Circuits, Systems, and Signal Processing (2022) 41:6266–6294
https://doi.org/10.1007/s00034-022-02082-8

A Novel Pitch Detection Algorithm Based on Instantaneous Frequency for Clean and Noisy Speech

Zied Mnasri[1,2] · Stefano Rovetta[2] · Francesco Masulli[2]

Received: 20 May 2021 / Revised: 31 May 2022 / Accepted: 4 June 2022 /
Published online: 25 June 2022

Abstract

In this paper, a novel pitch detection algorithm (PDA) is proposed. Actually, pitch detection is a classical problem that has been investigated since the very beginning of speech processing. However, the novelty of the proposed method consists in establishing an empirical relationship between fundamental frequency (f_0) and instantaneous frequency (f_i), which serves as a basis to develop the proposed PDA. Even though f_0 and f_i are defined as attributes of two different transforms, i.e., the Fourier transform and the Hilbert transform, respectively, the relationship proposed in this paper shows some interaction between both of them, at least empirically. The first step of this work consists in validating the proposed relationship on a large set of speech signals. Then, it is leveraged to develop an algorithm capable to (a) detect voiced/unvoiced parts of speech and (b) extract f_0 contour from f_i values in the voiced parts. For evaluation purposes, the yielding f_0 contour is compared to some well-rated state-of-the-art PDA's. The main findings show that the quality of pitch detection obtained by the proposed technique is as satisfactory as some of top PDA's, either in clean or in simulated noisy speech. In addition, one of the main advantages consists in bypassing the traditional short-time analysis required to assume local stationarity in speech signal.

Keywords Pitch detection algorithm (PDA) · f_0 contour · Instantaneous frequency · Voiced/unvoiced decision · Objective evaluation

✉ Zied Mnasri
zied.mnasri@enit.utm.tn

Stefano Rovetta
stefano.rovetta@unige.it

Francesco Masulli
francesco.masulli@unige.it

1 ENIT, University of Tunis El Manar, Tunis, Tunisia

2 DIBRIS, University of Genoa, Genoa, Italy

1 Introduction

Pitch is among the most prominent parameters in speech. From a phonological point of view, pitch is responsible of intonation and accentuation, whereas from the acoustic side, pitch is quantified by voiced/unvoiced (V/UV) decision and f_0 contour. Along with acoustic energy, it conveys most para-verbal content and may dramatically change the meaning of the verbal component, for instance by representing the interrogative form or ironic intent. It is also a major component of emotion, a key human–machine communication mode which is still in its infancy, and can be used to diagnose several neuropsychological conditions like early cognitive impairment or depression.

These application domains have received a strong boost in the past few years with the widespread diffusion of vocal assistants, vocal interfaces, personal health devices, and with the developments in collaborative and cooperative robotics, all of which call for detection methods that are both effective and computationally light-weight and more accurate than the state of the art.

Pitch detection is probably the speech processing problem which have had the biggest interest. Several techniques have been implemented during the last half century, to provide an accurate measure of such a highly variable speech feature. Actually, pitch depends on a variety of parameters, mainly the speaker's gender, age and the language type, i.e., tonal or non-tonal. A classification of the main pitch detection techniques can be made according to the domain of analysis, whether temporal, spectral or time-frequency [12]. In [17], another classification is proposed, dividing the pitch detection methods into event-detection techniques, like peak-picking and zero-crossing, and short-time average f_0 detection techniques, such as cepstrum [28], autocorrelation [32] and average magnitude difference functions (AMDF) [34], minimal distance methods [17] and harmonic analysis-based techniques [12, 16, 38]. As a common point, the aforementioned techniques are applied on short time frames, to reduce the effects of non-stationarity of the speech signal. However, such a short time processing may lead to errors while estimating the pitch periods [19]. On another side, multiresolution analysis methods, such as discrete wavelet transform (DWT), are utilized to extract pitch [21]. Nevertheless, their performance is influenced by their inherent defaults, mainly poor time-frequency resolution and spectral leakage, as noticed in [19].

To tackle these issues, another concept has emerged in the last two decades, based on techniques applied along the whole signal, instead of short-time analysis. The majority of these techniques are based on the analysis of instantaneous frequency (f_i), which is a theoretic concept. By definition, f_i is the time-derivative of the phase of the analytic signal. The latter is a complex signal obtained by Hilbert transform [6]. However, using f_i values to extract f_0 contour still suffers from the lack of a direct/explicit relationship between the two quantities.

Therefore, a novel relationship, although still empirical, is proposed in this work, in order to (a) determine the voiced vs. unvoiced parts of the speech signal, and (b) extract f_0 contour from f_i values in the voiced parts. This work is described as follows: Sect. 2 reviews the related work, with a focus on harmonic analysis-based and f_i-based pitch detection techniques; Sect. 3 presents the method adopted, including the proposed empirical relationship between f_i and f_0 in speech signal, and detailing the algorithm developed to extract f_0 from f_i through this relationship. Section 4 presents the objective evaluation protocol and the results yielding from the application of the proposed algorithm on clean and simulated noisy speech. Finally, the performance measures are commented and discussed, with some proposals for improvement.

We note that the empirical relationship between f_0 and f_i was already presented in [26], whereas the description of the proposed PDA has recently been accepted in [27]. The present paper contains an extended description of the utilized method, and especially a comprehensive evaluation, including not only the performance of the proposed algorithm with respect to its parameters, but also the comparison to state-of-the-art techniques, in various noise conditions and separately for male and female speakers.

2 Related Work

Even though pitch detection is a classical audio and speech processing problem, research in this field has never given up. In fact, several challenges are still open, such as multi-pitch detection, accurate pitch detection in noisy environments and real-time pitch tracking. Thus, two main directions are followed, (i) the classical harmonic analysis based on short-time Fourier transform (STFT) and (ii) the instantaneous spectrum based on Hilbert transform (HT).

2.1 Pitch Detection by Harmonic Analysis

The use of harmonics to detect f_0 in speech and music signals has been the key idea of several pitch detection algorithms, such as the subharmonic summation (SHS) [16], subharmonic-to-harmonic ratio (SHR) [38] and residual harmonics [12].

In [40], a harmonic model is applied to estimate voiced speech parameters. In particular, a maximum a posteriori probability is estimated to compute the fundamental frequency (f_0). In [45], an algorithm named YAAPT (Yet Another Algorithm for Pitch Tracking) leverages the STFT spectrum of the filtered squared signal to provide a primary estimation of f_0. Then, the f_0 candidate values are refined, and dynamic programming is applied to select the path containing the series of f_0 candidates that minimizes a cost function composed of a merit term and a transition term. This method has been particularly robust to white and babble noise.

In [5], the BaNa algorithm is presented. This algorithm is qualified as hybrid since it combines two main pitch detection approaches, i.e., harmonic ratios, like in SHR [38], and cepstrum analysis [28]. The final pitch value is calculated using a *Viterbi* algorithm to decode the optimal path search, where the cost function is defined as a sum of the log-ratio of each pair of f_0 candidates plus a weighted confidence score. The evaluation of this method to some state-of-the-art PDA's such as PRAAT [7] and YIN [10] shows that it is more robust to noise.

In [42], a method based on multi-band summary correlogram (MBSC) is developed for the problem of pitch detection in noisy environments. Thus, the input signal is filtered by multiple wide-band FIR filters. The extracted envelope at each frequency band is filtered again with a multi-channel comb filter and harmonic-to-subharmonic ratio (HSR) computation. Then, the MBSC is computed on the samples selected by HSR. Finally, the pitch value is retained as the candidate that has the smallest MBSC.

In [14], a PDA named PEFAC (Pitch Estimation Filter with Amplitude Compression) has been proved to be quite efficient in noisy speech, even with negative SNR. This PDA proceeds by: (i) normalizing the signal to remove channel dependencies, (ii) attenuating the strong noise components by harmonic summation and (iii) applying temporal continuity constraints to the selected pitch candidates.

More recently, [44] has proposed harmonic enhancement to cope with the issue of missed and submerged harmonics in spectrum before performing pitch detection; and lately, [33] has developed a spectrum-based PDA for singing voices. The latter method is based on candidate harmonic partials detection using a random forest classifier. Each triplet of successive f_0 candidates undertakes a harmonicity/ unharmonicity check using adapted thresholds.

2.2 Instantaneous Frequency-Based Pitch Detection

Using instantaneous frequency (f_i) for pitch detection is an alternative way to get around some problems of conventional methods, such as harmonic analysis and multiresolution analysis. In fact, f_i pattern can be continuously analyzed along the whole signal, which allows avoiding some constraints, such as short-time analysis, that is usually required to reduce the effect of non-stationarity of the speech signal, and wavelet scale adjustment, which is necessary to enhance the time-frequency resolution [19].

Even though it is less frequent to use instantaneous analysis for pitch detection, a few PDA's based on f_i analysis were proposed in the literature [1, 2, 18, 31] and more recently in [22], with valuable performance. These f_i-based methods extract f_0 contour as a continuous function of time in voiced regions. For instance, Qiu et al. [31] proceed as follows: First, the harmonics are attenuated using a band-pass filterbank; secondly, the discrete instantaneous frequency (DIF) is estimated at different scales of the band-pass filterbank; and finally, the V/UV decision is taken upon certain criteria related to: (i) the DIF value (unvoiced if DIF $\leq$ 50 Hz or DIF $\geq$ 500 Hz), or (ii) to the variation between neighboring DIF's (unvoiced if Δ(DIF) $\geq$ 1.4 Hz), or (iii) to the duration of sustained DIF (unvoiced it is less than 20 ms). However, using this technique is likely to cause the problem of pitch halving/doubling, where the low harmonics, i.e., the multiples of f_0 that are less than 500 Hz, could be also taken for f_0 values. To cope with this issue, multiple scales of the filterbank are used, to retain the smallest non-zero DIF as f_0 value.

In [1], Abe et al. used f_i pattern to extract f_0 by tracking the harmonics. To achieve this goal, the signal is decomposed into harmonic components by applying a filterbank with a variable center frequency. Then, f_i values of each component are considered as the harmonic pattern. Finally, the lowest f_i pattern, i.e., the lowest harmonic, is retained as the f_0 contour [1]. In continuation to this work, the same authors proposed in [2] an IF-based method where IF is extracted from the spectrum of the short-time Fourier transform (STFT) to enhance harmonics, by suppressing aperiodic components. This method was reported to perform well in presence of noise, in comparison to its contemporaneous state of the art, such as the dynamic programming-based cepstrum methods, proposed by [23].

In [18], the Hilbert-Huang transform (HHT) is applied for pitch detection from f_i pattern. Originally, HHT is a twofold process, that is performed first by applying empirical mode decomposition (EMD), and then by decomposing the signal into intrinsic mode functions (IMF) through a special process called *sifting*. Each resulting IMF is characterized by its instantaneous frequency (f_i) and its instantaneous amplitude (A_i). After extracting all IMF's, f_0 and V/UV decision are estimated, first by filtering all IMF's, where only f_i values between 50 Hz and 600 Hz are kept, and where f_i values are set to zero if $\Delta f i \geq 100$ Hz in a 5 ms frame or when the instantaneous amplitude $A_i(t) \leq \frac{\max(A_i)}{10}$. At each instant, the f_i value corresponding to the highest A_i value in all IMF's, is retained as f_0 value. Finally, the extracted f_0 contour is merged and smoothed by post-filtering.

More recently, [22] leveraged aperiodicity bands and short-time Fourier transform (STFT)-based channel-wise instantaneous frequency, defined as in (1)

$$f_i(t, \omega) = \frac{1}{2\pi} \frac{\partial \Phi(t, \omega)}{\partial t} \tag{1}$$

where $\Phi(t, \omega)$ is the STFT phase spectrum, to estimate f_0 for speech signal synthesis, using a three-stage process. In the first stage, aperiodicity bands are detected using a wavelet-based analysis filter with a highly selective temporal and spectral envelope. In this stage, instantaneous frequency is filtered to yield the periodicity probability map. The second stage generates a first estimate of f_0 trajectory from the periodicity probability map and signal power information. Finally, the third stage refines the estimated f_0 trajectory using the deviation measure of each harmonic component and f_0 time warping. It is worth noting that this PDA has been included to Google's vocoder named YANG (Yet ANother Generalized Vocoder) [3].

2.3 Interaction Between Fundamental Frequency and Instantaneous Frequency

Most of the aforementioned f_i-based pitch extraction techniques have been successfully compared to the rest of state-of-the-art methods, yielding a very accurate V/UV decision and f_0 values, which proves that using f_i is a good alternative to extract f_0 without taking care of the non-stationarity of the speech signal. Nevertheless, these methods are mostly based on empirical hypotheses, such as considering f_0 as a filtered discrete instantaneous frequency [31], or as the smallest harmonic [1], or as the instantaneous frequency matching to the greatest instantaneous amplitude of the intrinsic mode functions (IMF), which are extracted from the signal by empirical mode decomposition (EMD) [18]. Thus, none of these techniques is based on a direct or an explicit relationship between f_i and f_0, even though in each case, f_0 contour is extracted from f_i values.

Such a relationship could fill the gap between accurate empirical methods and the lack of a theoretical link between both quantities, i.e., f_0 and f_i. It should be noted that in [24], some interaction between Fourier transform and Hilbert transform is proposed for harmonic signals. Actually, [24] relates from that Fourier transform (FT) weakly generates Hilbert transform (HT) for any well-defined function $g(x)$, as follows:

$$\mathrm{FT}(\mathrm{HT}(g(x)) = \mathrm{i}\sigma(x)\mathrm{FT}(g(x)) \tag{2}$$

where i is the imaginary number and $\sigma(.)$ is the sign function. However, (2) is not enough to generalize the relationship between f_0 and f_i for two major reasons: First, speech signal is far from being strictly harmonic (notwithstanding the possibility to model speech by a harmonic-plus-noise model (HNM) [36]); and secondly, f_0 contour is not continuous over the speech signal due to binary voiced/unvoiced (V/UV) decision.

Also, Shimauchi et al. [35] have recently established an explicit relationship between the STFT magnitude spectrum and the channel-wise instantaneous frequency, defined by (1), such that:

$$f_i(t,\omega) = \frac{1}{2\pi\sigma^2}\frac{\partial log(A(t,\omega))}{\partial\omega} + \frac{\omega}{2\pi}. \tag{3}$$

where $\Phi(t,\omega$ and $A(t,\omega)$ are the STFT phase and magnitude, respectively; t and ω are the time and the angular frequency in the Fourier domain, respectively. This relationship has been used for phase retrieval, i.e., phase estimation given the magnitude spectrum only. Nevertheless, this result does not lead to a direct/explicit relationship between the STFT-based channel-wise instantaneous frequency and the Hilbert transform-based one.

3 Method

In this work, a direct relationship is proposed, albeit it is still empirical. This relationship relies on the same assumptions utilized in the aforementioned f_i-based techniques. Then, this relationship is used to implement an algorithm able to: (a) determine the voiced/ unvoiced parts and (b) extract f_0 contour from f_i values in the voiced regions.

3.1 Computation of Instantaneous Frequency

Instantaneous frequency (f_i) is a theoretic concept defined as the time-derivative of the phase of the analytic signal $z(t)$. The latter is the complex signal given by:

$$z(t) = s(t) + js_H(t) = a(t)e^{j\phi(t)}, \tag{4}$$

where

$$s_H(t) = \mathrm{HT}(s(t)) = \mathrm{pv}\left(\int_{-\infty}^{+\infty}\frac{s(t-\tau)}{\pi\tau}\,\mathrm{d}\tau\right). \tag{5}$$

HT and pv denote the Hilbert transform and the Cauchy principal value, respectively, whereas $a(t)$ and $\phi(t)$ are the instantaneous amplitude and the instantaneous phase, respectively. As $z(t)$ is unique for a given $s(t)$ [13], then:

$$s(t) = a(t)\cos(\phi(t)), \tag{6}$$

Since no restrictions are required concerning the stationarity or the linearity of the system that generates $s(t)$, (6) is valid for any natural signal. In [6], based on the earlier works of [30, 43], the generalized instantaneous phase $\phi(t)$ can be written as:

$$\phi(t) = 2\pi \int_0^t f(t)\,\mathrm{d}t. \tag{7}$$

It is obvious that $\phi(t)$ would have the classical formula $\phi(t) = 2\pi f t + \phi_0$ in case of a simple harmonic signal. Hence, the instantaneous frequency f_i can be defined as the time-derivative of the instantaneous phase ϕ as in (8), based on [6]:

$$f_i(t) = \frac{1}{2\pi}\frac{\mathrm{d}\phi(t)}{\mathrm{d}t} = \frac{1}{2\pi}\frac{\mathrm{darg}(z(t))}{\mathrm{d}t}. \tag{8}$$

For discrete signals, f_i is easily calculated by (9), where $z(n)$ is the associated discrete analytic signal and f_s is the sampling frequency (for $n \geq 1$):

$$f_i(n) = \frac{f_s}{4\pi}(\arg(z(n+1)) - \arg(z(n-1))). \tag{9}$$

3.2 Proposed Empirical Relationship Between Pitch and Instantaneous Frequency

In spite of the absence of a direct relationship between f_i and f_0, both types of frequency share a common point, which is continuity over time, at least in the regions where f_0 contour is defined, such as the voiced parts of a speech signal. This suggests that in such a region, the observed instantaneous frequency can be a relative integer multiple of f_0 with or without some residual frequency (note that in (9), f_i can be negative). Starting from this assumption, some working notations are defined in the following, with the sole aim to describe the proposed method.

3.2.1 Instantaneous Pitch

It can be defined as the value of f_0 at every discrete instant n inside the voiced regions only (f_0 is undefined in unvoiced segments). This is different from conventional PDA's, where pitch is usually obtained by one value at each frame and then f_0 contour is obtained by interpolation.

3.2.2 Instantaneous Pitch Multiples

They are defined at each instant n as the positive integer multiples of instantaneous pitch $f_0(n)$ below $|f_i(n)|$. The highest instantaneous multiple is defined as the closest one to $|f_i(n)|$. Consequently, the maximum order of instantaneous pitch multiples, denoted $H_{\max}(n)$, is defined as:

$$H_{\max}(n) = \left\lfloor \frac{|f_i(n)|}{f_0(n)} \right\rfloor . \tag{10}$$

We also highlight that in this particular case and for mathematical rigor, we avoided using the term *harmonics* to refer to pitch multiples for the following reasons: (a) Harmonics are related to Fourier transform, whereas f_i is obtained by Hilbert transform; (b) to the best of our knowledge, no explicit relationship has been proved so far between f_0 and f_i, even though some interaction may exist in harmonic signals [24].

3.2.3 Instantaneous Residual Frequency

It is defined as the difference between $|f_i|$ and the instantaneous pitch multiple:

$$f_{ir}(n) = |f_i(n)| - H(n) f_0(k) \;\forall\; H(n) \leq H_{\max}(n), \tag{11}$$

where $1 \leq H(n) \leq H_{\max}(n)$ are the orders of the instantaneous pitch multiples at time n.

3.3 Estimation of Instantaneous Pitch from Residual Instantaneous Frequency

It is obvious that for the highest instantaneous pitch multiple order $H_{\max}(n)$, the residual instantaneous frequency f_{ir} is minimal and we have:

$$f_{ir}(n) \leq f_0(n).$$

In this particular case, we empirically notice that f_0 contour can be obtained as the upper bound of the envelope of the instantaneous residual frequency f_{ir}. This upper bound is calculated on overlapping frames of small duration (less than 40 ms):

$$f_{0,\text{est}}(n_k) = \max_{n_k - \frac{L}{2} \leq l < n_k + \frac{L}{2}} f_{ir}(l), \tag{12}$$

where n_k and L are the center and the length of the kth frame, respectively.

Figure 1 shows the results for a speech signal, as follows:

- Figure 1 (top plot) shows the instantaneous frequency (f_i) of a speech signal, calculated by (9).
- Figure 1 (middle plot) illustrates the residual instantaneous frequency (f_{ir}) and its envelope, calculated by (11) and (12), respectively.
- In Fig. 1 (bottom plot), the ground-truth f_0 contour is superposed with the envelope of f_{ir}, to show a quasi-superposition between both patterns. This leads us to consider the envelope of f_{ir} as an estimated value of the f_0, as proposed in (12).

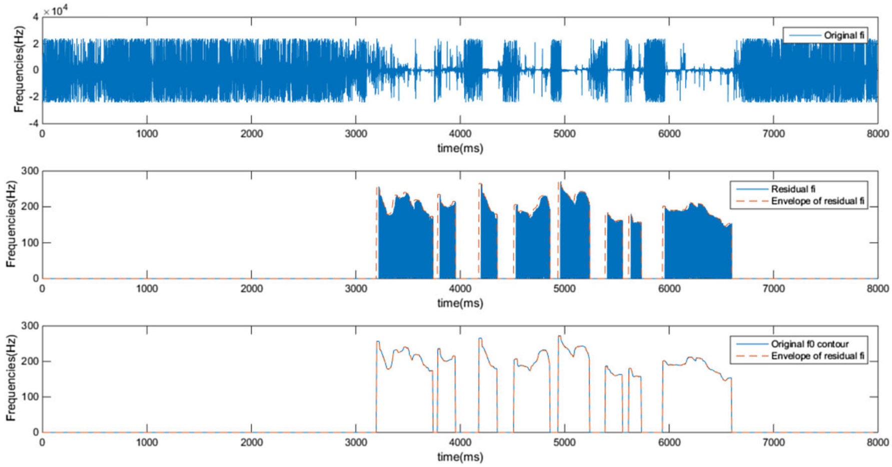


Fig. 1 Example of the established relationship between ground-truth f_0 and the instantaneous frequency f_i extracted using (4)–(9)

To validate the results given by (12), the ground-truth f_0 values provided by the standard pitch tracking database, PTDB-TUG [29], were utilized. Therefore, the ground-truth f_0 contour was first aligned to the instantaneous frequency f_i; then the residual frequency f_{ir} and the estimated fundamental frequency $f_{0_{\text{est}}}$ were calculated using (10)–(12) for different values of the order of instantaneous pitch multiples $(H(n))$, to confirm that maximizing $H(n)$, i.e., using $H_{\max}(n)$, obtained by (10), in (11), improves the superposition between the ground-truth f_0 and $f_{0,\text{est}}$ given by (12), as shown in Fig. 2.

To check further the validity of this empirical result, the root mean square error (RMSE) were measured between both contours, i.e., ground-truth f_0 and $f_{0,\text{est}}$ contour, for a large subset of signals from PTDB-TUG database [29]. In addition, the areas covered by both contours are compared, to confirm their superposition (cf. Table 1).

The test signals correspond to randomly selected 400 speech signals, uttered by 10 male and 10 female speakers. The results mentioned in Table 1 show that increasing the maximum order of instantaneous pitch multiples $H_{\max}$ in (10)–(12) makes the difference between the ground-truth f_0 contour and the estimated contour $f_{0,\text{est}}$, small enough to consider them as superposed. However, since ground-truth f_0 is already used to calculate $f_{0_{\text{est}}}$ (cf. (10)–(12)), it means that there is a recursive relationship between both, so the problem is how to extract $f_{0_{\text{est}}}$ directly from the instantaneous frequency f_i, such that it approximates the ground-truth f_0.

3.4 Proposed Pitch Detection Algorithm

To[1] extract f_0 from instantaneous frequency f_i using equations (10)–(12), the following algorithm is proposed. The algorithm is divided into three main steps: (a)

[1] MATLAB code is available at [25].

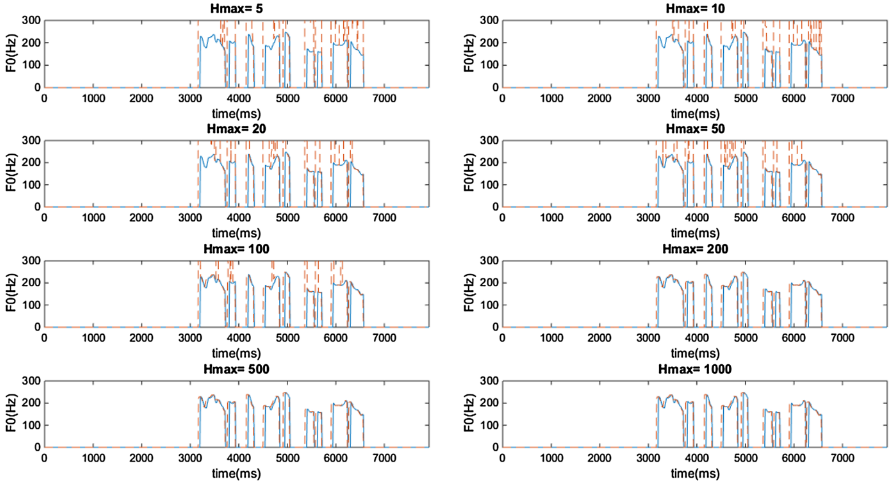


Fig. 2 Estimated f_0 contour (dotted line) versus ground-truth f_0 (continuous line) for different orders of instantaneous pitch multiples ($H_{\max}$) using (10)–(12)

Table 1 Covered area error and root mean square error (RMSE) between the contours of ground-truth f_0 and $f_{0,\mathrm{est}}$ estimated using (10)–(12)

Maximum order of instantaneous pitch multiples	Mean area error	Std area error	Mean RMSE (Hz)	Std RMSE (Hz)
5	70.9	23.6	5907.9	1255.5
10	66.1	23.2	5638.6	1221.4
20	58.3	22.2	5123.3	1157.7
50	40.2	19.1	3716.4	1008.7
100	19.5	13.9	1867.3	872.9
200	3.1	4.6	356.5	389.8
500	0.2	0.1	0.6	0.6
1000	0.2	0.1	0.6	0.6

preprocessing, where the instantaneous frequency is calculated using (4)–(9) and V/UV decision is evaluated, (b) f_0 extraction using (10)–(12), and (c) postprocessing, where the obtained f_0 contour is smoothed and segmented into voiced and unvoiced parts. The pseudocode is detailed in Algorithms 1–3 whereas Table 2 lists the settings of the parameters and thresholds used.

Step 1: Preprocessing

– Initialization

1. Extract f_i from a digital speech signal using (4)–(9).
2. Set the range of minimum and maximum f_0 values $[f_{0_{\min}}, f_{0_{\max}}]$.
3. Set the sweeping step of f_0 candidates ($f_{0_{\mathrm{cand}}}$) within the range $[f_{0_{\min}}, f_{0_{\max}}]$.

– V/UV decision

4. At each time n, the differential instantaneous frequency defined as

$$\Delta f_i(n) = (f_i(n+1) - f_i(n-1))/2,$$

is calculated. If $\Delta f_i(n) \geq Th_1$ then the point n is considered as unvoiced (cf. Table 2), otherwise voiced.

5. If the ratio of points marked as voiced within a frame is higher than the threshold Th_2 (cf. Table 2), then the whole frame is marked as voiced, otherwise unvoiced.

Step 2: f_0 extraction

6. Fix a set of $M \geq 1$ values of f_0 candidates, equally spaced by $f_{0,step}$ and ranging between $f_{0_{\min}}$ and $f_{0_{\max}}$ (cf. Table 2), $f_m = (f_{0_{\max}} - f_{0_{\min}})(m - 1)/(M-1)) + f_{0_{\min}}$, $m = 1, .., M$.
7. Set the maximum order of instantaneous pitch multiples $H_{\max}$ to be calculated at each instant n.
8. For each instant n, calculate the vector of the orders of instantaneous pitch multiples $1 \leq (H_m)_{m=1,..,M} \leq H_{\max}$ corresponding to each f_0 candidate value ($f_{0_{cand}}(n, m)$) such that

$$H_{\max,m}(n) = \min\left(H_{\max}, \lfloor \frac{|f_i(n)|}{f_{0_{cand}}(n, m)} \rfloor\right).$$

9. For each f_0 candidate value $f_{0_{\text{cand}}}(n, m)$ and each corresponding maximum harmonic order $H_{\max,m}(n)$, calculate the instantaneous residual frequency corresponding to each f_0 candidate value $(f_{ir}(n, m))_{m=1,..,M}$, using (11).
10. Calculate the value of $f_{0_{\text{cand}}}(n, \hat{m})$ at instant n such that

$$\hat{m} = \arg \min_{m=1...M} (|f_{ir}(n, m) - f_{0_{\text{cand}}}(n, m)|).$$

11. If $|f_{ir}(n, \hat{m}) - f_{0_{\text{cand}}}(n, \hat{m})| \leq Th_3$ (cf. Table 2) then $f_{0_{\text{cand}}}(n, \hat{m})$ is kept as a potential f_0 value at point n.
12. For each set of potential f_0 values kept at time n, i.e., $\{f_{0_{\text{cand}}}(n, \hat{m})\}_{\hat{m}=1,...,\hat{M}}$, if a subset of values are multiples of other ones, then keep only the lowest value within this subset, e.g., if {80 Hz, 160 Hz, 240 Hz} and {90 Hz, 180 Hz, 270 Hz} satisfy the conditions of passes (8–10) in *Step 2*, then the kept f_0 candidates are {80 Hz, 90 Hz}. Note that to bypass strict numerical inaccuracies, a kept f_0 candidate value ($f_{0,\text{cand}}(n, \hat{m}_2)$) is considered as an integer multiple of a smaller $f_{0,\text{cand}}(n, \hat{m}_1)$ if $\mod \left(\frac{f_{0,\text{cand}}(n,\hat{m}_2)}{f_{0,\text{cand}}(n,\hat{m}_1)}\right) < Th_4$ (cf. Table 2).
13. At the end of this process, if there are still ($\hat{M} > 1$) f_0 candidate values at point n that still satisfy the conditions above, then choose the f_0 candidate

value which highest multiple is the closest to $|f_i(n)|$, i.e.,

$$f_0(n) = \arg \min_{\hat{m}=1\ldots\hat{M}} \left(\mod \left(\frac{|f_i(n)|}{f_{0_{\mathrm{cand}}}(n,\hat{m})} \right) \right).$$

Step 3: Postprocessing

14. Smoothing: Apply a smoothing filter, i.e., median or linear, to the extracted f_0 values to smooth the obtained f_0 contour.
15. V/UV segmentation: Apply element-wise multiplication of the smoothed f_0 contour and the voiced/unvoiced vector obtained at *Step 1*, to set f_0 to zero in the unvoiced frames.

```
Result: VUV_frame: V/UV decision by frame
for n = 1 : length(s) − 1 do
    Calculate the Hilbert transform: s_H(n) = HT(s(n));
    Calculate the analytic signal: z(n) = s(n) + j s_H(n);
    Calculate the instantaneous frequency: f_i(n) = fs/(4π) (arg(z(n + 1)) − arg(z(n − 1)));
    Calculate the differential f_i: Δf_i(n) = (|f_i(n + 1) − f_i(n − 1)|)/2;
    if Δf_i(n) ≥ Th_1 then
        VUV(n) = 0 (The point n is unvoiced);
    else
        VUV(n) = 1 (The point n is voiced);
    end
end
for k = 1 : Number_frames do
    if (Σ_{i=1}^{L_frame} VUV(i))/L_frame ≥ Th_2 then
        VUV_frame(k) = 1 (Frame k is voiced);
    else
        VUV_frame(k) = 0 (Frame k is unvoiced);
    end
end
```

Algorithm 1: Preprocessing

4 Objective Evaluation

Before undertaking objective evaluation, an experimental protocol has been set in order to meet the requirements of such an evaluation, following recent PDA reviews [20, 37].

4.1 Evaluation Protocol

1. Select a random subset from the standard pitch tracking database, PTDB-TUG [29], containing 400 signals equally divided between the 10 male and the 10

Result: $f_{0,\mathrm{est}}$: Estimated f_0 contour
for $k = 1 : Number_{frames}$ **do**
  for $l = 1 : L_{frame}$ **do**
    $n \leftarrow (k-1) * L_{Shift} + l$;
    Initialize $f_{0,cand,min}$ vector as empty ;
    for $m = 1 : M$ **do**
      $H_{max,m}(n) = \min(H_{\max}, \lfloor \frac{|f_i(n)|}{f_{0,cand}(n,m)} \rfloor)$;
      $f_{ir}(n,m) = |f_i(n)| - H_{max,m}(n) \times f_{0,cand}(n,m)$;
      $\hat{m} = \arg \min_{m=1\ldots M} (|f_{ir}(n,m) - f_{0,cand}(n,m)|)$;
    end
    if $|f_{ir}(n,\hat{m}) - f_{0_{cand}}(n,\hat{m})| \le Th_3$;
    then
      Append $f_{0_{cand,min}}$ vector with $f_{0_{cand}}(n,\hat{m})$;
    end
    if $f_{0,cand,min}$ *vector is empty* **then**
      $f_{0,\mathrm{est}}(n) = 0$;
    else
      for $h = 1 : Length(f_{0_{cand,min}} vector)$ **do**
        for $l = h+1 : Length(f_{0_{cand,min}} vector)$ **do**
          if $mod(\frac{f_{0_{cand,min}}(l)}{f_{0_{cand,min}}(h)}) \le Th_4$ **then**
            Remove $f_{0_{cand,min}}(l)$;
          end
        end
      end
      $f_{0,\mathrm{est}}(n) = \arg \min_{\hat{m}=1\ldots\hat{M}} (mod(\frac{|f_i(n)|}{f_{0_{cand}}(n,\hat{m})}))$;
    end
  end
end

Algorithm 2: f_0 extraction

Result: Smoothed f_0 contour with V/UV decision
Apply linear or median smoothing to f_0 contour;
$f_{0,smooth} \leftarrow smoothing(f_0)$;
for $k = 1 : N_{frames}$ **do**
  for $l = 1 : L_{frame}$ **do**
    $n \leftarrow (k-1) * L_{Shift} + l$;
    $f_0(n) = f_{0,smooth}(n) * VUV_{frame}(k)$;
  end
end

Algorithm 3: Postprocessing

female speakers of the database. In fact, this database provides also ground-truth f_0, extracted from the high-pass-filtered laryngograph (LAR) signals.

2. Mix the evaluation wave files, containing initially clean speech, with babble noise and Gaussian white noise, at different SNR levels, ranging from 20 dB to 0 dB, to obtain simulated noisy speech signals.
3. Extract the f_0 contour from the input signals using state-of-the-art PDA's, namely PRAAT [7], RAPT [41], SWIPE [8], YIN [10], SHR [38], YANG [22], and finally

Table 2 Specific parameters and thresholds of the proposed algorithm

Parameter	Description	Value
L_{frame}	Frame length (in number of samples)	Corresponding to the range of 10–40 ms
L_{shift}	Frame shift length (hop size)	from 25 to 50 % of frame length
$[f_{0,\min}, f_{0,\max}]$	Range of search of f_0	[80 Hz, 270 Hz] for male, [120 Hz, 400 Hz] for female voices
$f_{0,\text{step}}$	Step of choosing f_0 candidates by sweeping the search range $[f_{0,\min}, f_{0,\max}]$	from 0.1 to 2 Hz (for reasonable computational load)
$H_{\max}$	Maximum order of instantaneous pitch multiples	From 50 to 1000 (upper-bounded by the result of (10))
Th_1	Threshold for instantaneous VUV decision: If $\Delta f_i(n) \geq Th_1$ then the point n is considered as unvoiced, otherwise the point n is voiced	For clean speech, from 1 KHz for male to 1.5 KHz for female voices, whereas for noisy speech, use a dynamic threshold (e.g., mean of Δf_i within the frame)
Th_2	Threshold for the VUV decision for each frame: If $(\sum_{i=1}^{L_{\text{frame}}} VUV(i))/L_{\text{frame}} \geq Th_2$ then Frame k is voiced; otherwise Frame k is unvoiced	$0.8 \leq Th_2 \leq 0.95$
Th_3	Tolerance to keep an f_0 candidate as a potential f_0 value: If $\lvert f_{ir}(n, \hat{m}) - f_{0_{\text{cand}}}(n, \hat{m})\rvert \leq Th_3$ then $f_{0_{\text{cand}}}(n, \hat{m})$ is kept as a potential f_0 value at point n	$0 \text{ Hz} \leq Th_3 \leq 1 \text{ Hz}$
Th_4	Tolerance to remove a multiple of an f_0 candidate at each point: If $\bmod\ (f_{0,\text{cand}}(n, \hat{m}_2)/f_{0,\text{cand}}(n, \hat{m}_1)) \leq Th_4$ then $f_{0,\text{cand}}(n, \hat{m}_2)$ is considered as an integer multiple of $f_{0,\text{cand}}(n, \hat{m}_1)$	$0 \leq Th_4 \leq 10$

the proposed algorithm (Prop.) [25]. More details about the aforementioned PDA's are in Table 3. We note that these PDA's have been selected for benchmarking based on their high performance for clean and noisy speech in a recent review [20].

4. For each pair of ground-truth and extracted f_0 contours calculate the standard metrics used in pitch detection evaluation, i.e., V/UV decision error (VDE (%)), gross pitch error (GPE (%)), f_0 frame error (FFE (%)) and fine pitch error (FPE

Table 3 Description of the PDA's used for benchmarking

PDA	Description	Version and implementation
PRAAT [7]	A toolbox of speech analysis, including pitch and formants detection. Pitch is detected using the autocorrelation-based method (AC) [32]	PRAAT version provided in [7]
RAPT [41]	A robust PDA that estimates the overall periodicity of the analysis frame using the normalized cross-correlation function (NCCF). RAPT has also been proved to provide a good estimate of instantaneous pitch [4]	RAPT version provided in SPTK toolkit [15]
SWIPE [8]	A sawtooth-inspired PDA. It estimates f_0 as that of a sawtooth waveform whose spectrum approximates best that of the input signal	SWIPE version provided in SPTK toolkit [15]
YIN [10]	A popular PDA using a combination of modification of the AC function in order to prevent errors, as described by [12]	MATLAB implementation provided by the respective authors [11]
SHR [38]	A spectral method, based on calculating the subharmonic-to-harmonic ratio and its comparison to a threshold to select the candidate f_0	MATLAB implementation provided by the respective authors [39]
YANG [22]	A PDA based on channel-wise instantaneous frequency. The instantaneous frequency is not obtained by Hilbert transform, but as time derivative of the STFT phase spectrum (cf. (1). Note that YANG is used by Google's open source vocoder [3]	MATLAB implementation provided by the respective authors [3]
Prop. [27]	The proposed algorithm (cf. Algorithms 1–3)	MATLAB code of the proposed algorithm is available at [25]

(cents)) [9, 12]. These standard measures are usually used to assess pitch detection quality. They are calculated as follows [9]:

- V/UV decision error (VDE(%)): The rate of misclassified V/UV decisions, i.e., $\mathrm{V} \rightarrow \mathrm{U}(\%)$ and $\mathrm{U} \rightarrow \mathrm{V}(\%)$, respectively, the rate of voiced frames detected as unvoiced (false negatives) and of unvoiced frames detected as voiced (false positives).

$$\mathrm{VDE}(\%) = \mathrm{V} \rightarrow \mathrm{U}(\%) + \mathrm{U} \rightarrow \mathrm{V}(\%), \tag{13}$$

where

$$\mathrm{V} \rightarrow \mathrm{U}(\%) = \frac{\mathrm{N}_{\mathrm{V}\rightarrow\mathrm{U}}}{\mathrm{N}} \times 100, \tag{14}$$

and

$$\mathrm{U} \rightarrow \mathrm{V}(\%) = \frac{\mathrm{N}_{\mathrm{U}\rightarrow\mathrm{V}}}{\mathrm{N}} \times 100. \tag{15}$$

$\mathrm{N}_{\mathrm{V}\rightarrow\mathrm{U}}$, $\mathrm{N}_{\mathrm{U}\rightarrow\mathrm{V}}$ and N are the number of false negatives, false positives and total frames, respectively.

- Gross pitch error (GPE(%)): The rate of voiced frames that are detected as voiced, where in addition the relative error between ground-truth f_0 and $f_{0,\mathrm{est}}$ is higher than 20%, among the total true positives $\mathrm{N}_{\mathrm{V}\rightarrow\mathrm{V}}$:

$$\mathrm{GPE}(\%) = \frac{\mathrm{N}_{\mathrm{GPE}}}{\mathrm{N}_{\mathrm{V}\rightarrow\mathrm{V}}} \times 100. \tag{16}$$

- f_0 frame error (FFE(%)): The rate of frames concerned by either a VDE error, i.e., $(\mathrm{N}_{\mathrm{V}\rightarrow\mathrm{U}} + \mathrm{N}_{\mathrm{U}\rightarrow\mathrm{V}})$ or by a GPE error, i.e., $\mathrm{N}_{\mathrm{GPE}}$, among all frames:

$$\mathrm{FFE}(\%) = \mathrm{VDE}(\%) + \frac{\mathrm{N}_{\mathrm{V}\rightarrow\mathrm{V}}}{\mathrm{N}} \times \mathrm{GPE}(\%). \tag{17}$$

- Fine pitch error (FPE(cents)): The standard deviation of the relative error (in cents) of pitch values in the voiced frames where there is no gross pitch error:

$$\mathrm{FPE(cents)} = \sqrt{\frac{\sum_{i=1}^{\mathrm{N}_{\mathrm{FPE}}} (\epsilon_i - \overline{\epsilon})^2}{\mathrm{N}_{\mathrm{FPE}}}} \times 100 \tag{18}$$

$\mathrm{N}_{\mathrm{FPE}}$ is the number of true positives for which there is no gross pitch error, ϵ is the absolute error between ground-truth f_0 and $f_{0,\mathrm{est}}$ for such a frame and $\overline{\epsilon}$ is the mean of this error over all concerned frames.

4.2 Evaluation Results

For coherence of error measures, the same values of frame and shift duration used for extraction of ground-truth f_0 were set for all the evaluated algorithms, i.e., 32 ms and 10 ms, respectively. Also, the same f_0 boundaries were set, i.e., [80 Hz, 270 Hz] for male speakers and [120 Hz, 400 Hz] for female ones. Evaluation has been made by

Table 4 Average pitch error measures of benchmarking PDA's for clean speech over all speakers

PDA	VDE (%) ($V \rightarrow U\,(\%) + U \rightarrow V\,(\%)$)	GPE (%)	FFE (%)	FPE (cents)
RAPT [41]	**4.70** (**2.63** + 2.17)	**4.77**	**5.92**	42.78
PRAAT [7]	4.96 (2.89+2.07)	4.96	6.24	42.86
YANG [22]	5.87 (4.56 + **1.31**)	7.61	7.44	39.22
YIN [10]	7.15 (3.18 + 3.97)	7.22	8.77	41.34
SWIPE [8]	7.20 (3.35 + 3.85)	7.20	8.91	41.57
SHR [38]	16.02 (14.34 + **1.68**)	19.47	20.94	41.03
Prop. [27]	7.38 (3.01 + 4.37)	12.78	10.32	**37.39**

Bold characters indicate the best result for each pitch error measure

Table 5 Average pitch error measures of the proposed algorithm for noisy speech over all speakers

Noise	SNR (dB)	VDE (%) ($V \rightarrow U(\%) + U \rightarrow V(\%)$)	GPE (%)	FFE (%)	FPE (cents)
Babble	20	14.99 (4.32 + 10.67)	21.22	19.05	33.59
	15	22.68 (11.72 + 10.96)	26.91	26.87	31.72
	10	27.82 (15.52 + 12.30)	31.75	32.36	30.41
	5	31.64 (16.60 + 15.04)	34.33	35.60	28.35
	0	37.80 (14.37 + 23.43)	37.80	38.99	17.20
White	20	7.31 (0.74 + 6.57)	13.77	10.14	34.18
	15	11.28 (0.37 + 10.91)	17.83	14.07	28.72
	10	16.23 (0.18 + 16.05)	20.96	18.38	23.31
	5	23.08 (0.00 + 23.08)	24.38	23.92	14.53
	0	26.54 (0.00 + 26.54)	26.54	26.54	0.00

measuring VDE, GPE, FFE and FPE rates for clean and noisy speech calculated using (13)–(18) in a systematic way, as follows:

1. Performance for clean vs. noisy speech for all benchmarking PDA's is reported in Tables 4 and 5, respectively.
2. The results are analyzed in different qualitative aspects for all benchmarking PDA's, i.e., by type of noise (cf. Fig. 3) and by gender of speaker, (cf. Fig. 4).
3. The performance of the proposed algorithm is evaluated in respect to its intrinsic parameters, i.e., by the preset maximum order of instantaneous pitch multiples ($H_{\max}$) (cf. Fig. 6) and by the sweeping step of f_0 candidates ($f_{0,step}$) (cf. Fig. 7).

4.2.1 Performance for Clean Versus Noisy Speech

Performance for clean speech: The analysis of Table 4 shows that the proposed algorithm is as good as the state-of-the-art algorithms YIN [10] and SWIPE [8] in V/UV decision detection, i.e., VDE(%), especially thanks to its low rate of false negatives, i.e., (V $\rightarrow$ U(%)). However, its rate of false positives, i.e., U $\rightarrow$ V(%), is slightly higher. Also, it should be noted that the low performance of SHR algorithm[38] is due to the unified frame length and shift imposed to all algorithms during evaluation. Actually, SHR should give better results for shorter frames.

A finer analysis shows the performance of (Prop.) slows down when looking to FFE rate. This should be due to its higher GPE rate, as FFE is a weighted mean of VDE and GPE, cf. (17). Nevertheless, the proposed PDA (Prop.) provides the best FPE in clean speech. This confirms the effect of good voicing detection, since FPE concerns only the true positives where there is no gross pitch error cf. (18). This means also that for a region detected as voiced, if the gross pitch error is less than 20%, then f_0 contour estimated by (Prop.) is closer to the ground truth than all other PDA's.

Performance for noisy speech: Table 5 shows the performance of the proposed PDA (Prop.) in different noise conditions, namely babble and white noise, with SNR ranging from 20 dB to 0 dB. Results show that for both types of noise, (Prop.) is doing well only for low noise levels (SNR $\geq$ 15 dB) whereas for higher noise levels, all rates are less satisfactory. In particular, VDE rate is as good as for clean speech, which proves that the proposed methods succeeds to (a) detect the presence of speech activity, particularly in white noise, and (b) make a distinction between the right voice and other voices in babble noise. Figure 3 shows the measured rates for all PDA's in both babble and white noise. The following remarks can be noticed:

- For SNR $\geq$ 15 dB, most PDA's are performing as well as in clean speech, i.e., PRAAT and RAPT, and in a lesser degree SWIPE, YIN and (Prop.); whereas for higher noise levels, all PDA's performance gets worse.
- For higher noise levels, (Prop.) succeeds to keep an intermediate position, especially for babble noise, whereas the top PDA's like PRAAT and RAPT lose their efficiency.
- For all tested PDA's, FPE rate gets too low at SNR $\leq$ 5 dB (cf. Fig. 3g, h).

Comparison with f_i-based PDA's: YANG is a PDA that has recently been proposed by [22] and utilized in Google's vocoder for speech synthesis. The particularity of this

PDA lies in the fact that is based on another type of instantaneous frequency, i.e., the channel-wise STFT instantaneous frequency (cf. (1)), which makes it interesting to compare it to the proposed approach. For clean speech, Table 4 shows that YANG outperforms the proposed PDA in all metrics except FPE. However, the main difference lies in GPE, which influences also FFE, whereas the difference between both PDA's in VDE is less sensitive.

For noisy speech, Fig. (3a–h) shows that even if YANG does better than the proposed PDA for low noise levels, i.e., SNR $\geq$ 15 dB, its performance decreases for higher levels of both types of noise, i.e., babble and white, whereas the proposed PDA remains more stable. This confirms the robustness of the proposed PDA to high noise levels.

4.2.2 Qualitative Performance Evaluation

Evaluation by type of noise: First, for babble noise (cf. Table 5 and Fig. 3a, c, e, g), the proposed algorithm is among the top PDA's at low noise levels, i.e., SNR $\geq$ 15 dB. This means that it is capable to distinguish the pitch of the right speaker among other voices. Also for high noise levels, i.e., SNR $\leq$ 10 dB , the proposed PDA is ranked among the top ones, even though all benchmarking PDA's are not so efficient.

Secondly, for white noise (cf. Table 5 and Fig. 3b, d, f, h), the proposed algorithm is interestingly efficient for low noise levels, with error rates close to clean speech, cf. Table 5. However, this trend is less maintained when dealing with high noise levels, i.e., SNR $\leq$ 10 dB, where the proposed algorithm is less efficient than some benchmarking PDA's such as RAPT, SWIPE and YIN.

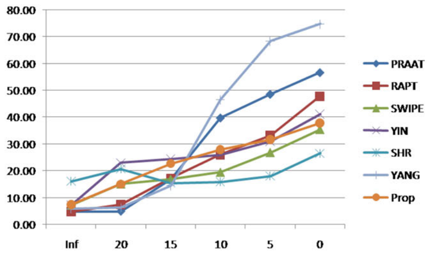
Finally, it is important to note that the low FPE value for SNR $\leq$ 15 dB for all benchmarking PDA's (cf. Fig. 3g, h) is rather caused by the poor V/UV estimation, i.e., a high VDE (cf. Fig. 3a, b) than to a good pitch estimation. Actually, if most of frames are detected as unvoiced, the overall FPE would be calculated only on a few voiced frames, where there is no gross pitch error.

Evaluation by gender of speaker: Figure 4 shows that there is no substantial difference in the performance of all PDA's between male and female speakers. In particular, the proposed algorithm is registering similar levels for each type of error measure for both genders. This means that the parameters are set correctly. In fact, the search range $[f_{0,min}, f_{0,max}]$ and the voicing threshold ($Th1$), both for clean and for noisy speech, depend on the speaker's gender (cf. Table 2).

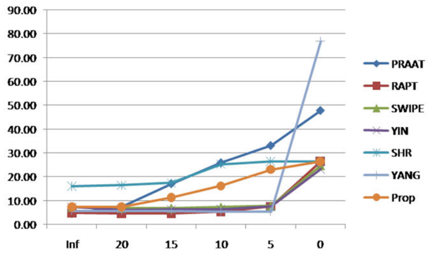
4.2.3 Evaluation of the Proposed PDA in Respect to Its Intrinsic Parameters

Evaluation by the preset maximum order of instantaneous pitch multiples: Figure 6 illustrates the results of the proposed algorithm for different values of the maximum order of instantaneous pitch multiples ($H_{\max}$), for both babble and white-noised speech, and at different SNR levels (from clean speech to a high noise level, i.e., SNR $=$ 0 dB). It is also worth noting that voicing decision error VDE does not depend on $H_{\max}$, since V/UV it is calculated using f_i and its first derivative Δ_{fi} (cf. Algorithm 1), and therefore only GPE and FPE are mentioned in Fig. 6a–d.

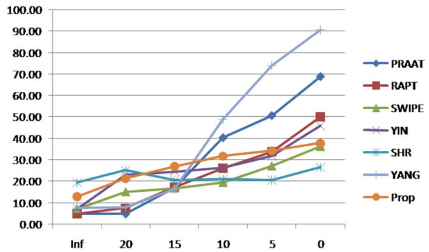
The main observation is that both evaluation metrics, i.e., GPE and FPE, depend more on the level of noise than on the maximum order of instantaneous pitch multiples



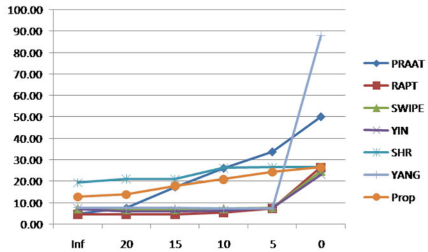
(a) VDE(%) of babble-noised speech



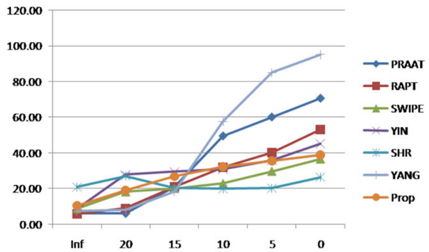
(b) VDE(%) of white-noised speech



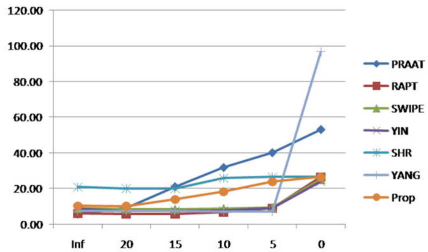
(c) GPE(%) of babble-noised speech



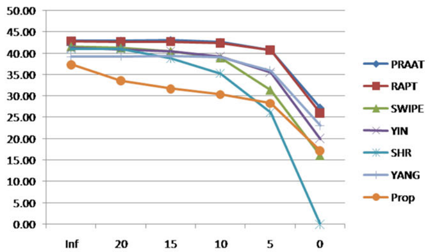
(d) GPE(%) of white-noised speech



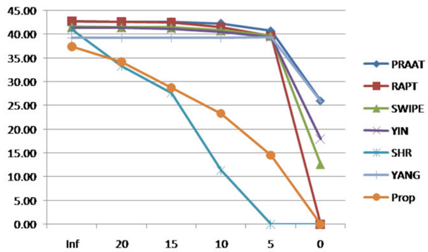
(e) FFE(%) of babble-noised speech



(f) FFE(%) of white-noised speech

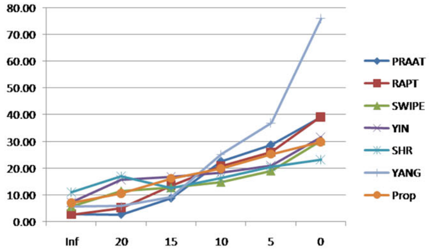


(g) FPE(cents) of babble-noised speech

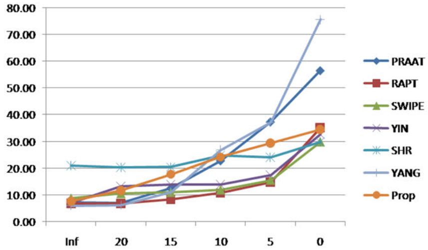


(h) FPE(cents) of white-noised speech

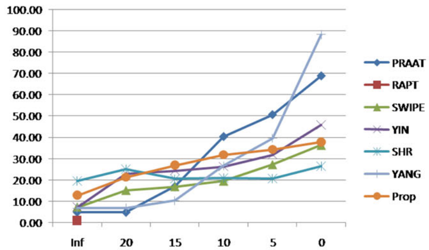
Fig. 3 Performance of benchmarking PDA's by type of noise for all speakers



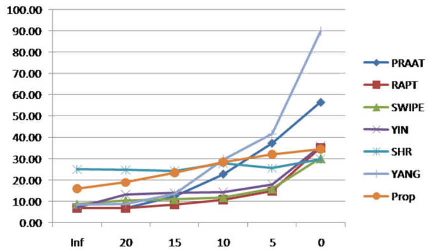
(a) VDE(%) of male speakers



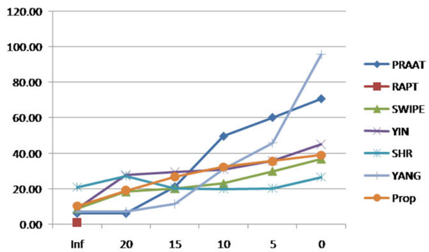
(b) VDE(%) of female speakers



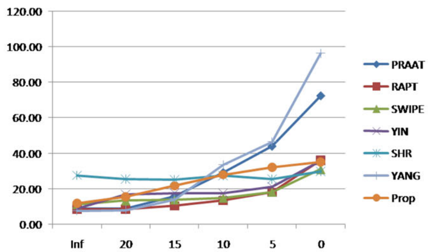
(c) GPE(%) of male speakers



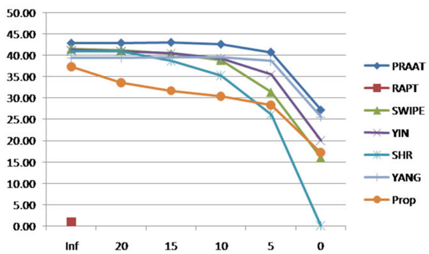
(d) GPE(%) of female speakers



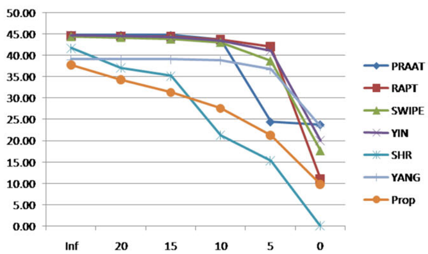
(e) FFE(%) of male speakers



(f) FFE(%) of female speakers



(g) FPE(cents) of male speakers



(h) FPE(cents) of female speakers

Fig. 4 Performance of benchmarking PDA's by gender of speaker for all types of noise

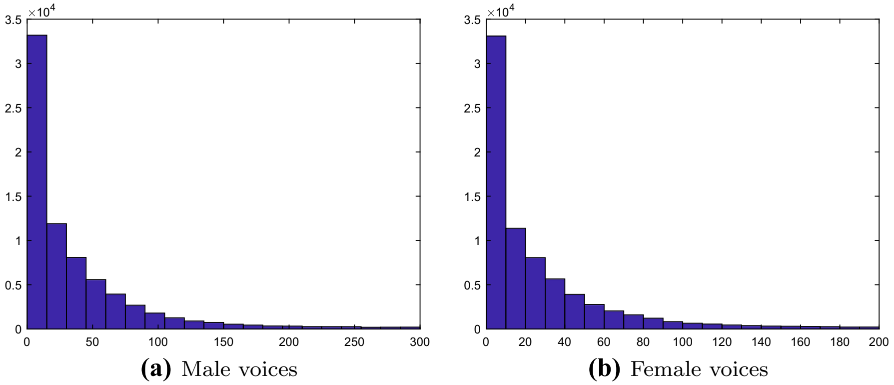
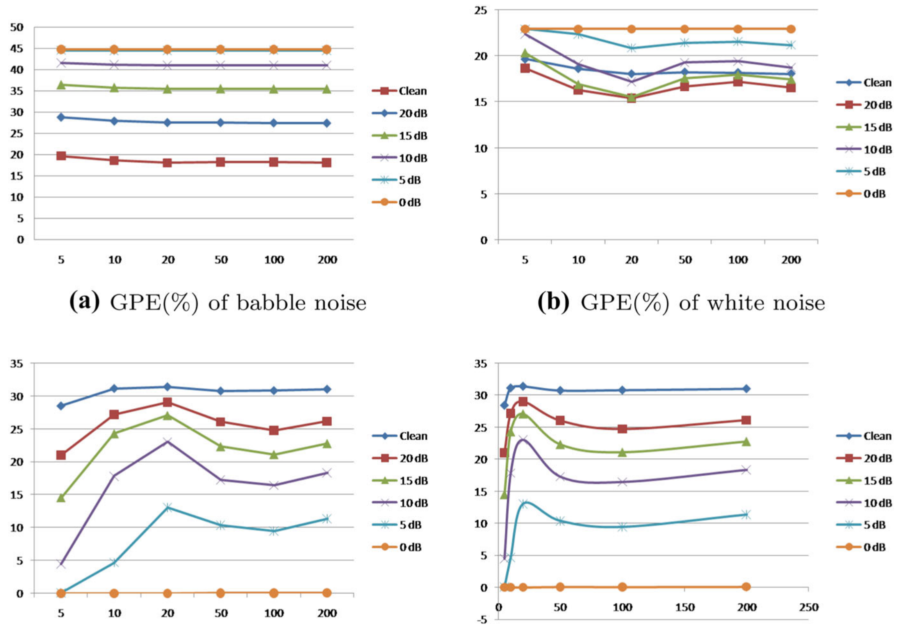


Fig. 5 Distribution of the effective number of pitch multiples $H_{max,m}(n)$ at each time (n) and for each $f_{0,cand}$ candidate value (m)

$H_{\max}$, that is preset as a parameter using values as mentioned in Table 2. Figure 6a–d shows that from $H_{\max} = 50$, the performance of the proposed PDA does not alter remarkably, for any type or level of noise. This can be accounted as an advantage, since setting a low $H_{\max}$ as an upper bound for instantaneous pitch multiples reduces the number of potential f_0 candidates (cf. Algorithm 2), hence reducing significantly the computational load.

Also, we checked out the effective number of pitch multiples used $H_{max,m}(n)$ at each time n and for every $f_{0,cand}$ index $m = 1, \ldots, M$ by setting $H_{\max}$ to the maximum value, i.e., 1000 (cf. Algorithm 2). The histograms shown in Fig. 5, indicate the distribution of the effective $H_{max,m}(n)$ calculated for speech signals uttered by male and female speakers. It is interesting to note that for both genders, the majority of effective $H_{max,m}(n)$ are less than 50, and that for a few cases only, it reaches high values, i.e., more than 100. In particular, for male voice, most values of $H_{max,m}(n)$ are around 15, whereas for female voices, the most common value is 10. Following the equation setting $H_{max,m}(n) = min(H_{\max}, \frac{f_i(n)}{f_{0,cand}(n,m)})$ (cf. Algorithm 2), the obtained histograms confirm that: a) the choice of $H_{\max}$ is not critical if it is high enough, and b) most of values of instantaneous frequency $f_i(n)$ fall in the range of $15 \times f_{0,cand}$ for male voices and $10 \times f_{0,cand}$ for female ones, i.e., in a bandwidth bounded by nearly 4 KHz, if we set $f_{0,max}$ to 270 Hz for male and 400 Hz for female speakers, which is, interestingly, the same bandwidth containing f_0 and the three main formants.

Evaluation by sweeping step of f_0 candidates: Another parameter that may influence the quality of the extracted f_0 contour is the sweeping step ($f_{0,step}$). As explained in Table 2, this parameter defines the precision of f_0 candidate selection within the interval of search $[f_{0,min}, f_{0,max}]$. For each f_0 candidate value, the algorithm decides whether it corresponds to the best fitting f_0 value at each instant n. Therefore, it is mentioned in Table 2 that such a step should be within the interval [0.1 Hz, 2 Hz], so that the tradeoff between the computational load and the precision of the extracted f_0 contour is preserved. Actually, precision standards of f_0 detection would not require less than 0.1 Hz, as pitch variation is not perceptible below 1 Hz. However, a precision



(a) GPE(%) of babble noise **(b)** GPE(%) of white noise

(c) FPE(cents) of babble noise **(d)** FPE(cents) of white noise

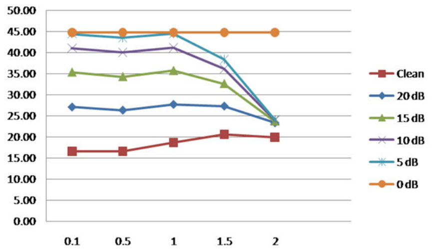
Fig. 6 Performance of the proposed PDA by maximum order of instantaneous pitch multiples ($H_{\max}$) for all speakers and each type of noise

higher than 2 Hz would be perceptible. Figure 7 confirms this trend, since for $f_{0,step}$ within the interval [0.1 Hz,2 Hz], most of GPE and FPE measures are stable. For the same reason as for the maximum order of pitch multiples, only GPE and FPE are mentioned in Fig. 7a–d.

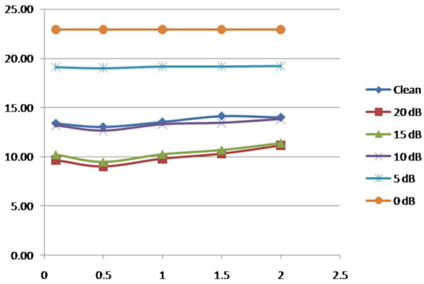
The analysis of these figures shows that a small sweeping step, i.e., $f_{0,step} \leq 1$ Hz, the pitch errors, whether gross, GPE or fine, FPE are smaller for any level of noise. For a bigger step, i.e., $f_{0,step} > 1$ Hz, both pitch errors are decreasing. Nevertheless, this is not due to a better estimation of f_0, but rather to a high rate of GPE, since FPE is calculated only for frames where there is no gross pitch error. An exception is registered for GPE of babble noise (cf. Fig. 7a) which decreases when $f_{0,step}$ increases for 20 dB $\geq$ SNR $\geq$ 5 dB. This is due to the high VDE for noisy speech, which makes many voiced frames classified as unvoiced. Finally, the very low value of FPE for high noise (cf. Fig. 7c, d) cannot be accounted as a positive result since it, as already explained, comes from the high VDE at that noise level (cf. Table 5).

4.3 Discussion

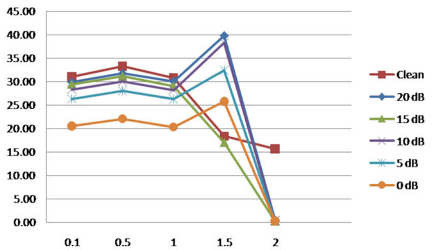
The main advantages and shortcomings of the proposed PDA can be opposed face-to-face as follows, with some proposed solutions.



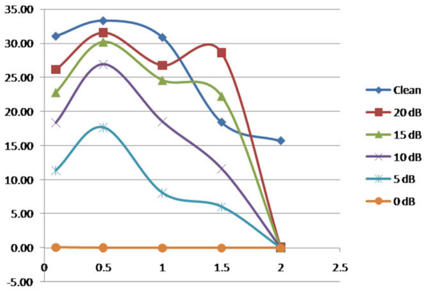
(a) GPE(%) of babble noise



(b) GPE(%) of white noise



(c) FPE(cents) of babble noise



(d) FPE(cents) of white noise

Fig. 7 Performance of the proposed PDA by f_0 candidate selection step ($f_{0,step}$) for all speakers and each type of noise

V/UV decision: It is satisfactory for both clean and noisy speech, at least for low noise levels (SNR $\geq$ 15 dB). This confirms the role of instantaneous frequency to detect periodicity in speech signal. On the other hand, it is not clearly outperforming the top state-of-the-art methods. In particular, it is less efficient in high noise levels (SNR $\leq$ 10 dB), even though this is a common notice for all benchmarking PDA's.

The parametric structure of the proposed algorithm: It allows improving its performance through combining the values of different parameters and thresholds using a grid search. In particular, a small step of f_0 candidates ($f_{0,step}$) and a high order of instantaneous pitch multiples ($H_{\max}$) should improve the overall performance. Nevertheless, this may lead to increasing the computational load, which makes it difficult to run online for some real-time application such as on-the-fly pitch tracking. To cope with such a shortcoming, some solutions can be suggested to reduce the computational load, e.g., by setting the optimal parameters corresponding to type and level of noise, gender of speaker, etc. into a look-up table, or by implementing an adaptive parameter adjustment solution.

The instantaneous frequency: It is computed for each sample along the whole signal, hence there is no need for short-time analysis of the signal, which avoids assuming local stationarity. However, such a sample-wise procedure is computationally heavy, especially for a high sampling rate. An intermediate solution, to keep the tradeoff between temporal and frequency resolution could be subsampling the signal before computing the instantaneous and the fundamental frequencies.

Shortcomings and proposals: The method, as proposed, may be considered as rather heuristic than rigorously theoretic. In fact, while studying this problem, we reviewed the past works/elements that help finding a mathematical proof; however, all what we found were some results in limited cases, as mentioned in Sect. 2.3. Therefore, we believe that, in spite of this limitation, this method may be useful for the following reasons: a) It shows that natural signals, such as speech may have some properties that are still to investigate, to provide more accurate models to represent either the signal itself or its parameters, such as f_0; b) The results obtained by this method, at least during the validation of the proposed relationship (cf. Subsection 3.3.) may hopefully spark the curiosity of the signal processing community in general, and speech/audio processing in particular, to study this problem and to precise if it holds for any type of signals and at which conditions, or if it is just a propriety of a particular class of signals; c) Finally, and in case no explicit mathematical proof could be sorted out for the proposed relationship between f_0 and f_i, machine learning could be an alternative to set a model to create a mapping between both, taking into consideration the particularities of every type of signals.

5 Conclusion

In this paper, a novel pitch detection algorithm was presented. The key idea relies on proposing an empirical relationship between fundamental frequency f_0 and instantaneous frequency f_i. This relationship stipulates that f_0 contour could be approximated as the smoothed envelope of the residual f_i, which is calculated as the rest of the division of the absolute value of f_i by the highest pitch multiples at each instant. The superposition of the so-estimated f_0 and the ground-truth values was verified. Then, an algorithm was implemented based on this relationship, in order to detect voiced/unvoiced regions and then to extract f_0 contour from f_i values in the voiced parts. In comparison to some well-rated state-of-the-art PDA's, the proposed algorithm has been highly successful in taking accurate V/UV decision, and quite satisfactory in approximating f_0 values in voiced parts, either in clean or in simulated noisy speech at low SNR levels.

The proposed algorithm has two major advantages: (a) It does not rely on short-time signal analysis and thus is able to perform instantaneous pitch detection, (b) its parametric structure, which allows its adaption for several considerations, such as type and level of noise, gender of speakers, etc., through fine-tuning its specific parameters, such as $H_{\max}$ and $f_{0,step}$, in addition to its thresholds. Further improvement can be achieved through investigating more in depth the proposed empirical relationship between f_0 and f_i, in order to make it more explainable and interpretative. Finally, the proposed method can be useful for audio and speech signal analysis, reconstruction and synthesis, using an f_i-based vocoder, like in [22]. Besides, it can be extended to other audio and speech applications such as compressive sensing, where only a few amount of data is required to reconstruct the signal.

Acknowledgements This work was supported by the University of Tunis El Manar, Tunisia, and by the University of Genoa, Italy.

Data Availability The datasets analyzed during the current study are available in the PTDB-TUG repository [29]. These datasets were derived from the following public domain resources: https://www2.spsc.tugraz.at/databases/PTDB-TUG/

References

1. T. Abe, T. Kobayashi, S. Imai, Harmonics tracking and pitch extraction based on instantaneous frequency, in *1995 International Conference on Acoustics, Speech, and Signal Processing*, IEEE, vol. 1, pp. 756–759 (1995)
2. T. Abe, T. Kobayashi, S . Imai, Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency, in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. IEEE, vol. 2, pp. 1277–1280 (1996)
3. Y. Agiomyrgiannaki, Yang: Yet-another-generalized vocoder. https://github.com/google/yang_vocoder/, last accessed: 31-05-2022 (2017)
4. E. Azarov, M. Vashkevich, A. Petrovsky, Instantaneous pitch estimation based on rapt framework, in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 2787–2791 (2012)
5. H. Ba, N. Yang, I. Demirkol, W. Heinzelman, Bana: a hybrid approach for noise resilient pitch detection, in *2012 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, pp. 369–372 (2012)
6. B. Boashash, Estimating and interpreting the instantaneous frequency of a signal. II. Algorithms and applications. Proc. IEEE **80**(4), 540–568 (1992)
7. P. Boersma, D. Weenink, Praat: doing phonetics by computer. https://www.fon.hum.uva.nl/praat/, last accessed: 31-05-2022 (2006)
8. A. Camacho, J.G. Harris, A sawtooth waveform inspired pitch estimator for speech and music. J. Acoust. Soc. Am. **124**(3), 1638–1652 (2008)
9. W. Chu, A. Alwan, Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend, in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 3969–3972 (2009)
10. A. De Cheveigné, H. Kawahara, Yin, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. **111**(4), 1917–1930 (2002)
11. A. De Cheveigné, H. Kawahara, Yin algorithm. https://labrosa.ee.columbia.edu/doc/yin.html, last accessed: 31-05-2022 (2002)
12. T. Drugman, A . Alwan, Joint robust voicing detection and pitch estimation based on residual harmonics, in *Proceedings of the Interspeech 2011*, Florence, Italy. IEEE, pp. 1973–1976 (2011)
13. D. Gabor, Theory of communication. Part 1. The analysis of information. J. Inst. Electr. Eng. Part III Radio Commun. Eng. **93**(26), 429–441 (1946)
14. S. Gonzalez, M. Brookes, Pefac: a pitch estimation algorithm robust to high levels of noise. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(2), 518–530 (2014)
15. S.W. Group et al., Speech signal processing toolkit (sptk) version 3.3, https://sourceforge.net/projects/sp-tk//, last accessed: 31-05-2022 (2009)
16. D.J. Hermes, Measurement of pitch by subharmonic summation. J. Acoust. Soc. Am. **83**(1), 257–264 (1988)
17. W. Hess, Manual and instrumental pitch determination, voicing determination, in *Pitch Determination of Speech Signals*. Springer, pp. 92–151 (1983)
18. H. Huang, J. Pan, Speech pitch determination based on Hilbert–Huang transform. Signal Process. **86**(4), 792–803 (2006)
19. N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Tung, H.H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **454**(1971), 903–995 (1998)
20. D. Jouvet, Y. Laprie, Performance analysis of several pitch detection algorithms on simulated and real noisy speech data, in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1614–1618 (2017)

21. S. Kadambe, G.F. Boudreaux-Bartels, Application of the wavelet transform for pitch detection of speech signals. IEEE Trans. Inf. Theory **38**(2), 917–924 (1992)
22. H. Kawahara, Y. Agiomyrgiannakis, H. Zen, Using instantaneous frequency and aperiodicity detection to estimate f0 for high-quality speech synthesis, in *9th ISCA Speech Synthesis Workshop (SSW9)*, ISCA, pp. 221–228 (2016)
23. A. Kissling, R. Kompe, N. Niemann, A. Batliner, Dp-based determination of f0 contours from speech signals, in *Acoustics, Speech, and Signal Processing, 1992. Proceedings. (ICASSP'92)*, IEEE, vol. 1, pp. 1–4 (1992)
24. E. Liflyand, Interaction between the Fourier transform and the Hilbert transform. Acta et Commentationes Universitatis Tartuensis de Mathematica **18**(1), 19–32 (2014)
25. Z. Mnasri, Proposed algorithm, https://github.com/zied-mnasri/f0_IF_model, last accessed: 31-05-2022 (2021)
26. Z. Mnasri, H. Amiri, On the relationship between instantaneous frequency and pitch in speech signals, in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018*. ed. by A. Berton, U. Haiber, W. Minker (TUDpress, Dresden, 2018), pp. 23–29
27. Z. Mnasri, S. Rovetta, F. Masulli, A novel pitch detection algorithm based on instantaneous frequency, in *2021 29th European Signal Processing Conference (EUSIPCO)*, IEEE, pp. 16–20 (2021). http://doi.org/10.23919/EUSIPCO54536.2021.9616047
28. A.M. Noll, Cepstrum pitch determination. J. Acoust. Soc. Am. **41**(2), 293–309 (1967)
29. G. Pirker, M. Wohlmayr, S. Petrik, F. Pernkopf, A pitch tracking corpus with evaluation on multipitch tracking scenario, in *Twelfth Annual Conference of the International Speech Communication Association*. http://doi.org/10.21437/Interspeech.2011 (2011)
30. B. Van der Pol, The fundamental principles of frequency modulation. J. Inst. Electr. Eng. Part III Radio Commun. Eng. **93**(23), 153–158 (1946)
31. L. Qiu, H. Yang, S.N. Koh, Fundamental frequency determination based on instantaneous frequency estimation. Signal Process. **44**(2), 233–241 (1995)
32. L. Rabiner, On the use of autocorrelation analysis for pitch detection. IEEE Trans. Acoust. Speech Signal Process. **25**(1), 24–33 (1977)
33. P. Rengaswamy, K.S. Rao, P. Dasgupta, Songf0: a spectrum-based fundamental frequency estimation for monophonic songs. Circuits Syst. Signal Process. **40**(2), 772–797 (2021)
34. M. Ross, H. Shaffer, A. Cohen, R. Freudberg, H. Manley, Average magnitude difference function pitch extractor. IEEE Trans. Acoust. Speech Signal Process. **22**(5), 353–362 (1974)
35. S. Shimauchi, S. Kudo, Y. Koizumi, K. Furuya, On relationships between amplitude and phase of short-time Fourier transform, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 676–680 (2017)
36. Y. Stylianou, Modeling speech based on harmonic plus noise models, in *International School on Neural Networks, Initiated by IIASS and EMFCSC*. (Springer, 2004), pp. 244–260
37. L. Sukhostat, Y. Imamverdiyev, A comparative analysis of pitch detection methods under the influence of different noise conditions. J. Voice **29**(4), 410–417 (2015)
38. X. Sun, A pitch determination algorithm based on subharmonic-to-harmonic ratio, in *Sixth International Conference on Spoken Language Processing* (2000)
39. X. Sun, Pitch determination algorithm. https://www.mathworks.com/matlabcentral/fileexchange/1230-pitch-determination-algorithm, last accessed: 31-05-2022 (2002)
40. J. Tabrikian, S. Dubnov, Y. Dickalov, Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model. IEEE Trans. Speech Audio Process. **12**(1), 76–87 (2004)
41. D. Talkin, W.B. Kleijn, A robust algorithm for pitch tracking (rapt). Speech Coding Synth. **495**, 518 (1995)
42. L.N. Tan, A. Alwan, Multi-band summary correlogram-based pitch detection for noisy speech. Speech Commun. **55**(7–8), 841–856 (2013)
43. J. Ville, Theorie et application de la notion de signal analytique. Câbles et transmissions **2**(1), 61–74 (1948)
44. K. Wu, D. Zhang, G. Lu, Ipeeh: improving pitch estimation by enhancing harmonics. Expert Syst. Appl. **64**, 317–329 (2016)
45. S.A. Zahorian, H. Hu, A spectral/temporal method for robust fundamental frequency tracking. J. Acoust. Soc. Am. **123**(6), 4559–4571 (2008)