INSTANTANEOUS PITCH ESTIMATION ALGORITHM BASED ON MULTIRATE SAMPLING

Elias Azarov, Maxim Vashkevich, Alexander Petrovsky

Department of Computer Engineering, Belarusian State University of Informatics and Radioelectronics
6, P.Brovky str., 220013, Minsk, Belarus

ABSTRACT

The paper presents an algorithm for accurate pitch estimation that takes advantage of the sinusoidal model with instantaneous parameters. The algorithm decomposes the signal into subband components, extracts their instantaneous parameters and evaluates period candidate generating function (PCGF). In order to achieve high accuracy for low and high-pitched sounds it is assumed that possible pitch variation range is proportional to current pitch value. The bandwidths of the decomposition filters and length of the analysis frame are scaled for each period candidate by multirate sampling. The algorithm is compared to other widely used pitch extractors on artificial quasiperiodic signals and natural speech. The proposed algorithm shows a remarkable frequency and time resolution for pitch-modulated sounds and performs well both in clean and noisy conditions.

Index Terms— pitch extraction, instantaneous pitch, sinusoidal/harmonic modeling of speech

1. INTRODUCTION

Accurate pitch detection/extraction is required in many speech processing applications. The majority of parametrical models that are applied to coding, morphing and synthesis imply voiced/unvoiced decision and pitch extraction for voiced sounds. Using a pitch detector with fine time resolution is advantageous for analysis/synthesis of unsteady voiced sounds that usually occur at the borders of voiced segments and transitions. Accurate estimation of pitch contour is crucial for pitch-synchronous frequency analysis [1]. The notion of instantaneous frequency can be naturally applied to pitch assuming that the signal is represented by harmonic model [2,3]. Considering voiced speech as a continuous time-varying process it is possible to extract instantaneous pitch as was shown in [4,5].

1.1. Relation to prior work

The idea of fine pitch detection developed in [6,7] relies on decomposition of the signal into narrow-band components and using their instantaneous frequencies as initial data. The idea proved to be useful for speech and singing voice applications [8,9] and designing robust pitch estimation algorithms [10]. However, there is a fundamental limitation of the approach, which originates from the uncertainty principle: it is impossible to provide equally high resolution over the whole pitch range using a filter bank with fixed parameters. Performing time-frequency transformation on long analysis frames is beneficial for low-pitched sounds while using short analysis frames and wider frequency bands is beneficial for high-pitched sounds. The compromise choice made in [7] (50ms frames, 70Hz bands) provided accurate tracking of female voices, however turned out to be prone to gross errors for low male voices [11].

The present paper focuses on developing a fine pitch extraction algorithm based on a multirate analysis scheme. The main idea is achieving improved performance by adjusting parameters of the analysis filter bank for each period candidate. The algorithm provides appropriate processing of pitch-modulated sounds for high and low-pitched sounds and assumes that possible pitch variation is proportional to current pitch value. Widening bandwidths for short period candidates results in mixing of harmonic components for low voices. To compensate the effect a special candidate generation function is proposed that is less sensitive to harmonic mixing. The paper includes theoretical considerations, implementation outline and the experimental performance evaluation in clean and noisy conditions.

2. FEATURE EXTRACTION

The proposed pitch estimator is based on the sinusoidal model which represents deterministic part of the signal as a sum of periodic components with time-varying parameters:

$$s(n) = \sum_{k=1}^{K} A_k(n)\cos(\varphi_k(n)) + r(n) \qquad (1)$$

where $\varphi_k(n) = \sum_{i=1}^{n} \omega_k(n) + \varphi_k(0)$, K – number of periodic components and $r(n)$ – residual part. Parameters of the model (instantaneous amplitude $A_k(n)$ and frequency $\omega_k(n)$ in radians per sample) are used as initial data for pitch estimation. In order to extract them the signal is decomposed into complex subband components by a DFT-modulated filter bank which is briefly described below.

Let us specify a uniform frequency grid that corresponds to center frequencies of the channels as $k\omega_{step}$, $k = 1,2,\dots,K$, $K \le \pi/\omega_{step}$ where ω_{step} is frequency step in radians per sample. Impulse response of each channel of the filter bank is given by the following equation:

$$h_k(n) = 2\frac{\sin(\omega_{bw} n)}{\pi n} w(n) e^{jkn\,\omega_{step}} \quad (2)$$

where ω_{bw} – half of the bandwidth and $w(n)$ – an even window function.

The output of each channel is an analytical band-limited signal $S_k(n)$ that can be expressed as convolution of the input signal $s(n)$ with the impulse response:

$$S_k(n) = \sum_{i=-\infty}^{\infty} h_k(i) s(n-i). \quad (3)$$

Instantaneous parameters of subband components are available from the following expressions:

$$A_k(n) = \sqrt{R^2(n) + I^2(n)}, \quad (4)$$

$$\varphi_k(n) = \arctan\left(\frac{-I(n)}{R(n)}\right), \quad \omega_k(n) = \varphi_k'(n), \quad (5)$$

where $R(n)$ and $I(n)$ are real and imaginary parts of $S_k(n)$ respectively. To avoid phase discontinuities in (5) phase is unwrapped.

Instantaneous parameters are used as initial data for PCGF evaluation. Assuming that possible pitch variation is proportional to the current pitch value, parameters of the filter bank should be scaled for each period candidate:

$$\omega_{step}(\omega_0) = \omega_0, \quad \omega_{bw}(\omega_0) = \alpha\omega_0 \quad (6)$$

where ω_0 – frequency of the candidate in radians per sample and α – allowed relative pitch variation. Time duration of the analysis frame should be adjusted accordingly and contain a fixed number of periods:

$$N = 2\pi L/\omega_0, \quad (7)$$

where N – number of samples and L – number of periods in the frame. The idea is illustrated in figure 1.

Changing parameters of the filter bank for each period candidate is computationally expensive. Alternatively it is possible to use a filter bank with fixed parameters and change sampling frequency of the signal. Let us specify sampling frequency F_s as a multiple of the frequency of the period candidate:

$$F_s = Rf_0\,, \quad (8)$$

where f_0 – frequency of the candidate in Hz and R – integer factor. Using (8) and considering that $f_0 = \omega_0 F_s/(2\pi)$ eq. (7) results in fixed analysis frame length for all period candidates:

$$N = RL\,. \quad (9)$$

Factor R determines how many harmonic are retained in the resampled signal:

$$K = \begin{cases} (R-1)/2\,, & \text{for odd } R \\ R/2 - 1, & \text{for even } R. \end{cases} \quad (10)$$

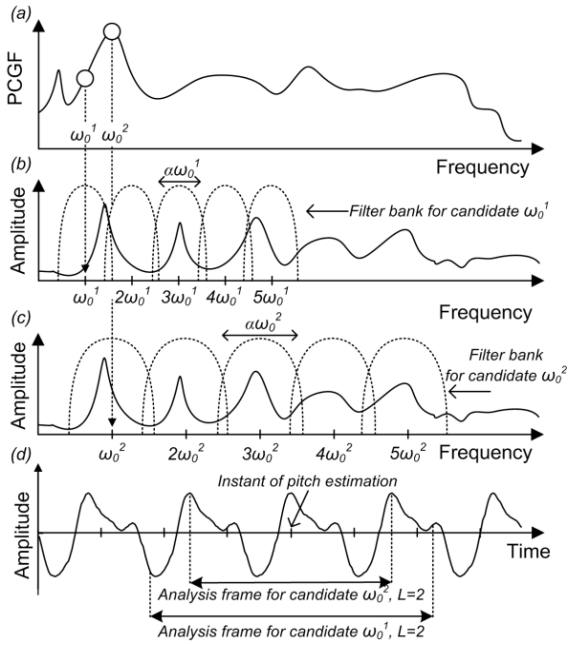

Figure 1 – Adjusting analysis filter bank to each period candidate. (a) – period candidate generation function, (b) – amplitude spectrum and filter bank for candidate ω_0^1, (c) – filter bank for candidate ω_0^2, (d) – source signal

Considering that only a few harmonics are of practical importance it is possible to use very short analysis frames. Assuming (8) the parameters of the filter bank become

$$\omega_{step} = 2\pi/R, \ \omega_{bw} = \alpha\omega_{step} \quad (11)$$

The desired parameters are extracted from the signal using the multirate sampling scheme as shown in figure 2.

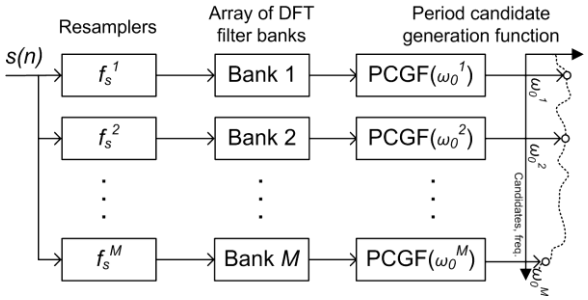

Figure 2 – Multirate sampling scheme for pitch period candidates generation (M – number of period candidates)

3. PERIOD CANDIDATE GENERATION FUNCTION

For period candidates generation an autocorrelation-based measure is used. In [12] was introduced the normalized cross-correlation function (NCCF)

$$\phi(l) = \frac{\sum_{i=0}^{N+l-1} s(n)s(n+l)}{\sqrt{e(0)e(l)}}, \quad (12)$$

where l – lag in samples, $e(l) = \sum_{i=l}^{N+l-1} s^2(i)$. The function averages data within analysis frame and gives smoothed

values. In order to improve time resolution the instantaneous model-based version of the function can be used [7]:

$$\phi_{inst}(n,l) = \frac{\sum_{k=1}^{K}[A_k(n)]^2 \cos(\omega_k(n)l)}{\sum_{k=1}^{K}[A_k(n)]^2}. \quad (13)$$

The function assumes that the bandwidth of each analysis filter is narrower than the minimum possible pitch value and therefore the harmonics are always separated. The multirate analysis scheme, described above, suffers from harmonic mixing (that occurs for high frequency candidates when processing low-pitched sounds) which causes sporadic amplification of high frequency regions of $\phi_{inst}()$. In order to reduce the impact of harmonic mixing the following period candidate generating function is used that multiplies obtained measures of $2V+1$ adjacent samples:

$$\phi_{ms}(n,l) = \prod_{v=-V}^{V}\sum_{k=1}^{K} A_k(n+v)\cos(\omega_k(n+v)l). \quad (14)$$

For each lag l the model parameters A_k and ω_k are estimated on a separate channel of the multirate scheme with sampling factor $\frac{R}{l}$ according to (8). Energy of each resampled frame is normalized to 1 in order to equalize values of $\phi_{ms}()$ for different candidates. Using non-squared amplitudes in (14) instead of squared amplitudes in (13) generally is more robust since contribution of harmonic amplitudes becomes more balanced. Normally the effect of harmonic mixing emerges on short time periods and can be significantly reduced by multiplication of just a few terms. Figure 3 shows period candidates, generated by $\phi()$, $\phi_{inst}()$ and the proposed function $\phi_{ms}()$ for a short speech fragment. Function $\phi_{ms}()$ clearly provides much higher frequency and time resolution compared to $\phi()$ and $\phi_{inst}()$.

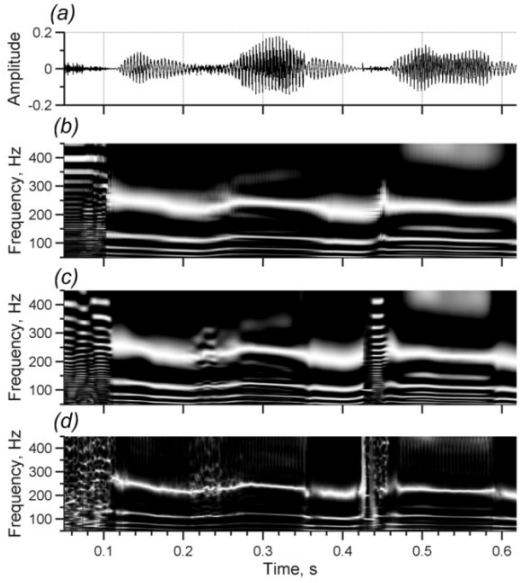

Figure 3 – Period candidates generation. (a) – source signal, (b) – NCCF $\phi()$, (c) –model-based NCCF $\phi_{inst}()$, (d) – proposed function $\phi_{ms}()$ ($V=1$)

4. PITCH ESTIMATION ALGORITHM

The proposed algorithm consists of the following steps[1]:

1) resample the input signal frame $s(n)$ for each period candidate using corresponding sampling rate (8);

2) normalize energy of each resampled frame to 1;

3) estimate instantaneous harmonic parameter using equations (2)–(5); this step is implemented using $2V+1$ overlapping discrete Fourier transforms for each resampled frame; the Hamming widow is used as window function $w(n)$ in (2);

4) evaluate (14) for each period candidate from the correspondent set of parameters;

5) multiply obtained period candidate values with a weighting window, penalizing low-frequency candidates: $w_{wight}(\omega_0) = 0.2\frac{\omega_0}{\pi} + 0.8$;

6) find the best locally continuous contour maximizing total period candidate values for adjacent frames with dynamic programming; this step results in selection of the best current candidate $\omega_{0,best}(n)$ – a rough pitch estimate;

7) calculate fine pitch value $\omega_{0,fine}(n)$ using instantaneous harmonic parameters extracted for the best candidate using weighted sum:

$$\omega_{0,fine}(n) = \frac{1}{\sum_{k=1}^{K}A_k(n)}\sum_{k=1}^{K}\frac{1}{k}\omega_k(n)A_k(n) \quad (15)$$

Considering that the filter bank is implemented using the fast Fourier transform overall computational complexity of the algorithm (multiplications per one pitch estimate) can be approximately expressed as $O(IKN + 2(V+1)KNlog(N))$, where I – length of the low-pass interpolation filter used for resampling.

For practical implementation of the algorithm we used the following values $K=8$, $L=4$, $R=2K+1=17$, $N=68$, $M=100$, $I=121$, $V=1$. The allowed pitch range is between 50 and 450Hz. The pitch range is separated uniformly in logarithmic scale into 100 points each corresponding to one candidate. Duration of resampled frames vary from 80ms (the longest period candidate) to 9ms (the shortest period candidate).

5. SIMULATION RESULTS

We evaluated five known pitch extractors RAPT [12], YIN [13], SWIPE' [14], IRAPT [7], PEFAC [15] and the proposed algorithm (denoted as 'Halcyon') in terms of gross pitch error (GPE) and mean fine pitch error (MFPE). GPE was calculated as percentage of voiced frames with estimated pitch error higher than $\pm 20\%$ of the true pitch value, for MFPE calculation gross errors were omitted.

In order to explore time resolution of the algorithms and their robustness against pitch variations we synthesized

[1] We call it 'halcyon'. Matlab implementation of the algorithm can be found at http://dsp.tut.su/halcyon.html

artificial signals with changing pitch in the range from 100 to 350 Hz. All obtained measurements were separated into six groups distinguished by variation rate: 0–0.3, 0.3–0.6, 0.6–0.9, 0.9–1.2, 1.2–1.5, >1.5 percent of pitch change per millisecond. Averaged errors are shown in figure 4.

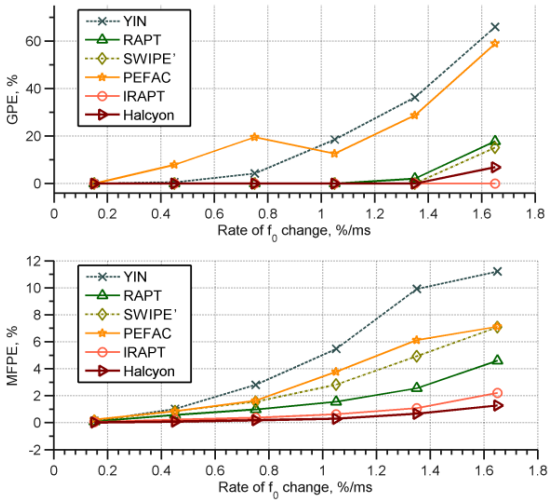

Figure 4 – Pitch estimation error (synthetic signals)

IRAPT and Halcyon showed very high robustness against pitch variations – percentage of gross pitch errors is negligible up to 1.5% of pitch change per millisecond. According to MFPE values the proposed algorithm exceeds other competitors in terms of time/frequency resolution. An example of analysis of an artificial signal with rapid pitch change is given in figure 5. IRAPT, SWIPE' and Halcyon approach the actual contour very closely, while other pitch estimators are not that accurate.

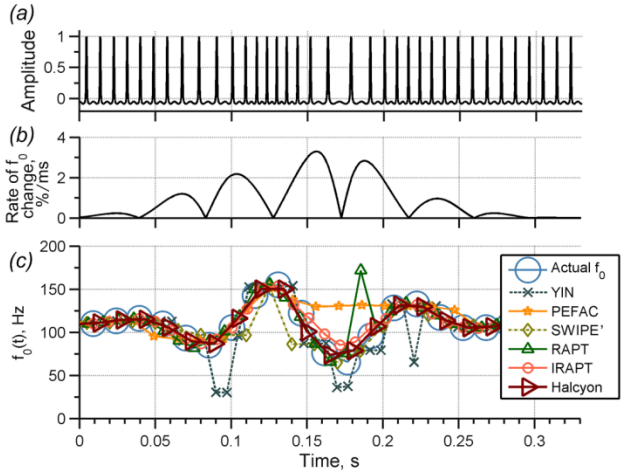

Figure 5 – Analysis of a signal with pitch variations.
(a) – source signal, (b) – rate of pitch change,
(c) – actual and extracted pitch contours

For natural speech experiments the PTDB-TUG speech database [16] was used. Obtained averaged results for clean speech are given in table 1. Compared to IRAPT 1 the proposed algorithm provides two times smaller GPE for low-pitched sounds due to multirate analysis scheme.

	Male speech		Female speech	
	GPE%	MFPE%	GPE%	MFPE%
RAPT	3.687	1.737	6.068	1.184
YIN	3.184	1.389	3.960	0.835
SWIPE'	0.756	1.505	4.273	**0.800**
PEFAC	20.521	1.383	31.192	0.972
IRAPT 1	1.625	1.608	3.777	0.977
Halcyon	**0.743**	**1.268**	**3.600**	1.039

Table 1 – Pitch estimation error (natural speech)

Noisy test samples were generated using two types of noise (white and babble) with SNRs from -20 to 20dB. The averaged results for noisy samples are shown in figures 6,7.

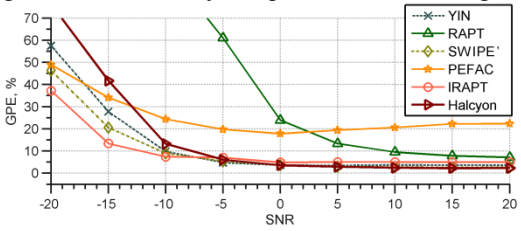

Figure 6 – Pitch estimation error (additive white noise)

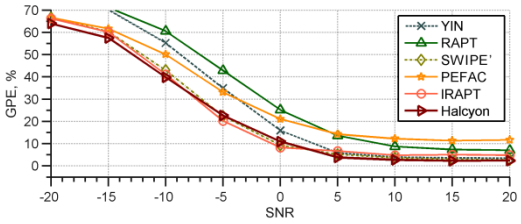

Figure 7 – Pitch estimation error (additive babble noise)

For white noise all algorithms except RAPT do not degrade much up to -10dB SNR. For babble noise all algorithms degrade rapidly at -0dB SNR. In whole the proposed algorithm showed decent robustness against additive noise, considering that it extracts instantaneous frequency and uses very short analysis frames (up to 9ms) for high frequency pitch candidates.

6. CONCLUSIONS

The paper presents a model-based pitch estimation algorithm that can be advantageous for fine harmonic modeling of stable and unsteady voiced speech sounds. The algorithm decomposes the signal into narrow-band components which are represented in terms of instantaneous sinusoidal parameters. The analysis filter bank is scaled for each period candidate providing accurate frequency analysis for low and high-pitched sounds. According to experiments, the algorithm is robust to pitch modulations and provides high frequency and time resolution.

7. ACKNOWLEDGEMENTS

This work was supported by ITForYou company and Belarusian Republican Foundation for Fundamental Research (grant No F14MV-014).

8. REFERENCES

[1] F. Zhang, G. Bi, Y. Q. Chen, "Harmonic transform," in *Vision, Image and Signal Processing, IEE Proceedings*, vol.151, no.4, pp.257–263, 2004.

[2] R. J. McAulay, T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[3] J. Laroche, Y. Stylianou, E. Moulines, "HNS: Speech modification based on a harmonic+noise model," in *ICASSP-93 – IEEE International Conference on Acoustic, Speech, and Signal Processing, April 27-30, Minneapolis, USA, Proceedings*, 1993. – pp. 550–553.

[4] J. O. Hong, P. J. Wolfe, "Model-based estimation of instantaneous pitch in noisy speech," in *INTERSPEECH 2009 – 10th Annual Conference of the International Speech Communication Association, September 6–10, Brighton, UK, Proceedings*, 2009. – pp. 112–115.

[5] B. Resch, M. Nilsson, A. Ekman and W. B. Klejin "Estimation of the Instantaneous Pitch of Speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 15, pp. 819–822, 2007.

[6] T. Abe, T. Kobayashi, S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," in *ICASSP-95 – IEEE International Conference on Acoustic, Speech, and Signal Processing, May 9-12, Detroit, USA, Proceedings*, 1995. – pp. 756–759.

[7] E. Azarov, M. Vashkevich, A. Petrovsky, "Instantaneous pitch estimation based on RAPT framework," in *EUSIPCO'12 –European Signal Processing Conference, August 27-31, Bucharest, Romania, Proceedings*, 2012. – pp. 2787–2791.

[8] E. Azarov, M. Vashkevich, A. Petrovsky, "Instantaneous harmonic representation of speech using multicomponent sinusoidal excitation," in *INTERSPEECH 2013 –14th Annual Conference of the International Speech Communication Association, August 25–29, Lyon, France, Proceedings*, 2013. – pp. 1697–1701.

[9] E. Azarov, M. Vashkevich, A. Petrovsky, "Guslar: A framework for automated singing voice correction," in *ICASSP-2014 – IEEE International Conference on Acoustic, Speech, and Signal Processing, May 4-9, Florence, Italy, Proceedings*, 2014. – pp. 7919–7923.

[10] K. Hotta, K. Funaki, "On a Robust F_0 Estimation of Speech based on IRAPT using Robust TV-CAR Analysis," in *APSIPA 2014 – Annual Summit and Conference Asia-Pacific Signal and Information Processing Association, 2014, December 9–12, Siem Reap, Cambodia, Proceedings*, 2014. – pp. 1–4.

[11] E. van den Berg, B. Ramabhadran, "Dictionary-based pitch tracking with dynamic programming," in *INTERSPEECH 2014 –15th Annual Conference of the International Speech Communication Association, September 14–18, Singapore, Proceedings*, 2014. – pp. 1347–1351.

[12] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)" in "Speech Coding & Synthesis," *W B Kleijn, K K Paliwal eds, Elsevier* ISBN 0444821694, 1995.

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

[14] A. Camacho, J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 1638–1652, 2008.

[15] S. Gonzalez, M. Brookes, "PEFAC – A Pitch Estimation Algorithm Robust to High Levels of Noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no.2, pp. 518–530, 2014.

[16] G. Pirker, M. Wohlmayr, S. Petrik, F. Pernkopf "A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario," in *INTERSPEECH 2011 –12th Annual Conference of the International Speech Communication Association, August 28–31, Lyon, France, Proceedings*, 2011. – pp. 1509–1512.