ON THE RELATIONSHIP BETWEEN INSTANTANEOUS FREQUENCY AND PITCH IN SPEECH SIGNALS

Zied Mnasri[1], *Hamid Amiri*[1]

[1]*Electrical engineering dept, National School of Engineering in Tunis, University Tunis El Manar, Tunisia*
zied.mnasri@enit.utm.tn, hamid.amiri@enit.utm.tn

Abstract: In this paper, a novel relationship between instantaneous frequency (IF) and fundamental frequency (F0) in voiced parts of speech signals is presented. IF is calculated as the time-derivative of the phase of the analytic signal, yielding from Hilbert transform. Whereas F0 can be extracted using any classical pitch tracking technique (e.g. autocorrelation, cepstrum, subharmonic-to-harmonic ratio (SHR) …etc.), 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.

1 Introduction

Pitch is one of the most prominent parameters in speech. Phonologically, pitch is related to intonation and accentuation, and phonetically, pitch is expressed by F0 values in voiced parts. Hence, information about pitch may be useful for any speech processing application, such analysis, recognition, synthesis …etc. Therefore, a variety of techniques were developed to accurately measure pitch for speech signals. The main techniques could be classified according to their domain, whether temporal, spectral or both [1]. Another classification, proposed by [2], splits the pitch tracking into event-detection techniques, like peak-picking and zero-crossing, and short-time average F0 detection techniques, like autocorrelation [3], minimal distance methods [2], cepstral analysis [4] and harmonic analysis.

However, most of these techniques are applied in a short time manner, in order to reduce the effects of non-linearity and non-stationarity of speech. This short time processing usually leads to errors while estimating the pitch periods [5]. Also, wavelets are used to extract pitch, but with their inherent defaults, mainly spectral leakage and poor time-frequency resolution [5].

Therefore, a new set of techniques applied along the whole signal have been developed during the last two decades. Most of them rely on the notion of instantaneous frequency (IF), which is defined as the time-derivative of the phase of the analytic signal, obtained through Hilbert transform [6]. Three main IF-based technique were developed [7], [8] and [9] with recognized performance. However, most of these methods are based on empirical assumptions, where F0 is either as the smallest harmonic [8], or as a filtered discrete IF [7] or as the IF corresponding to the greatest instantaneous amplitude of the signal IMF's (Intrinsic mode functions) components, obtained by EMD (Empirical mode decomposition) [5].

Whereas F0 is accurately extracted from IF by these techniques a direct relationship between IF and F0 is still looked for, to fill the gap between successful empirical approaches and the lack of explaining theory. Therefore a novel relationship between IF and F0 in voiced parts of speech signal is proposed in this work.

This paper is organized as follows; section 2 presents the main IF-based pitch tracking techniques, section 3 details the mathematical formulation and the physical interpretation of IF, then section 4 presents a proposition for a direct relationship between IF and F0 in case of speech signals and an algorithm to implement this relationship. The main findings will be commented and discussed in section 5.

2 Instantaneous frequency-based Pitch tracking

Instantaneous frequency (IF) offers the possibility to avoid the issues of conventional techniques, since IF pattern is continuously examined along the signal, and then there's no need to use truncated segments to reduce non-stationarity effect, nor to adjust the wavelet scale to enhance time-frequency resolution.

Most of these technique start from the IF values to extract F0 contour as a continuous function of time (F0 is considered null if unvoiced segments). Qiu & Al start by a) attenuating the harmonics, through a band-pass filter bank, b) estimating the discrete IF (DIF) at different scales of the band-pass filter bank, c) deciding about voicing based on a set of criteria related to the DIF value (less than 50Hz or greater than 500Hz) or to the variation between neighboring DIF's (greater than 1.4Hz) or to the duration of sustained DIF (if it's less than 20ms) [7]. Nevertheless, in this technique, low harmonics (less than 500Hz) may be confounded with F0 values. Therefore, multiple scales of the filter bank are used. Then the smallest non-zero DIF is retained as F0.

In a similar approach, Kobayashi & Al. used IF pattern to track harmonics and extract F0. In this technique, a band-pass filter bank with variable center frequency is applied to decompose the signal into harmonic components. Then the IF of each component is considered as the harmonic pattern. Hence the lowest IF pattern (i.e. the lowest harmonic) is considered as the F0 contour [8].

Huang & Al. proposed another IF-based technique, as a direct application of the Huang-Hilbert Transform (HHT) [9]. Actually, HHT is a two-fold process performed by a) EMD (empirical mode decomposition) where the signal is decomposed into IMF's (Intrinsic mode functions) by 'Sifting', where each IMF is characterized by its IF and IA. Then, to extract F0 (and also voicing decision), first a filtering phase is applied to all IMF's, where only IF values between 50Hz and 600Hz are kept, and where IF values are set to zero if $\delta f_i \geq 100Hz$ in a 5-ms frame or when $A_i(t) \leq \frac{MaxA_i}{10}$. Then F0 value is selected as the IF value corresponding to the highest IA value in all IMF's. Finally post filtering is applied to merge and smooth the obtained F0 values.

All of the aforementioned IF-based pitch extraction techniques were tested and compared to classical methods, giving very accurate voicing decision and F0 values, which proves that IF succeeds to reduce the effect of non-linearity and non-stationarity on pitch tracking.

However, most of these techniques are based on empirical assumptions, where F0 is either taken as the smallest harmonic [Kobayashi95], or as the filtered discrete IF [7], or as the filtered IF having the highest IA in all IMF's obtained by EMD [5]. Thus, none of these techniques propose a direct or an analytic relationship between IF and F0, though in each case, F0 is considered as a particular value of IF. Therefore, a direct relationship is proposed in this paper, which actually starts from the same assumptions in all the described IF-based techniques. Actually, F0 will be described as the local maximum of the residual IF pattern, which is the difference between IF and the highest harmonics. Then an algorithm is proposed to determine F0 from IF, according to this relationship.

3 Instantaneous frequency and its physical interpretation

Though IF physical meaning is still controversial, its existence is mathematically proven, since it's considered as the time-derivative of the phase of the analytic signal.

3.1 Definition of instantaneous frequency

The analytic signal $z(t)$ is obtained from a signal $s(t)$ by

$$z(t) = s(t) + js_H(t) \quad (1)$$

Where

$$\begin{aligned} s_H(t) &= H.T\big(s(t)\big) \\ &= p.v\int_{-\infty}^{+\infty}\frac{s(t-\tau)}{\pi\tau}d\tau \end{aligned} \quad (2)$$

Where H.T denotes the Hilbert transform and p.v. the Cauchy principal value of $\int_{-\infty}^{+\infty}\frac{s(t-\tau)}{\pi\tau}d\tau$

An important consequence is that

$$z(t) = a(t)e^{j\varphi(t)} \quad (3)$$

Since $z(t)$ is unique for a given $s(t)$ [10], then

$$s(t) = a(t)\cos\big(\varphi(t)\big) \quad (4)$$

$a(t)$ and $\varphi(t)$ respectively defined as the instantaneous amplitude and phase.

It should be noted that this definition does not require neither the stationarity nor the linearity of the system producing s(t), which makes it valid for any natural signal.

In a generalization of the phase in case of non-harmonic signal, $\varphi(t)$ can be written as in (5). [6]

$$\varphi(t) = 2\pi\int_0^t f(t)dt \quad (5)$$

It's obvious that $\varphi(t)$ would have the classical formula $\varphi(t) = 2\pi ft$ in case of a harmonic signal.

Here came the idea to define the instantaneous frequency as the derivative over time [11], [12], [6], as in (6)

$$\begin{aligned} f_i(t) &= \frac{1}{2\pi}\frac{d\varphi(t)}{dt} \\ &= \frac{1}{2\pi}\frac{d\,arg\big(z(t)\big)}{dt} \end{aligned} \quad (6)$$

Then for a discrete signal, the IF is easily calculated by (7)

Where z(n) is the associated discrete analytic signal and f_s is the sampling frequency (for $n \geq 1$).

$$f_i(n) = \frac{f_s}{4\pi}\big[arg\big(z(n+1)\big) - \arg(z(n-1))\big] \quad (7)$$

3.2 Physical usefulness of instantaneous frequency in speech signal

Whereas F0 is defined as the proper frequency of a phenomenon, matching to the local peak of Fourier magnitude spectrum in case of a harmonic signal, or the pitch period in case of speech, it's more difficult to find a physical interpretation of IF. Actually, there's no evident and direct relationship between Fourier and Hilbert spectra, though some interaction may exist [13].

Meanwhile, IF can be regarded as the carrier of harmonics, since IF exists at every instant, including those corresponding to the period of each harmonic. Then one can look at F0 and its harmonics as special values of IF.

4 Established relationship between pitch and instantaneous frequency

4.1 Proposed relationship

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

4.1.1 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).

4.1.2 Instantaneous harmonic

It is the multiple of the instantaneous pitch. Then IF is again defined as the closest end to the highest instantaneous harmonic. Consequently, the instantaneous harmonic order is defined as the floor of IF divided by F0, as in (8):

$$N_h(k) \quad = \quad floor\left(\frac{f_i(k)}{f_0(k)}\right) \quad (8)$$

4.1.3 Instantaneous residual frequency

It is defined as the difference between IF and the largest harmonic at each instant, as in (9)

$$f_{ir}(k) \quad = \quad f_i(k) \quad - \quad N_h(k) f_0(k) \quad (9)$$

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).

$$\begin{array}{ccccc} f_{0est}(n_k) & & = & & max(f_{ir}(n_k)) \\ (k-1)shift & \leq & n_k & \leq & (k-1)shift + frame\_length \end{array} \quad (10)$$

This relationship between IF and F0, as given in (9) and (10), was verified and validated on a large set of signals. Actually, F0 used in (8) and (9) are extracted by any conventional technique of pitch tracking. In the case of this study, SHR algorithm [14] 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.

The next step is to align F0 contour, so that each extracted F0 value is affected to all the instants along the frame.

Figure 1 show the results for a speech signal, where *fir* denotes the residual *IF*, *F0* the SHR-extracted value and $F0_{est}$ the re-estimated *f0* values by (10).

4.2 Experimental implementation

The IF-F0 relationship check was implemented in a 3-step algorithm

4.2.1 Step 1- Check voicing which was realized using the CV-option, i.e. check voicing, in the SHR algorithm, which values were used as reference. Actually, SHR algorithm was opted for since it's based on studying the ratio of harmonics, though in the Fourier domain, and therefore it looks the most similar approach to the present one.

4.2.2 Step 2- Calculating the number of instantaneous harmonics and the residual IF: Only in voiced parts, the number of instantaneous harmonics and the residual IF was calculated using equations (8) and (9)

4.2.3 Step 3- Calculating the instantaneous F0 at each frame: the instantaneous F0 is calculated as the maximum of the residual IF at each sliding frame.

4.3 Experimental results

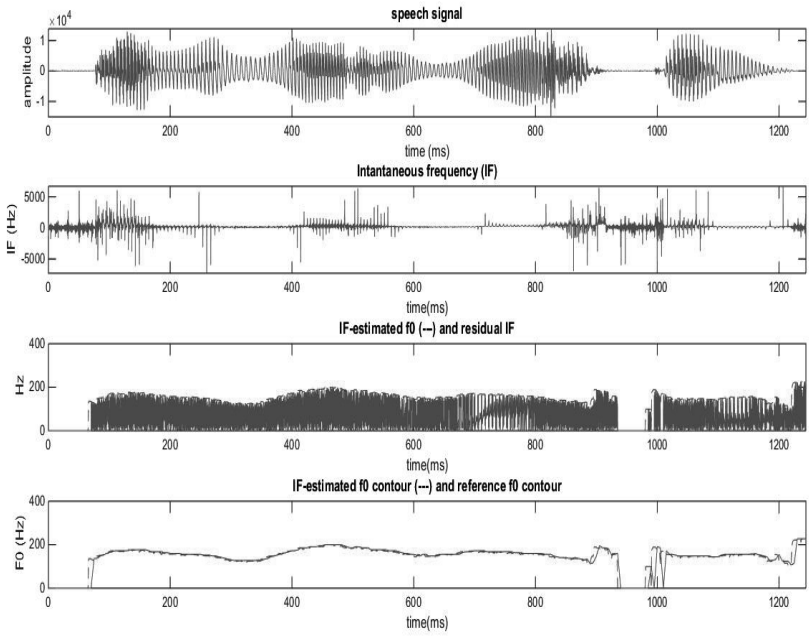

Figure 1 – Instanteneous frequency (IF), fundamental frequency (F0) and residual IF of the Arabic speech signal /laa lan yudhia alkhabara/ (No, he won't diffuse the news)

Figure 1 shows a sample of F0 extraction using the instantaneous frequency. Subplot 2 shows the IF pattern directly obtained as the time derivative of the phase of the analytic signal. In Subplot 3, the curve of the frame-maxima of the residual IF is considered as the estimated f0 contour. Then Subplot 4 shows a quite superposition between the estimated F0 contour and the reference f0 contour extracted by SHR algorithm [14], using a 20ms frame length with 5ms shift, for SHR and 5ms frame length and 1ms shift for the IF-based F0.

Since the frame length is not compulsory the same, as used to extract F0 from the IF pattern, or using the SHR-algorithm, then it would be difficult to measure the mean square error. Therefore, another measure, consisting in the relative absolute error between the areas swept by reference and estimated f0 contours could be used. Whereas the SHR-algorithm frame length was fixed at 20ms with a 5ms-shift, as it gives the best F0 values and voicing decisions, the frame length was varied in the IF-based f0 extraction algorithm.

Table 1 shows the statistics obtained through the application of both f0 extraction algorithms on four sets of speech signals, each containing 10 samples.

Table 1 – Statistical measures between IF-based and SHR-based F0 for different frame lengths

Speech DB	Voice	Fs	Frame length	Shift	Relative absolute error
DB1 [15]	Female	16 KHz	20 ms	5 ms	17.5 %
			10 ms	2.5 ms	9.3 %
			5 ms	1 ms	4.1%
DB1 [15]	Male	16 KHz	20 ms	5 ms	27.1%
			10 ms	2.5 ms	15.4 %
			5 ms	1 ms	7.8%
DB2 [16]	Female	48 KHz	20 ms	5 ms	30.1 %
			10 ms	2.5 ms	16.3 %
			5 ms	1 ms	8.9%
DB3 [16]	Male	48 KHz	20 ms	5 ms	56.8 %
			10 ms	2.5 ms	33.6 %
			5 ms	1 ms	19.4 %

5 Discussion and conclusion

In this paper, a novel relationship between IF and F0 was proposed for speech signals. Many IF-based pitch extraction methods were developed by [5], [7] and [8]. However, none of these works mentioned a direct relationship between IF and pitch, but a successful empirical technique to extract F0 from IF pattern. In this work, such a relationship is established, allowing to propose an algorithm where F0 would be directly estimated from the IF pattern of speech signals. Based on the experimental results, the smaller is the frame length; the better is the extraction performance. Then further developments could improve the algorithm, especially in terms of reducing its complexity for a small frame length.

Literature

[1] DRUGMAN, T. and ALWAN, A.: *Joint robust voicing detection and pitch estimation based on residual harmonics*. In: *Twelfth Annual Conference of the International Speech Communication Association*..

[2] HESS, W.: Manual and instrumental pitch determination, voicing determination. In *Pitch Determination of Speech Signals*, p. 92-151. Springer, Berlin, Heidelberg, 1983.

[3] RABINER, L.: *On the use of autocorrelation analysis for pitch detection.* In *IEEE transactions on acoustics, speech, and signal processing*, vol. 25, no 1, p. 24-33, 1977.

[4] NOLL, A. M.: *Cepstrum pitch determination*. In *The journal of the acoustical society of America*, vol. 41, no 2, p. 293-309, 1967.

[5] HUANG, H., PAN, J.: *Speech pitch determination based on Hilbert-Huang transform*. In *Signal Processing*, vol. 86, no 4, p. 792-803, 2006.

[6] BOASHASH, B.: *Estimating and interpreting the instantaneous frequency of a signal. II. Algorithms and applications*. In *Proceedings of the IEEE*, vol. 80, no 4, p. 540-568, 1992.

[7] QIU, L, YANG, H., KOH, S.: *Fundamental frequency determination based on instantaneous frequency estimation*. In *Signal Processing*, vol. 44, no 2, p. 233-241, 1995.

[8] ABE, T, KOBAYASHI, T., IMAI, S.: *Harmonics tracking and pitch extraction based on in-*

stantaneous frequency. In : *Acoustics, Speech, and Signal Processing, ICASSP-95., 1995 International Conference on*. IEEE, p. 756-759, 1995.

[9] HUANG, NORDEN E., ZHENG S., STEVEN R., MANLI C., HSING H., SHIH, ZHENG, Q., , Nai-YEN, C, TUNG, C. C., and LIU, H.: *The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis.* In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*, vol. 454, no. 1971, pp. 903-995. The Royal Society, 1998.

[10] GABOR, D.: *Theory of communication. Part 1: The analysis of information.* In *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no 26, p. 429-441., 1946.

[11] VAN DER POL, B.: *The fundamental principles of frequency modulation.* In *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no 23, p. 153-158., 1946.

[12] VILLE, J.: *Theorie et application de la notion de signal analytic, Cables et Transmissions,* 2A (1), 61-74, Paris, France, Translation by SELIN, I.*, Theory and applications of the notion of complex signal, Report T-92, RAND Corporation, Santa Monica, CA*., 1948.

[13] LIFLYAND, E.: *Interaction between the Fourier transform and the Hilbert transform.* In *Acta et Commentationes Universitatis Tartuensis de Mathematica* 18, no. 1 (2014): 19., 2014.

[14] SUN, X.: *A pitch determination algorithm based on subharmonic-to-harmonic ratio.* In *Sixth International Conference on Spoken Language Processing*. 2000.

[15] EUSTACE,: speech database available online at http://www.cstr.ed.ac.uk/projects/eustace

[16]PTDB-TUG,: Pitch tracking database of the T.U. Graz, available online at www.spsc.tugraz.at.