF, (2) residual IF(blue) and its envelope (red), (3) Envelope of residual IF (red) and original f0 contour (blue) for a speech signal