SPEECH ANALYSIS BASED ON SINUSOIDAL MODEL WITH TIME-VARYING PARAMETERS

E. Azarov, M. Vashkevich, A. Petrovsky

Computer Engineering Department,
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus

Work was supported by IT ForYou Company and Belarusian Republican Foundation for Fundamental Research (grant No F14MV-014)

1. Introduction

9267

The paper presents some techniques for extracting pitch and spectral envelope of a signal using sinusoidal model with instantaneous parameters.

The main features of the proposed algorithms are:

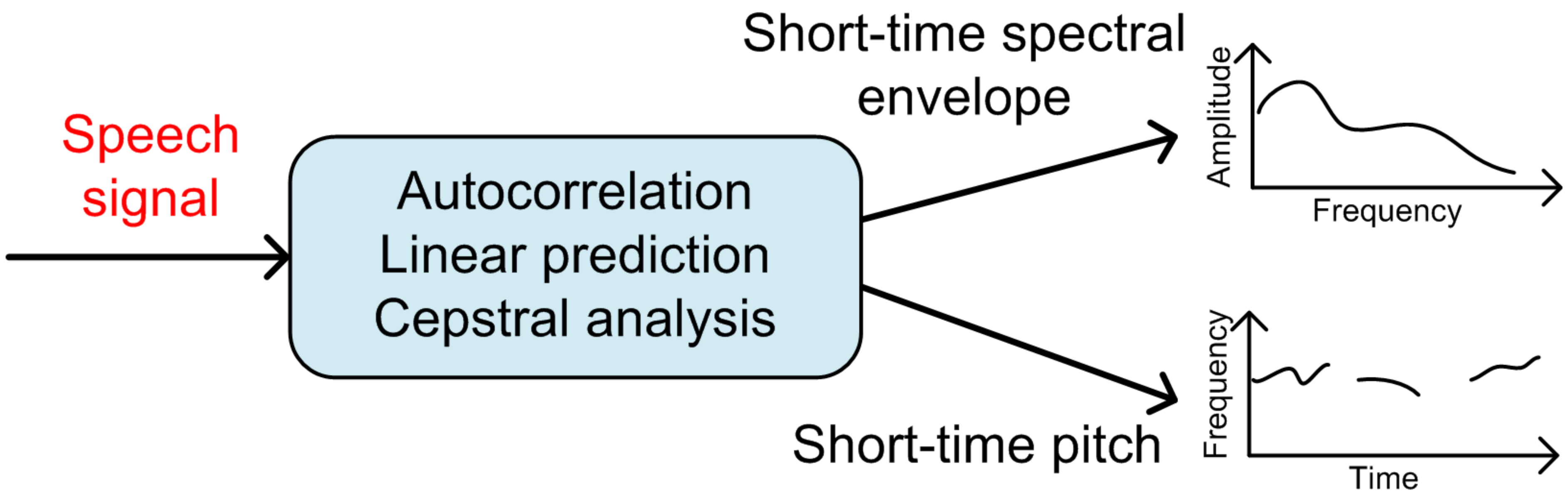
- model-based analysis that provides high time-frequency resolution;
- model-based estimation of instantaneous pitch
- model-based estimation of instantaneous spectral envelope;
- accurate envelope representation by linear predictors of high order.

Two-step analysis:

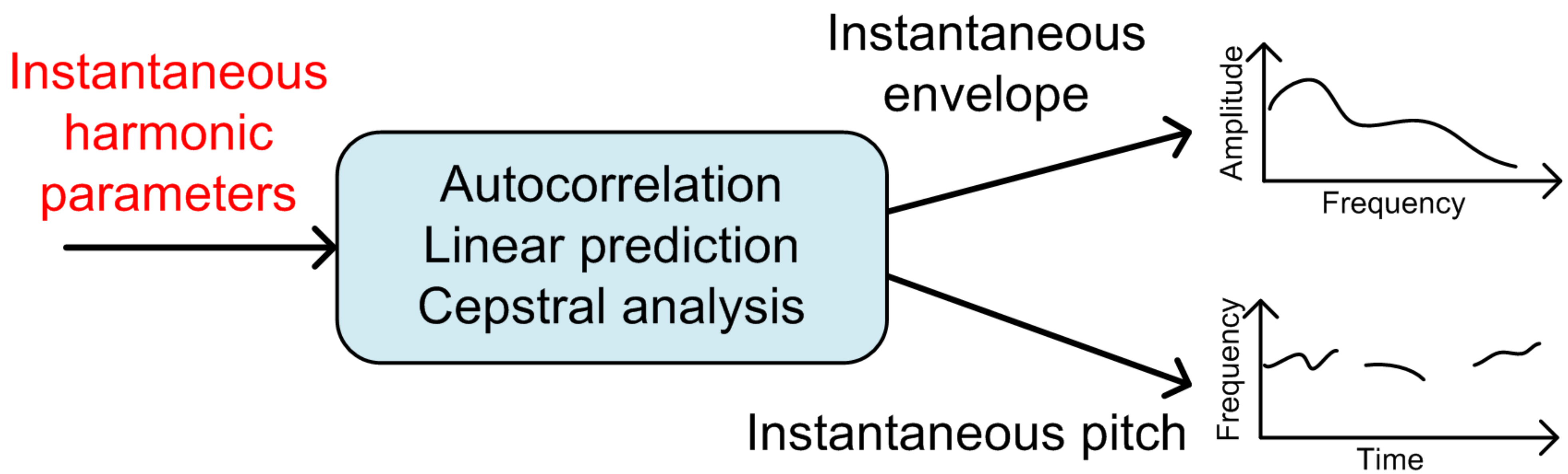
1. extraction of sinusoidal parameters;
2. parameters transformation to required characteristics

2. Model-based extraction of speech characteristics

Classical extraction techniques



Model-based extraction techniques



3. Estimation of instantaneous harmonic parameters

The signal $s(m)$ is decomposed into overlapping bandlimited analytical signals $S_{F_\Delta,F_c^i}(m)$:

$$S_{F_\Delta,F_c^i}(m) = \sum_{n=-\infty}^{\infty} \frac{\sin(F_\Delta n)}{n\pi} w(n)s(m-n)e^{-jF_c^i n} = A_{F_\Delta,F_c^i}(m)\cos\left(\varphi_{F_\Delta,F_c^i}(m)\right),$$

where $2F_\Delta$ - bandwidth and F_c^i - center frequency of the *i*-th band and $w(n)$ – an even window function. Then instantaneous parameters are evaluated as

instantaneous amplitude $\longrightarrow$ $A_{F_\Delta,F_c^i}(m) = \sqrt{R^2(m) + I^2(m)},$

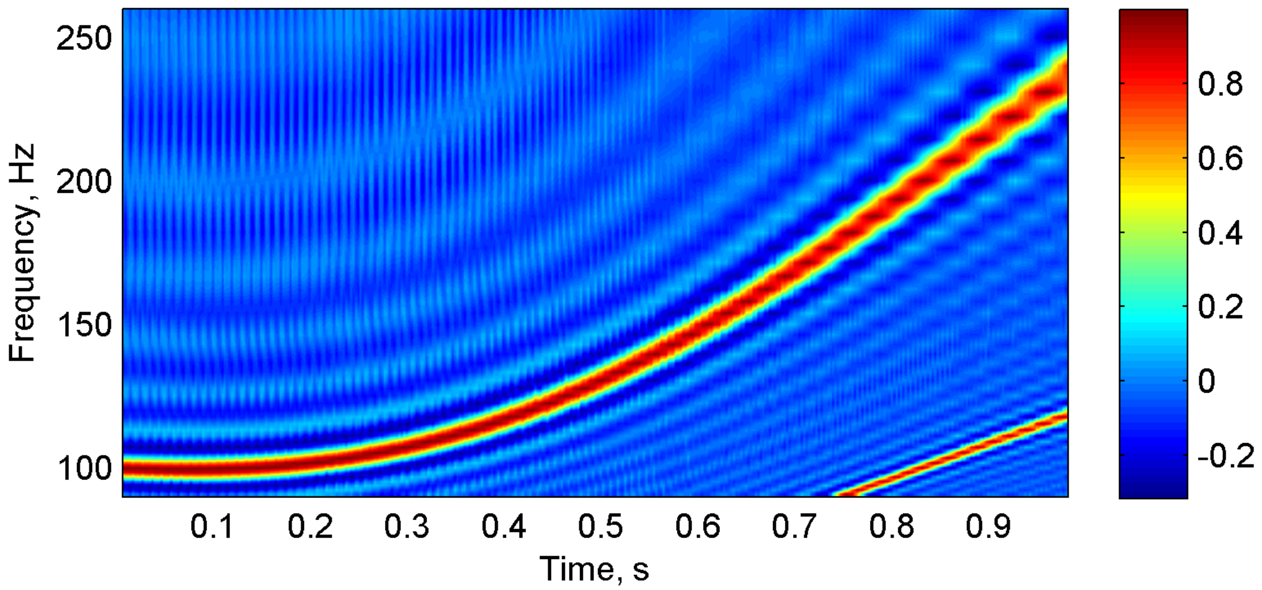
instantaneous phase $\longrightarrow$ $\varphi_{F_\Delta,F_c^i}(m) = \arctan\left(\frac{-I(m)}{R(m)}\right),$

instantaneous frequency $\longrightarrow$ $F_{F_\Delta,F_c^i}(m) = \varphi'_{F_\Delta,F_c^i}(m),$

where $R(m)$ and $I(m)$ are real and imaginary parts of $S_{F_\Delta,F_c^i}(m)$ respectively.

4. Normalized cross-correlation function (NCCF)

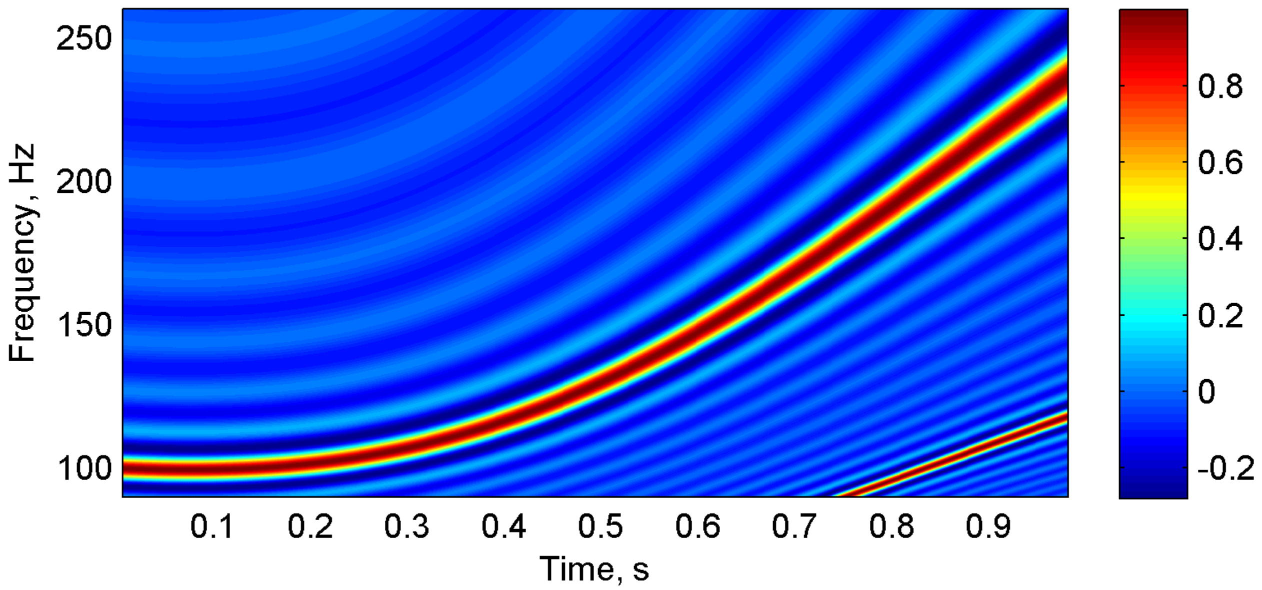
Time domain



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

where $e_i = \sum_{l=i}^{i+n-1} s_l^2$ and n – window size

Model-based estimation



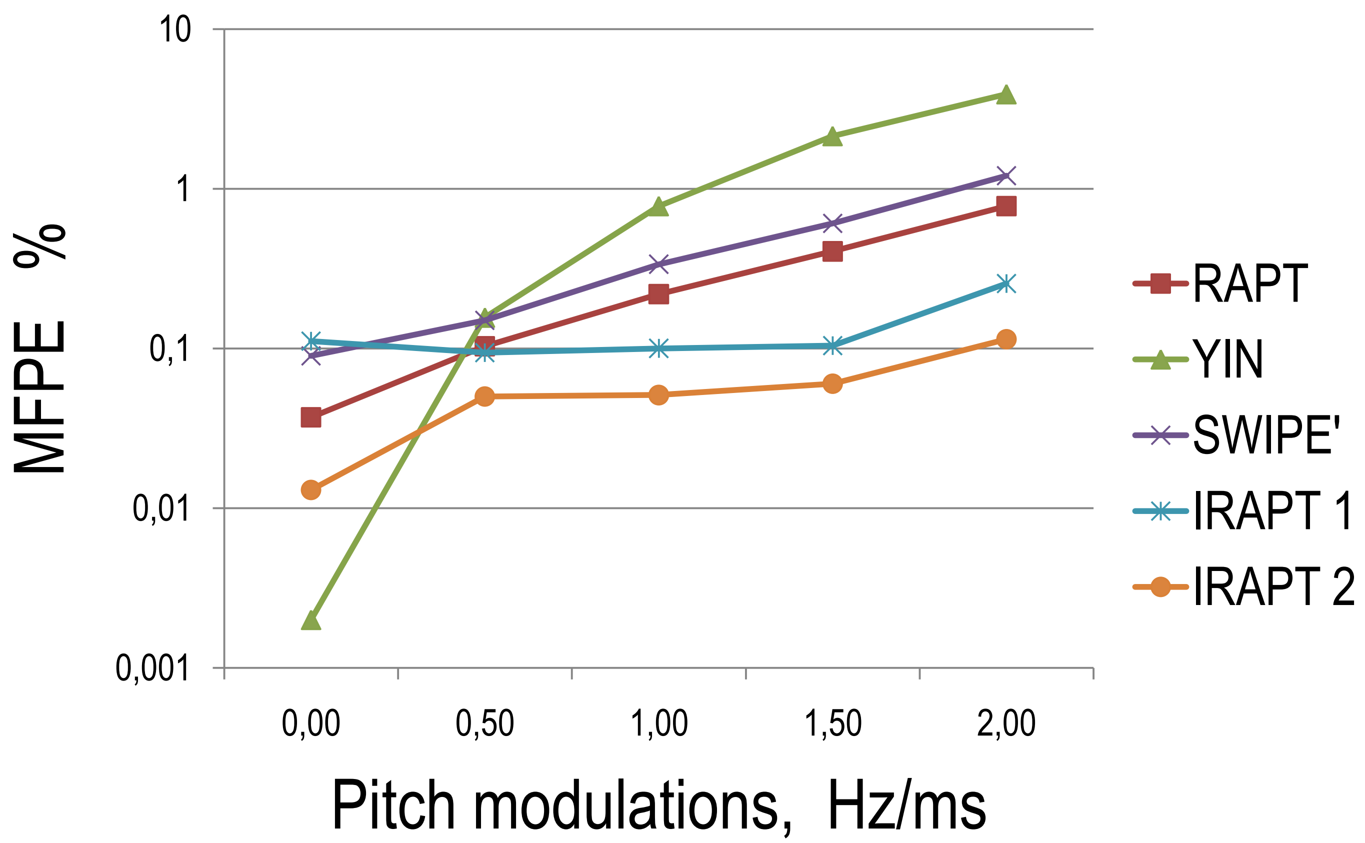
$$\phi_{inst}(m,k) = \frac{\sum_{p=1}^{P} A_p^2(m)\cos(F_p(m)k)}{\sum_{p=1}^{P} A_p^2(m)}$$

P – number of bandlimited analytical signals

5. Experimental results of pitch extraction

The proposed technique is compared with other pitch estimation algorithms in terms of gross pitch error (GPE, %) and mean fine pitch error (MFPE, %).

Artificial signals



Natural speech

	Male		Female	
	GPE	MFPE	GPE	MFPE
RAPT	3.69	1.74	6.07	1.18
YIN[3]	3.18	**1.39**	3.96	0.84
SWIPE'[4]	**0.78**	1.51	4.27	**0.80**
IRAPT 1	1.63	1.61	**3.78**	0.98
IRAPT 2	1.57	1.57	**3.78**	1.05

[3] 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.

[4] 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.

6. High-order linear prediction

The coefficients are evaluated using the following system:

$$Q = \begin{bmatrix} q(0) & \cdots & q(p-1) \\ \vdots & \ddots & \vdots \\ q(p-1) & \cdots & q(0) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} = -\begin{bmatrix} q(1) \\ \vdots \\ q(p) \end{bmatrix} \qquad q(l) = \sum_{i=1}^{K-1} D(l,i)$$

where $a_1, \dots, a_p$ – prediction coefficients, *p* – prediction order.

Each segment of the spectral envelope $f_i \le \omega \le f_{i+1}$, $1 \le i \le K-1$ is defined by a linear equation $A(\omega) = b_i\omega + c_i$

$$D(l,i) = \begin{cases} \frac{b_i}{l^2}[\cos(f_{i+1}l) + f_{i+1}l\sin(f_{i+1}l)] + \frac{c_i}{l}\sin(f_{i+1}l) - \\ \quad -\frac{b_i}{l^2}[\cos(f_i l) + f_i l\sin(f_i l)] - \frac{c_i}{l}\sin(f_i l) & l \ne 0 \\ \quad \frac{1}{2}b_i f_{i+1}^2 + c_i f_{i+1} - \frac{1}{2}b_i f_i^2 - c_i f_i & l = 0 \end{cases}$$

7. Model-based linear prediction of speech