INSTANTANEOUS PITCH ESTIMATION BASED ON RAPT FRAMEWORK

Elias Azarov[1], Maxim Vashkevich[1] and Alexander Petrovsky[2]

[1]Computer engineering department, Belarusian State University of Informatics and Radioelectronics
6, P.Brovky str., 220013, Minsk, Belarus
[2]Department of Digital Media and Computer Graphics, Bialystok University of Technology,
15-351 Białystok, Poland
email: a.petrovsky@pb.edu.pl

ABSTRACT

The paper presents a pitch estimation technique based on the robust algorithm for pitch tracking (RAPT) framework. The proposed solution provides estimation of instantaneous pitch values and is not sensitive to rapid frequency modulations. The technique utilizes a different period candidate generating function based on instantaneous harmonic parameters. The function is similar to normalized cross-correlation function, however it represents momentary periodicity with high frequency resolution. The second major revision of RAPT is an additional post-processing procedure that makes estimation values more accurate. The proposed algorithm is compared with other pitch detection algorithms using artificial and natural signals.

Index Terms— RAPT, instantaneous pitch estimation, instantaneous harmonic parameters

1. INTRODUCTION

Parametric representation of speech often implies pitch contour as a part of the model. Choosing an algorithm for pitch estimation is always a tradeoff between time and frequency resolution, robustness, delay and computational complexity. The present paper addresses speech processing applications where the most accurate pitch estimates are required and where pitch frequency is considered as a continuous function of time. Any application that uses hybrid deterministic/stochastic representation of wideband speech (such as parametric speech coding, speech synthesis or voice conversion) benefits from minimization of pitch estimation error. Pitch estimation accuracy determines how many high-frequency harmonics can be extracted from speech. Two different characteristics constitute the notion of accuracy in the context of speech processing: 1 – time resolution i.e. how fast the estimator can detect pitch changes, 2 – frequency resolution i.e. how small pitch variances can be detected. Both are sensitive to pitch modulations and noise (either background noise or speech with mixed excitation). Some original algorithms for instantaneous pitch estimation have been proposed recently [1-2]. In the present work RAPT [3] was chosen for the following reasons:

- RAPT is an extensively tested algorithm with known benefits and drawbacks;
- RAPT ensures robust pitch estimation in noisy conditions, relatively low delay and low computational cost;
- RAPT framework is widely used in speech applications and has some free open source implementations.

The main idea behind the proposed algorithm is as follows. It is possible to perform instantaneous pitch estimation with RAPT by using an instantaneous period candidate generating function instead of normalized cross-correlation (NCCF). Such function can be estimated from instantaneous harmonic parameters of speech in a way that provides much higher time and frequency resolution. Further improvement of pitch estimation accuracy is achievable in post-processing phase for the cost of additional computational complexity and time delay.

The paper presents essential theoretical considerations, optimized implementation outline and the results of practical experiments with artificial and natural speech signals. The robustness of the proposed algorithm against noise is also evaluated.

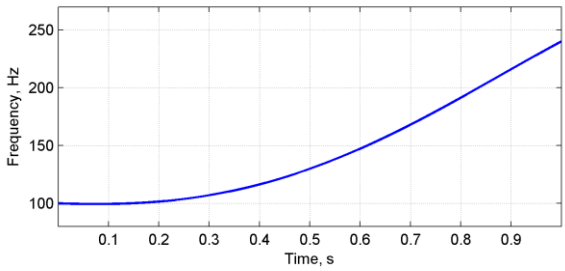
2. INSTANTANEOUS PERIOD CANDIDATE GENERATING FUNCTION

RAPT estimates overall periodicity of the analysis frame using NCCF function. Let $s(m)$ be a speech signal, z – step size in samples and n – window size. The NCCF $\phi(x,k)$ of K samples length at lag k and analysis frame x is defined as [3]:

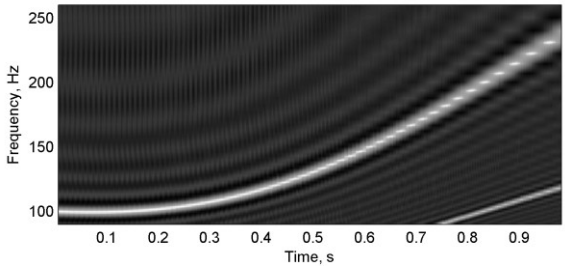
$$\phi(x,k) = \frac{\sum_{i=m}^{m+n-1} s(i)s(i+k)}{\sqrt{e_m e_{m+k}}}, \quad (1)$$

$$k = 0, K-1;\ m = xz;\ x = 0, M-1,$$

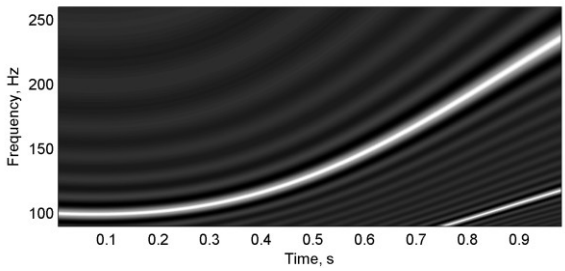
where $e_i = \sum_{l=i}^{i+n-1} s_l^2$.



a – pitch contour



b – NCCF



c – instantaneous version of NCCF

Figure 1 - Period candidate generating functions of a quasiperiodic signal sampled at 6kHz

Let us assume that the signal $s(m)$ can be presented as a sum of P harmonic components with instantaneous parameters (amplitude $A_p(m)$, frequency $F_p(m)$ and phase $\varphi_p(m)$):

$$s(m) = \sum_{p=1}^{P} A_p(m)\cos\varphi_p(m), \tag{2}$$

where $\varphi_p(m) = \sum_{i=0}^{m} F_p(i) + \varphi_p(0)$ and $F_p \in [0, \pi]$ (π corresponds to the Nyquist frequency).

Instantaneous harmonic parameters give spectral representation of the current sample $s(m)$ that can be utilized in order to estimate momentary autocorrelation function $R_{inst}(m,k)$. Using the Wiener-Khintchine theorem:

$$R_{inst}(m,k) = \frac{1}{2}\sum_{p=1}^{P} A_p^2(m)\cos\big(F_p(m)k\big). \tag{3}$$

$R_{inst}(m,k)$ corresponds to the autocorrelation function calculated on infinite window of periodic signal generated with specified harmonic parameters. As far as analysis window is infinite there is no difference between the autocorrelation function and NCCF [3]. Considering this fact it is possible to propose the instantaneous version of NCCF $\phi_{inst}(m,k)$ in the following form:

$$\phi_{inst}(m,k) = \frac{\sum_{p=1}^{P} A_p^2(m)\cos(F_p(m)k)}{\sum_{p=1}^{P} A_p^2(m)}. \tag{4}$$

It should be pointed out that unlike NCCF lag k in (4) does not need to be an integer, valid values can be produced for any desired frequency. The second feature worth mentioning is that the proposed function is immune to any rapid frequency modulations in the neighborhood of m provided that estimated instantaneous harmonic parameters are accurate enough. Figure 1 shows that NCCF function is prone to "stair-case effect" while its instantaneous version produces a smooth continuous pitch candidate contour.

3. ESTIMATION OF INSTANTANEOUS HARMONIC PARAMETERS

The instantaneous harmonic parameters can be estimated using the complex pass-band filter [4] with impulse response

$$h_{F_1,F_2}(n) = 2\frac{\sin(F_\Delta n)}{n\pi}w(n)e^{jF_c n}. \tag{5}$$

where F_1 and F_2 are low and high cut-off frequencies respectively, $F_\Delta = \frac{F_2 - F_1}{2}$, $F_c = \frac{F_1 + F_2}{2}$ and $w(n)$ – an even window function.

The output of the filter is an analytical band-limited signal $S_{F_\Delta,F_c}(m)$ that can be expressed as the convolution of the input signal $s(m)$ with the impulse response:

$$S_{F_\Delta,F_c}(m) = 2\sum_{n=-\infty}^{\infty}\frac{\sin(F_\Delta n)}{n\pi}w(n)s(m+n)e^{-jF_c n}. \tag{6}$$

Instantaneous parameters are available from the following expressions:

$$A_{F_\Delta,F_c}(m) = \sqrt{R^2(m) + I^2(m)}, \tag{7}$$

$$\varphi_{F_\Delta,F_c}(m) = \arctan\left(\frac{-I(m)}{R(m)}\right), \tag{8}$$

$$F_{F_\Delta,F_c}(m) = \varphi'_{F_\Delta,F_c}(m), \tag{9}$$

where $R(m)$ and $I(m)$ are real and imaginary parts of $S_{F_\Delta,F_c}(m)$ respectively. In (9) the unwrapped phase is used to avoid phase discontinuities of π.

It is clear that (6) can be calculated using a fast Fourier transform (FFT) algorithm given that the central frequencies of pass-band filters are uniformly distributed from zero to the Nyquist frequency. This fact will be useful in the next section.

The choice of cut-off frequencies is defined by the following conditions. Bandwidths F_Δ should be wide enough to let through rapid frequency variations and, at the same time, the maximum value of F_Δ is limited by half of the minimum possible pitch frequency - this constraint ensures that only monocomponent subband signals are analyzed. Center frequencies F_c should be close to the frequencies of analyzed harmonics. On the stage of pitch estimation their locations are unknown and they are chosen uniformly with F_Δ spacing ensuring that each harmonic has a pass-band with suitable center frequency.

In the case of rapid frequency modulations it is not possible to get accurate estimates of instantaneous harmonic parameters because of the aforementioned limits of the bandwidths F_Δ. In order to overcome this issue the frequency-warped impulse response is applied that allows narrow-band filtering of frequency modulated components [4]:

$$S_{F_\Delta,F_c}(m) = 2\sum_{n=-\infty}^{\infty}\frac{\sin(F_\Delta n)}{n\pi}w(n)s(m+n)e^{-j\varphi_c(n)} \quad (10)$$

where $\varphi_c(n) = \sum_{l=0}^{n} F_c(l)$.

The last equation is computationally inefficient and not used directly in the practical implementation described below. The same effect is achieved by applying time-warping to the input signal. The time warping function $W(\psi) = m$ matches samples of the signal $s(m)$ to the phase parameter of pitch ψ. The warped signal $s_W(\psi) = s(W(\psi))$ is calculated at equal phase steps and then processed using (6) - figure 2.

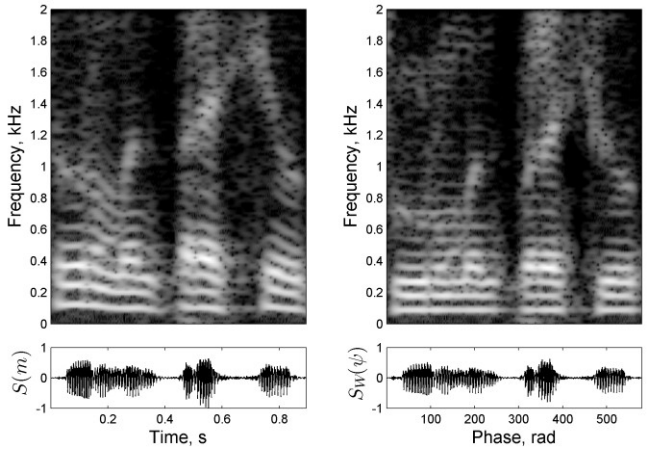


Figure 2 - Time warping of a speech signal

The operation can improve performance of the estimator for signals with high frequency modulations however it implies an additional delay and computational cost.

Frequency contour is not known from the outset. It means that the time warping can be carried out only after some preliminary pitch estimation. The algorithm proposed in the paper optionally applies time-warping in the post-processing stage.

4. PITCH ESTIMATION ALGORITHM

4.1. Algorithm outline[1]

As the original RAPT, the algorithm described below does not require any preprocessing and provides good results for wide range of signal sample rates ($6kHz \le F_s \le 44kHz$). The pitch estimation scheme is given in figure 3. The possible pitch range is 50 to 500Hz. In the current implementation values of instantaneous pitch are evaluated every 5 ms. The algorithm consists of the following steps.

1) In order to reduce the computational cost the signal is downsampled to $6kHz$. It is assumed that harmonics above *3kHz* do not influence the overall periodicity much and can therefore be discarded. It should be pointed out that unlike RAPT downsampling of the signal does not lead to any loss in frequency resolution.

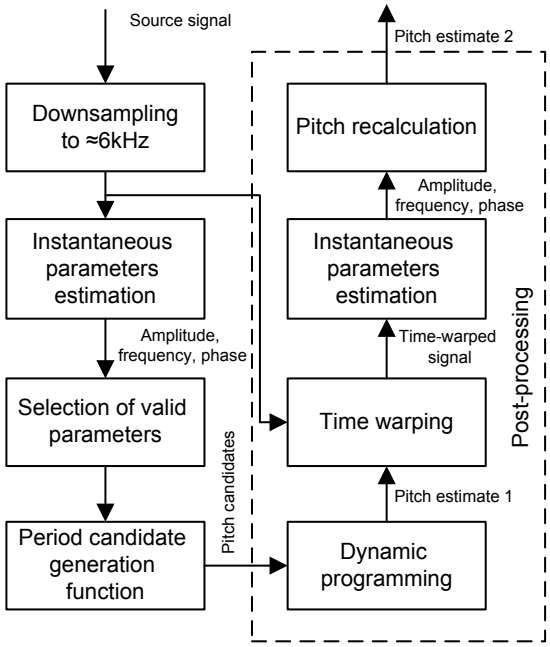


Figure 3 - Pitch estimation scheme

2) Estimation of instantaneous harmonic parameters related to the current analysis moment is carried out. The equation (6) is evaluated using analysis DFT-modulated filter bank with the prototype filter

$$h_p(n) = 2\frac{\sin(F_\Delta n)}{n\pi}w(n), \quad (11)$$

[1] Matlab implementation of the algorithm is available at http://dsp.tut.su/irapt.html

where $w(n)$ – zero-centered Hamming window. As has been said above channels passband $2F_\Delta$ equals to the minimum possible pitch value. The frequency spacing between channels is set equal to F_Δ. The channels of the filter bank are overlapped in frequency domain.

3) The obtained harmonic frequencies are checked to discard possible duplicates that can occur because of the frequency overlap.

4) Instantaneous period candidate generation function $\phi_{inst}(m,k)$ is approximately calculated. The approximation is used instead of the direct evaluation of (4) in order to reduce computational cost. The approximation is done by inverse fast Fourier transform (IFFT):

$$\hat{\phi}_{inst}(m,k) = IFFT\left(x_{amp}\right), \tag{12}$$

where x_{amp} is the signal that consists of zeros and normalized amplitude values $A^2_{p\,norm}(m) = \frac{A^2_p(m)}{\sum_{p=1}^{P} A^2_p(m)}$ placed at discrete frequency positions closest to $\pm F_p(m)$. The IFFT size specifies frequency resolution of the function.

5) Dynamic programming technique is applied to select the best pitch candidate.

After this step an initial estimation of pitch is available (denoted as "Pitch estimate 1" on the scheme). The pitch accuracy is limited by the IFFT size used in the previous step and can be degraded more by high frequency modulations.

6) Using estimated pitch contour the time-warped signal $s_W(\psi)$ is obtained from the source downsampled signal using all-pass sinc-filters.

7) Estimation of instantaneous harmonic parameters from $s_W(\psi)$ is carried out. As in the step 2) a DFT-modulated filter bank is used. The difference is that there is no overlapping between channels and that each channel of the filter bank processes only one correspondent harmonic.

8) The new pitch values (denoted as "Pitch estimate 2" on the scheme) are calculated as follows

$$F_0(m) = \sum_{p=1}^{P} \frac{F_p(m)\mathrm{A}_p(m)}{p\sum_{j=1}^{P}\mathrm{A}_j(m)}. \tag{13}$$

4.2. Computational complexity

The computational complexity of downsampling (step 1) depends linearly on the sampling frequency of the source signal F_s. The complexity of step 2 and 7 is equal to $O(N + M\log_2 M)$, where N – prototype filter order and M – number of channels of the DFT filter bank. The complexity of instantaneous period candidate generation function approximation (step 4) is $O(B\log_2 B)$, where B is the IFFT size. The overall complexity of the algorithm is of quasilinear form.

The first pitch estimate "Pitch estimate 1" is available with 50ms inherent delay. Pitch values with improved accuracy "Pitch estimate 2" require additional 43ms.

5. SIMULATION RESULTS

In order to evaluate true performance of the proposed algorithm a set of artificial signals with predefined instantaneous pitch is used. The pitch change rate differs from 0 to 2 Hz/ms. The pitch values are within the range of 100-350Hz. The signals are sampled at 44.1kHz and corrupted with additive white noise of different intensity. The amount of noise is specified by harmonic-to-noise rate (*HNR*)

$$HNR = 10\lg {\sigma_H^2}\Big/{\sigma_e^2}, \tag{14}$$

where σ^2_{H} – the energy of the harmonic signal and σ^2_{e} – the energy of the noise. The range of *HNR* is from 25dB to 5dB. The lower edge of 5dB is limited by voiced/unvoiced decision that tends to classify all the frames with lower *HNR* as unvoiced.

The performance of the algorithms is evaluated in terms of 1) gross pitch error (GPE) and 2) mean fine pitch error (MFPE).

Percentage of GPE is calculated as

$$GPE(\%) = \frac{N_{GPE}}{N_v} \times 100 \tag{15}$$

where N_{GPE} - the number of frames with estimated pitch error higher than $\pm 20\%$ of the true pitch value, N_v – overall number of voiced frames.

Mean fine pitch error is calculated on voiced frames where no gross pitch errors occur.

$$MFPE(\%) = \frac{1}{N_{FPE}} \sum_{n=1}^{N_{FPE}} \frac{|F_0^{true}(n) - F_0^{est}(n)|}{F_0^{true}(n)} \times 100 \tag{16}$$

where N_{FPE} – number of voiced frames without GPE, $F_0^{true}(n)$ – true pitch and $F_0^{est}(n)$ - estimated pitch.

Five different algorithms are compared: RAPT [3], YIN [6], SWIPE' [7] and two versions of the proposed instantaneous pitch estimation technique – one without pitch recalculation (IRAPT 1) and one with pitch recalculation (IRAPT 2).

The simulation results obtained for artificial signals are brought together in table 1. The results show that all of the algorithms have a very close performance in the case of stationary pitch and the advantage of IRAPT 1-2 grows with increase of pitch modulations. In the presence of white noise the proposed algorithms keep their benefits, however at low *HNR* values the advantage of IRAPT 2 over IRAPT 1 decreases.

The algorithms are compared using natural speech samples from the PTDB-TUG speech database [5]. The

database contains 2342 sentences taken from the TIMIT corpus that have been read by 10 male and 10 female speakers. The database includes ground truth signals obtained from a laryngograph and their estimated pitch values. These values cannot be considered as instantaneous and therefore it is not possible to compare the algorithms regarding their ability to estimate instantaneous pitch. However, the experiment shows that the proposed technique has a close performance to other pitch detection algorithms – table 2.

Table 1 - Performance comparison using artificial signals

		Pitch change rate Hz/ms				
		0	0.5	1	1.5	2
HNR 25dB						
RAPT	GPE	**0**	**0**	**0**	7.90	18.42
	MFPE	0.037	0.103	0.219	0.405	0.778
YIN	GPE	**0**	**0**	**0**	**0**	5.36
	MFPE	**0.002**	0.156	0.778	2.136	3.905
SWIPE'	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.09	0.150	0.337	0.607	1.206
IRAPT 1	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.111	0.094	0.100	0.104	0.255
IRAPT 2	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.013	**0.050**	**0.051**	**0.060**	**0.114**
HNR 15dB						
RAPT	GPE	**0**	**0**	**0**	7.90	18.42
	MFPE	0.053	0.108	0.217	0.415	0.778
YIN	GPE	**0**	**0**	**0**	**0**	5.16
	MFPE	**0.004**	0.154	0.785	2.103	3.803
SWIPE'	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.165	0.193	0.347	0.632	1.194
IRAPT 1	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.113	0.094	0.102	0.111	0.273
IRAPT 2	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.049	**0.056**	**0.065**	**0.074**	**0.148**
HNR 5dB						
RAPT	GPE	**0**	**0**	**0**	10.52	18.42
	MFPE	0.161	0.205	0.268	0.506	0.871
YIN	GPE	**0**	**0**	**0**	**0**	4.33
	MFPE	**0.019**	0.151	0.813	1.948	3.524
SWIPE'	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.316	0.253	0.373	0.706	1.307
IRAPT 1	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.143	**0.099**	**0.115**	**0.147**	0.356
IRAPT 2	GPE	**0**	**0**	**0**	**0**	**0**
	MFPE	0.162	0.131	0.145	0.164	**0.256**

6. CONCLUSIONS

An instantaneous pitch estimation algorithm has been proposed. The algorithm utilizes the standard RAPT framework with some major modifications. The first modification is that the period candidate generation function is calculated using instantaneous harmonic parameters. The periodicity of the signal thus is related to a time moment rather than to an entire analysis frame. This feature makes the function immune to any pitch modulations. The second modification is the additional post-processing step that provides more accurate instantaneous pitch values due to time-warping. The algorithm is implemented efficiently in the sense that its complexity is of quasilinear form. The algorithm has been tested using artificial harmonic signals with predefined modulated pitch and different *HNR* ratios. It has been shown that the accuracy of the algorithm degrades less than of RAPT, YIN and SWIPE' in the case of rapid pitch modulations. The algorithm also proved to be robust against additive noise. The experiments on natural speech conformed to these results. Though it was not the main concern of the authors the algorithm operates continuously as well as the original RAPT and can be used in real-time applications where latency of 50-90ms can be tolerated.

Table 2 - Performance comparison using natural speech

	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'	**0.783**	1.507	4.273	**0.800**
IRAPT 1	1.625	1.608	**3.777**	0.977
IRAPT 2	1.571	1.565	**3.777**	1.054

7. ACKNOWLEDGEMENT

Work was supported by Bialystok University of Technology under the grant S/WI/4/08.

8. REFERENCES

[1] J. O. Hong and P. J. Wolfe, "Model-based estimation of instantaneous pitch in noisy speech" in Proceedings of INTERSPEECH, 2009.

[2] B. Resch, M. Nilsson, A. Ekman and W. B. Klejin "Estimation of the Instantaneous Pitch of Speech", IEEE Trans. on Audio, Speech, and Lang. Process., 2007, vol. 15, no. 3, pp. 813-822.

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

[4] Al. Petrovsky, E. Azarov and A. Petrovsky, "Hybrid signal decomposition based on instantaneous harmonic parameters and perceptually motivated wavelet packets for scalable audio coding", Signal Processing, vol. 91, Issue 6, Fourier Related Transforms for Non-Stationary Signals, pp. 1489-1504, June 2011.

[5] G. Pirker, M. Wohlmayr, S. Petrik and F. Pernkopf, "A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario", in Proceedings of INTERSPEECH, 2011, p. 1509-1512

[6] A. Cheveigné and H. Kawahara "YIN, a fundamental frequency estimator for speech and music", *Journal Acoust. Soc. Am.*, vol. 111, no. 4, pp 1917-1930, Apr. 2002.

[7] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music", *Journal Acoust. Soc. Am.*, vol. 123, no. 4, pp 1638-1652, Sep. 2008.