A NOVEL PITCH DETECTION ALGORITHM BASED ON INSTANTANEOUS FREQUENCY

Zied Mnasri[a,b], Stefano Rovetta[a] and Francesco Masulli[a]
[a]DIBRIS, Università degli studi di Genova, Via Dodecaneso 35, 16146 Genova, Italy
[b]Electrical engineering department, ENIT, University Tunis El Manar, BP 37, 1002 Tunis, Tunisia
zied.mnasri@enit.utm.tn, {stefano.rovetta, francesco.masulli}@unige.it

***Abstract*—In this paper, a novel pitch detection algorithm (PDA) is presented. Though pitch detection is a classical problem that has been investigated since the very beginning of speech processing, the proposed algorithm is based on a novel approach relying on a proposed empirical relationship between fundamental frequency (f_0) and instantaneous frequency (f_i). Basically, f_0 is defined for periodic signals only, whereas f_i can be calculated for any type of signals using the Hilbert transform. Notwithstanding this substantial difference, the relationship described in this paper shows some interaction between them, at least empirically. Once this relationship was validated on a large set of speech signals, it has been exploited to implement an algorithm in order to (a) detect voiced parts of speech and (b) extract f_0 contour from f_i pattern in the voiced regions. The obtained results of the proposed method were compared to those of some well-rated state-of-the-art PDA's of different backgrounds, to show that the quality of pitch detection yielded by the proposed approach is quite satisfactory, both in clean and simulated noisy speech.**

***Index Terms*—Pitch detection algorithm (PDA), f_0 contour, instantaneous frequency, voicing decision.**

I. Introduction

Pitch is amongst 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 voicing decision and f_0 contour. Pitch detection is probably the speech processing problem which 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 signal feature. Actually, pitch depends on a variety of parameters, mainly the gender, male or female, the age, young or old, and the type of the language, 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 [1]. In [2], 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 cepstral analysis [3], autocorrelation [4], minimal distance methods [2], and harmonic analysis. 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 [5].

To tackle these issues, another concept has emerged in the last two decades, based on techniques applied along the entire signal. The majority of these techniques are based on the analysis of instantaneous frequency (f_i), which is a theoretic concept. In fact, f_i is, by definition, the time-derivative of the phase of the analytic signal. The latter is a complex signal obtained by Hilbert transform [6]. For discrete signals, f_i is calculated by (1), where $z(n)$ is the associated discrete analytic signal and f_s is the sampling frequency (for $n \geq 1$)

$$f_i(n) = \frac{fs}{4\pi}(\arg(z(n+1)) - \arg(z(n-1))). \qquad (1)$$

Three main pitch detection techniques based on f_i analysis were proposed by [7], [8] and [5], with valuable performance. In spite of the good accuracy of these methods to extract f_0 contour from f_i values, an explicit or a direct relationship is still missing. 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. Therefore, a novel relationship, although still empirical, is proposed in this work, in order to determine the voiced vs. unvoiced parts of the speech signal, and then to extract f_0 contour from f_i values in the voiced parts. Hence, this novel approach is proposed in the aim to improve pitch detection especially in noisy environment, where classical PDA's may be less efficient. Potential applications could vary from intonation change detection to expressive speech recognition in noisy environment.

This work is described as follows: Section II reviews the main f_i-based pitch detection techniques, Section III proposes an empirical relationship between f_i and f_0 in speech signals, and details an algorithm to implement the extraction of f_0 from f_i through this relationship. Section IV shows the main results of the application of this algorithm, in addition to other state-of-the-art PDA's, on a dataset of clean and simulated noisy speech. Finally, the performance measures are commented and discussed.

II. Related work

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

resolution, and (iii) spectral leakage, which is inevitable in multi-resolution analysis [5].

Most of f_i-based methods extract f_0 contour as a continuous function of time (f_0 is considered null in unvoiced segments). In [7], Qiu et al. proceed as follows: First, the harmonics are attenuated using a bandpass filter-bank, then the discrete instantaneous frequency (DIF) is estimated at different scales of the bandpass filter-bank, and finally voicing decision is taken upon certain criteria related to the DIF value (DIF ≤ 50Hz or DIF ≥ 500Hz) or to the variation between neighboring DIF's (Δ(DIF) ≥ 1.4Hz) or to the duration of sustained DIF (whether it is less than 20ms).

In [8], 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 filter-bank with a variable center frequency. The instantaneous frequency, f_i, of each component is considered as the harmonic pattern. Finally, the lowest f_i pattern, i.e. the lowest harmonic, is retained as the f_0 contour [8].

In [5], the well-known 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 voicing decision are estimated, first by filtering all IMF's, where only f_i values between 50Hz and 600Hz are kept, and where f_i values are set to zero if $\Delta fi \geq 100$Hz in a 5ms-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.

The aforementioned f_i-based pitch extraction techniques were successfully compared to the rest of state-of-the-art methods, yielding a very accurate voicing 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. However, none of these methods has been established upon a direct relationship, neither theoretically nor empirically, between f_i and f_0.

III. Method

In this work, firstly an empirical relationship between f_i and f_0 patterns in speech signals is proposed. Secondly, an algorithm is implemented based on this relationship in order to determine the voiced/unvoiced parts, and then to extract f_0 contour from f_i values in the voiced regions.

A. Proposed empirical relationship between pitch and instantaneous frequency

In spite of the absence of a direct relation between f_i and f_0, both quantities 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.

1) Definitions: Starting from the assumption that f_i observed at each instant n carries f_0 and its multiples, some novel notations are proposed in the following.

a) Instantaneous pitch: It is defined as the value of f_0 at every discrete instant n inside the voiced regions only, i.e. where f_0 exists. This is different from conventional PDA's where pitch is usually computed on short overlapping frames, by one value at each frame, and then f_0 contour is obtained by interpolation.

b) 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 highest instantaneous pitch multiple order, denoted $H_{max}(n)$, is defined as:

$$H_{max}(n) = \left\lfloor \frac{|f_i(n)|}{f_0(n)} \right\rfloor . \tag{2}$$

It should be emphasized that in this particular case, we avoided calling such f_0 multiples as harmonics for two major reasons: (a) The notion of harmonics is related to Fourier transform, whereas f_i is obtained from the analytic signal, yielding from Hilbert transform (cf. (1)), (b) To the best of our knowledge, no explicit relationship has been proved between f_0 and f_i, though some interaction may exist in harmonic signals [6], [9].

c) Instantaneous residual frequency: It is defined as the difference between $|f_i(n)|$ and the instantaneous pitch multiple:

$$f_{ir}(n) = |f_i(n)| - H(n)f_0(k) \;\forall\; H(n) \leq H_{max}(n), \tag{3}$$

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

2) Estimation of instantaneous pitch from residual 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 notice empirically that f_0 contour can be obtained as the upper bound of the envelope of f_{ir}. This upper bound is calculated on overlapping frames of short duration (less than 40ms):

$$f_{0,\text{est}}(n_k) = \max_{n_k - \frac{L}{2} \leq l < n_k + \frac{L}{2}} f_{ir}(l), \tag{4}$$

where n_k and L are the center and the length of the k^{th} frame, respectively.

To validate the result given by (4), the ground-truth f_0 values provided by PTDB-TUG database [10] 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 $f_{0_{est}}$ were calculated using (2)-(4) for different preset values of H_{max}. To evaluate the degree of superposition of ground-truth f_0 and the estimated pitch $f_{0_{est}}$, the root mean square error (RMSE) was measured between their respective contours for a large subset of signals from PTDB-TUG database [10]. The choice of this corpus is motivated by its original purpose, i.e. pitch tracking quality

assessment. The test set corresponds to a random selection of 20 phrases, each uttered by 10 male and 10 female speakers, thus yielding 400 signals. The results mentioned in Table I show that increasing the maximum order of instantaneous pitch multiples (H_{max}) in (2) and (3) makes the difference between the contours of ground-truth f_0 and the estimated pitch $f_{0,est}$, calculated by (4), small enough to consider them as superposing. However, the fact of utilizing ground-truth f_0 to calculate $f_{0_{est}}$ (cf. (2)-(4)), implies the existence of a recursive relationship between both of them, so the problem is how to extract $f_{0_{est}}$ directly from the instantaneous frequency f_i, such that it approximates the ground-truth f_0.

TABLE I: Root mean square error (RMSE) between ground-truth f_0 and f_0 contour estimated using (2)-(4) for different preset values of the maximum order of instantaneous pitch mutiples (H_{max})

H_{max}	Mean RMSE (Hz)	Std RMSE (Hz)
5	5907.9	1255.5
10	5638.6	1221.4
20	5123.3	1157.7
50	3716.4	1008.7
100	1867.3	872.9
200	356.5	389.8
500	0.6	0.6
1000	0.6	0.6

B. Proposed pitch detection algorithm [1]

To extract voicing decision (V/UV) and f_0 contour from f_i values using (1)-(4), the following three-stage method is implemented as an algorithm.

a) Stage 1-Preprocessing:

- Initialization

1) Extract f_i from a digital speech signal using (1).
2) Set the range of minimum and maximum f_0 values $[f_{0_{\text{min}}}, f_{0_{\text{max}}}]$, e.g. [80 Hz, 270 Hz] for a male voice and [12Hz, 400Hz] for female one.
3) Set the step of f_0 candidates ($f_{0_{\text{cand}}}$) within the range $[f_{0_{\text{min}}}, f_{0_{\text{max}}}]$, e.g. step $= 0.1$ Hz.

- V/UV decision

4) At each time index $n \geq 1$, calculate the differential instantaneous frequency defined as
$$\Delta f_i(n) = \frac{f_i(n+1) - f_i(n-1)}{2}.$$
5) If $\Delta f_i(n) \geq Th_1$ then the point n is considered as unvoiced. The choice of threshold Th_1 may differ between clean and noisy speech, male and female voices. However, a dynamic setting of Th_1 as the mean value (along the entire signal) of the differential instantaneous frequency Δf_i should give good results.
6) If the ratio of points marked as voiced within a frame is higher than the threshold Th_2 (generally set between 85% and 95%), then the whole frame is marked as voiced.

[1] Matlab code: https://github.com/zied-mnasri/f0_IF_model

b) Stage 2-f_0 extraction:

7) Fix a set of $M > 1$ values of equally-spaced f_0 candidates ranging between $f_{0_{\text{min}}}$ and $f_{0_{\text{max}}}$,
$$f_m = (f_{0_{\text{max}}} - f_{0_{\text{min}}})\frac{(m-1)}{(M-1)} + f_{0_{\text{min}}}, \forall m = 1, .., M.$$
8) Set the maximum order of instantaneous pitch multiples H_{max} to be calculated at each instant n.
9) For each instant n, calculate the vector of the instantaneous pitch multiples orders $1 \leq (H_m)_{m=1,..,M} \leq H_{max}$ for each f_0 candidate value ($f_{0_{cand}}(n,m)$) such that
$$H_{max,m}(n) = \min(H_{max}, \lfloor \frac{|f_i(n)|}{f_{0_{cand}}(n,m)} \rfloor).$$
10) For each f_0 candidate value and each candidate maximum pitch multiple order $H_{max,m}(n)$, calculate the corresponding residual frequency $(f_{ir}(n,m))_{m=1,..,M}$, using (3).
11) Calculate the value of $f_{0_{cand}}(n,\hat{m})$ at instant n such that
$$\hat{m} = \arg \min_{m=1...M}(|f_{ir}(n,m) - f_{0_{cand}}(n,m)|).$$
12) If $|f_{ir}(n,\hat{m}) - f_{0_{cand}}(n,\hat{m})| \leq Th_3$ then $f_{0_{cand}}(n,\hat{m})$ is kept as a potential f_0 value at point n. The threshold Th_3 is the desired tolerance in frequency identification, for instance $0 < Th_3 \leq 1$ Hz.
13) For each set of potential f_0 values kept at time n, i.e. $\{f_{0_{cand}}(n,\hat{m})\}_{\hat{m}=1,...,\hat{M}}$, 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 steps 9)-12), then the kept f_0 candidates are {80 Hz, 90 Hz}. It should be noted that to bypass strict numerical inaccuracies, a kept f_0 candidate value ($f_{0,\text{cand}}(n,\hat{m_2})$ is considered a multiple of a smaller one, $f_{0,\text{cand}}(n,\hat{m_1})$ if $\text{mod}\left(\frac{f_{0,\text{cand}}(n,\hat{m_2})}{f_{0,\text{cand}}(n,\hat{m_1})}\right) < Th_4$, where the threshold Th_4 is small enough ($0 \leq Th_4 \leq 10$).
14) 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.
$$f_0(n) = \arg \min_{\hat{m}=1...\hat{M}}(\text{mod}\left(\frac{|f_i(n)|}{f_{0_{cand}}(n,\hat{m})}\right)).$$

c) Stage 3-Postprocessing:

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

TABLE II: Pitch error measures of clean speech for all speakers

PDA	VDE(%)(V→ U(%)+U→V(%))	GPE(%)	FFE(%)	FPE(cents)
RAPT [11]	**4.70** (**2.63**+2.17)	**4.77**	**5.92**	42.78
PRAAT [12]	4.96 (2.89+2.07)	4.96	6.24	42.86
YIN [13]	7.15 (3.18+3.97)	7.22	8.77	41.34
SWIPE [14]	7.20 (3.35+3.85)	7.20	8.91	41.57
SHR [15]	16.02 (14.34+**1.68**)	19.47	20.94	41.03
Prop.	7.38 (3.01+4.37)	12.78	10.32	**37.39**

IV. Objective evaluation

A. Evaluation protocol

1) Select a random subset from PTDB-TUG [10] containing 400 signals, equally divided between the 10 male and the 10 female speakers of the database, i.e. nearly 10% of the whole database.
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 f_0 contour from the reconstructed signals using state-of-the-art PDA's, namely, RAPT [11], SWIPE [14], both provided in SPTK toolkit [16], YIN [13] and SHR [15], using the Matlab code supplied by their respective authors, and finally the proposed algorithm (Prop.).
4) For each pair of ground-truth and extracted f_0 contours, calculate the standard measures 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 (cents)). These standard measures are usually used to assess pitch detection quality. More details about how to calculate these metrics are in [17].

B. Evaluation results

For coherence of error measures, the same values of frame and hop 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 used, i.e. $[80\text{ Hz}, 270\text{ Hz}]$ for male speakers and $[120\text{ Hz}, 400\text{ Hz}]$ for female ones. Results for clean speech are reported in Table II, whereas Fig. 1a-Fig. 1d illustrate FFE and FPE rates for simulated babble and white noisy speech. It is worth noting that VDE and GPE are not explicitly shown for noisy speech, first for insufficient room, and secondly because FFE is already a weighted average of both rates [17].

1) Comparison to state-of-art PDA's: Globally, the results show that the proposed algorithm (Prop.) is in the second range of PDA's, with YIN [13] and SWIPE [14], after RAPT [11] and PRAAT [12], and clearly outperforming SHR [15]. It should be emphasized that these PDA's have been selected for benchmarking based on their high performance for clean and noisy speech as reported in a recent review [18].

2) Performance for clean speech: TABLE II shows that the proposed algorithm (Prop.) does as well as SWIPE and YIN in detecting voiced/unvoiced regions. i.e. VDE rate, especially thanks to its low rate of false negatives, (V $\rightarrow$ U(%)). Nevertheless, this trend slows down when looking to GPE and consequently to FFE. Finally, TABLE II shows that the proposed PDA (Prop.) provides the lowest FPE.

3) Performance for noisy speech: For noisy speech, the proposed algorithm seems to be amongst the top PDA's for babble noise, particularly at high noise levels, i.e. SNR $\leq$ 15 dB (cf. Fig. 1a, Fig. 1b). However, this trend is less sustained when dealing with white noise, where the proposed algorithm is more efficient only for low noise levels, i.e. SNR ≥ 15dB (cf. Fig. 1c, Fig. 1d).

C. Discussion

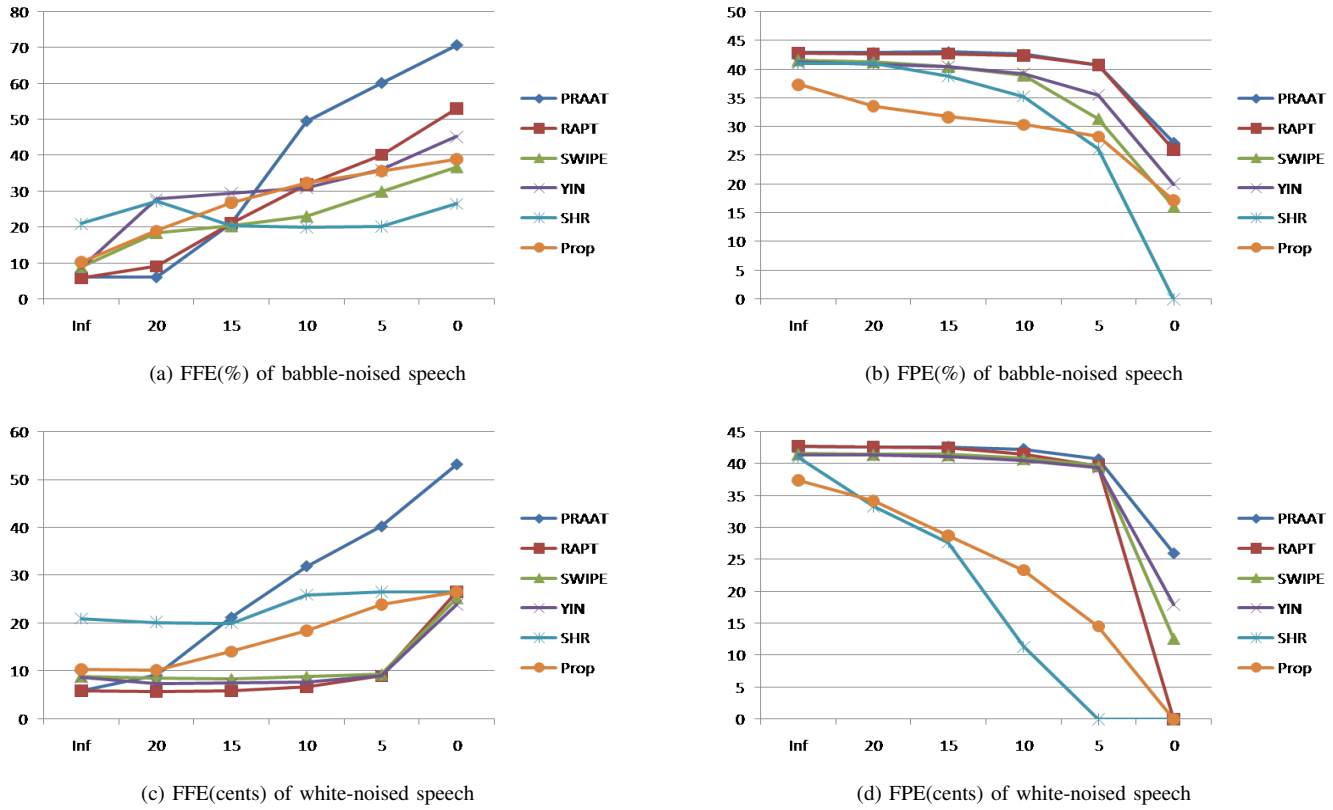
Objective evaluation shows that the proposed algorithm is as good as some recognized state-of-the-art PDA's such as YIN and SWIPE, mainly with a good voicing decision (VDE) and especially with the best fine pitch error (FPE). However, the weakest point of the proposed algorithm consists in its high GPE rate, which results in a relative increase of the FFE rate, even though it remains balanced by the low VDE rate. This may be corrected by tuning the smoothing filter parameters. In opposition, the strongest point is the good FPE rate obtained. This is a twofold advantage since (i) it confirms the good VDE rate, and (ii) it means that when there is no gross pitch error, the $f0$ extracted by (Prop.) is the closest to ground-truth values.

For noisy speech, firstly it looks that the proposed PDA's works better for babble noise than for white noise. This means that it is capable to distinguish the pitch of the right speaker amongst those of other voices. Secondly, for white noise, it seems more adapted to low levels, i.e SNR ≥ 15dB. However, a fine tuning of the algorithm's thresholds may lead to better pitch estimation for higher SNR levels.

Finally, to reduce the computational load of the instant-wise f_0 search, signal subsampling could be a convenient solution.

V. Conclusion

In this paper, a novel pitch detection algorithm was proposed. The key idea relies on a proposed 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 instantaneous frequency (f_{ir}), which is calculated as the rest of the division of the absolute value of f_i by the closest f_0 multiple 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 first detect voiced/unvoiced regions and then 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



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

(b) FPE(%) of babble-noised speech

(c) FFE(cents) of white-noised speech

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

Fig. 1: Performance of the benchmarking PDA's for babble and white noise at SNR levels ranging from Inf (clean) to 0 dB

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 simulated noisy speech.

The proposed algorithm has two major advantages: First, avoiding short-time analysis and thus the underlying approximations about local stationarity; secondly the parametric structure, that makes it possible to adapt pitch detection to several sound conditions, such as type and level of noise, gender of speaker, etc. Finally, investigating more in depth the proposed empirical relationship between f_0 and f_i may lead to make it more explainable and interpretative.

REFERENCES

[1] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," 2011.

[2] W. Hess, "Manual and instrumental pitch determination, voicing determination," in *Pitch Determination of Speech Signals*. Springer, 1983, pp. 92–151.

[3] A. M. Noll, "Cepstrum pitch determination," *The journal of the acoustical society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[4] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE transactions on acoustics, speech, and signal processing*, vol. 25, no. 1, pp. 24–33, 1977.

[5] H. Huang and J. Pan, "Speech pitch determination based on hilbert-huang transform," *Signal Processing*, vol. 86, no. 4, pp. 792–803, 2006.

[6] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. ii. algorithms and applications," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 540–568, 1992.

[7] L. Qiu, H. Yang, and S.-N. Koh, "Fundamental frequency determination based on instantaneous frequency estimation," *Signal Processing*, vol. 44, no. 2, pp. 233–241, 1995.

[8] T. Abe, T. Kobayashi, and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1995, pp. 756–759.

[9] E. Liflyand, "Interaction between the fourier transform and the hilbert transform," *Acta et Commentationes Universitatis Tartuensis de Mathematica*, vol. 18, no. 1, pp. 19–32, 2014.

[10] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[11] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[12] P. Boersma, "Praat: doing phonetics by computer," *http://www. praat. org/*, 2006.

[13] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[14] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[15] X. Sun, "A pitch determination algorithm based on subharmonic-to-harmonic ratio," in *Sixth International Conference on Spoken Language Processing*, 2000.

[16] T. Kobayashi, K. Tokuda, T. Masuko, K. Koishida *et al.*, "Speech signal processing toolkit (sptk), version 3.3," 2009.

[17] W. Chu and 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, 2009, pp. 3969–3972.

[18] D. Jouvet and 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, 2017, pp. 1614–1618.