16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

ESTIMATION OF THE INSTANTANEOUS HARMONIC PARAMETERS OF SPEECH

Elias Azarov, Alexander Petrovsky and Marek Parfieniuk

Department of Computer Engineering, Belarusian State University of Informatics and Radioelectronics
P.Brovky 6, 220027, Minsk, Belarus
Department of Real-Time Systems, Bialystok Technical University, Bialystok, Poland
Wiejska 45A, 15-351, Bialystok, Poland
phone: + (48) 85 746-90-50, fax: + (48) 85 746-90-57, email: palex@bsuir.by, palex@wi.pb.edu.pl

ABSTRACT

This paper describes a method of accurate estimation of the instantaneous speech signal harmonic parameters. The method is based on adaptive filtering of the speech signal along its harmonic components. A simple way of filter synthesis based on the Fourier transform is also proposed. The synthesized filters have a closed form impulse response which can be modulated in frequency domain to achieve better performance for components with high frequency alteration. This method is also applicable to give an accurate estimate of the fundamental frequency of speech.

1. INTRODUCTION

The Harmonic+Noise representation of a speech signal [1] is used effectively in many speech applications [2-4], for instance in speech synthesis, coding, recovery and recognition; in speaker identifying; in speech conversion and noise reduction. Accurate separation of the periodic and noise parts of the signal has been a fundamental problem for a few recent decades. The primary way to solve this problem is to use the DFT (Discrete Fourier Transform) or some of its modifications. However, such method always assumes a stationary character of the signal within an analysis frame. It means that within some short time interval the frequency and magnitude values of harmonic components are considered to be constant [5-7]. Besides the assumption of the stationarity most researchers assume that the frequencies of harmonics have values exactly divisible by the fundamental frequency of speech. Despite the fact the above mentioned assumptions can strongly simplify estimation methods they can cause worsening of the analysis accuracy, and lead to audible artifacts after the reconstruction of the signal.

In this paper we suggest a method for exact harmonic parameters estimate, assuming frequencies and magnitudes changes for every sample of the speech signal. Also we consider a possibility that the frequency of any harmonic could have some deviation from the fundamental frequency of speech. Similar approaches are proposed in [8-10], however they do not consider the instantaneous fundamental frequency modulations influence on the parameters' estimate. The energy separation technique proposed in [11] can be efficiently applied for the instantaneous frequency and magnitude calculation. However, for speech applications this method requires additional filtering and in [11] the Gabor filter is used for this purpose. This filter cannot provide accurate frequency tracking in the frames of rapid fundamental frequency changes.

For the accurate estimation we have developed the frequency-modulated filter [12]. Its closed form impulse response can be adjusted according to instantaneous frequencies of the harmonics and the fundamental frequency modulations of speech. Also we present closed form expressions for instantaneous phase, frequency, and magnitude obtained directly from the filter output. For the practical implementation we propose an algorithm to estimate the harmonic parameters that can be used for efficient speech signal separation. The algorithm evaluates the parameters sample per sample by adjusting filter parameters at every step according to estimated frequencies at the previous step and the fundamental frequency modulations of speech. We executed series of experiments and proved high efficiency of the proposed method for estimation of the instantaneous harmonic parameters. Experiments were performed by using synthetic signals with predefined parameters along with original speech signals. The method combines high accuracy and noise robustness. Because of its simplicity the method can be used in various speech applications.

2. HARMONIC MODEL

A speech signal can be efficiently represented as a sum of two basic components the periodic and the noise ones [13]. This representation can be expressed by the following formula:

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

where $s(n)$ is the source signal, A_k - the instantaneous magnitude of the k-th harmonic component, K is the number of the harmonic components, $r(n)$ is the noise component and $\varphi_k(n)$ is the instantaneous phase of the k-th harmonic component. There is a definite correlation between $\varphi_k(n)$ and the instantaneous frequency f_k. It can be presented in the following way:

$$\varphi_k(n)=\sum\nolimits_{i=0}^{n}\frac{2\pi f_k(i)}{F_s}+\varphi_k(0),$$

where F_s is the sampling frequency and $\varphi_k(0)$ the initial phase of k-th harmonic. The harmonic model assumes that the frequencies of the components are integer multiples of the fundamental frequency: $f_k = kf_0$, where f_0 is the fundamental frequency. In the present work we assume

$$\left|f_k - kf_0\right| < f_{tr}\,. \tag{2}$$

In other words, the instantaneous frequencies can deviate from the multiples of the fundamental frequency for the value less than some specified f_{tr}. To separate a certain harmonic from the rest ones it is necessary to use a bandpass filter [11]. Taking into account (2), an appropriate bandwidth that covers a single specified harmonic can be found. This assumption lets us synthesize a digital filter, which could be able to perform the harmonic separation. Here is the list of some special requirements for the filter:

- to provide an ability of filtering the signal in an arbitrary bandwidth. For this purpose the impulse response should be derived as a closed form expression that uses the passband center frequency and its width as parameters;
- to represent the output signal as a one-component periodic function to derive the instantaneous parameters expressions directly from the output signal;
- the impulse response should be continuous to implement the time warping procedure for frequency-modulated signals.

3. ESTIMATION OF THE INSTANTANEOUS HARMONIC PARAMETERS

3.1 Synthesis of the stationary filter

The N-point DFT can be considered as a finite impulse response (FIR) filter for a specified normalized frequency f:

$$S(f) = \frac{1}{N}\sum_{n=0}^{N-1} s(n)e^{\frac{-j2\pi nf}{N}}\,,$$

$$MAG[S(f)] = \sqrt{\operatorname{Re}S(f)^2 + \operatorname{Im}S(f)^2}\,,$$

$$\varphi[S(f)] = -\arctan\frac{\operatorname{Im}S(f)}{\operatorname{Re}S(f)},$$

the output signal can be written as a periodic function with the constant frequency f and the constant magnitude MAG in the following form:

$$\bar{s}(n) = MAG[S(f)]\cos(\frac{2f\pi n}{N} + \varphi[S(f)])\,.$$

The closed form impulse response $h(n)$ of this filter for frequency f in Hz is:

$$h(n) = \cos(\frac{2\pi}{F_s}nf)\,.$$

Let us generalize this expression considering a constant frequency band (from F_1 to F_2) instead of the constant frequency f. We can obtain the impulse response:

$$h(n) = \int_{F_1}^{F_2}\cos(\frac{2\pi}{F_s}nf)df\,,$$

Integrating the expression we will get the impulse response in the following form:

$$h(n) = \begin{cases} F_2 - F_1, & n = 0 \\ \frac{F_s}{n\pi}\cos(\frac{n\pi}{F_s}(F_2 + F_1))\sin(\frac{n\pi}{F_s}(F_2 - F_1)), & n \neq 0 \end{cases}$$

The output signal $\bar{s}(n)$ can be calculated as the convolution of $s(n)$ and $h(n)$. It can be expressed as the sum:

$$\begin{aligned}\bar{s}(n) = \sum_{i=0}^{N-1}\frac{s(i)F_s}{(n-i)\pi}\cos(\frac{(n-i)\pi}{F_s}(F_2 + F_1))\cdot \\ \cdot\sin(\frac{(n-i)\pi}{F_s}(F_2 - F_1))\end{aligned} \tag{3}$$

The last expression can be rewritten in the following form:

$$\bar{s}(n) = A(n)\cos(\frac{2\pi}{F_s}nF_c) + B(n)\sin(\frac{2\pi}{F_s}nF_c)\,,$$

where

$$A(n) = \sum_{i=0}^{N-1}\frac{s(i)F_s}{(n-i)\pi}\sin(\frac{2\pi}{F_s}F_\Delta(n-i))\cos(\frac{2\pi}{F_s}F_c i)\,,$$

$$B(n) = \sum_{i=0}^{N-1}\frac{s(i)F_s}{(n-i)\pi}\sin(\frac{2\pi}{F_s}F_\Delta(n-i))\sin(\frac{2\pi}{F_s}F_c i)\,,$$

$$F_c = \frac{F_2 + F_1}{2}\,,$$

$$F_\Delta = \frac{F_2 - F_1}{2}\,.$$

Thus the output signal of the filter can be written as the magnitude and frequency-modulated cosine function:

$$\bar{s}(n) = C(n)\cos(\frac{2\pi}{F_s}F_c n + \alpha(n))\,,$$

where $$C(n) = \sqrt{A^2(n) + B^2(n)}\,,$$

$$\alpha(n) = \arctan(-\frac{B(n)}{A(n)})\,.$$

Consequently, the instantaneous frequency F, magnitude MAG and phase φ can be determined as follows:

$$F(n) = \frac{\alpha(n+1) - \alpha(n)}{2\pi}F_s + F_c\,,\ MAG(n) = C(n)\,,$$

$$\varphi(n) = 2\pi F_c n + \alpha(n)\,.$$

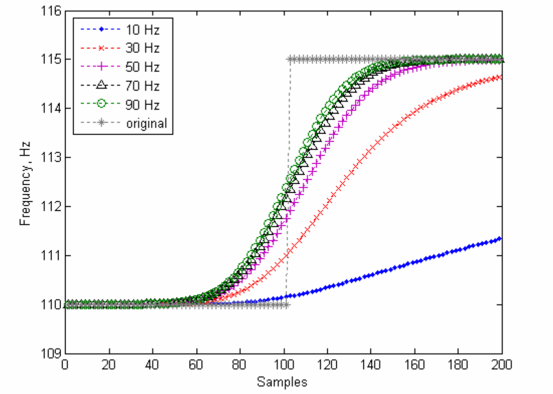


Figure 1 – Harmonic frequency tracking with different bandwidth filters

So then we have the required closed form filter expression and formulas that provide us with the instantaneous values of frequency, magnitude and phase of the harmonic component within the assigned passband. F_c should be selected close to the frequency of the harmonic component in order to provide an accurate parameters estimate. During the estimation process F_c could be set initially as $F_c = kf_0$, where k is the number of the harmonics and f_0 is the instantaneous fundamental frequency. Then F_c is set equal to the estimated frequency of the respective harmonic. Generally speaking, the analysis filter bank is not uniform. The result of the tracking method is demonstrated in Fig.1. The synthesized harmonic component has a discontinuity in order to show the inertness of the filters with various bandwidths. It is clear, that filters with wider bandwidths have a closer approach to the original frequency track; however it is not always possible to use wide filter band because of small distances between the adjacent components or because of noise in the filter bandwidth. In order to demonstrate the comparative noise sensitivity of the filters with different bandwidths we present Fig.2.

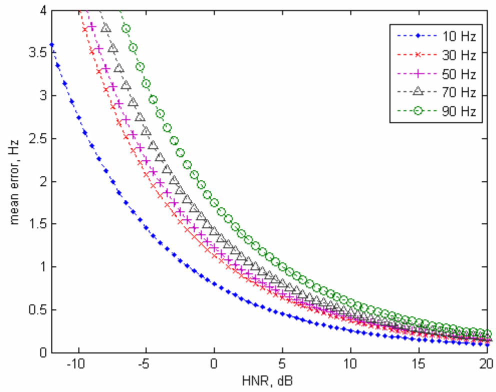


Figure 2 - Frequency mean error caused by the noise presence in the signal

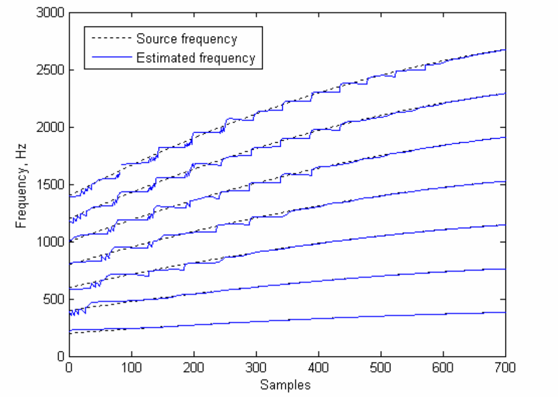


Figure 3 - Inaccurate estimation of the high order harmonics because of a rapid frequency changes

We use the harmonic noise ratio (HNR) as a measure of the noise amount in the signal. The $HNR = 10\lg E_h / E_r$, where E_h and E_r are the energies of the harmonic and noise components respectively. The passband width is restricted by parameter F_Δ, which cannot be chosen arbitrarily. In many cases, especially when dealing with a male voice, it cannot exceed 30 Hz. Also, as it was shown above, the filters with narrow bandwidth are more robust against noise. On the other hand, a lower bandwidth filter has a higher inertness and if the frequency of the harmonic component changes rapidly, it may lead to tracking failure. We established that the use of the stationary filter can provide accurate results for estimation of the fundamental frequency, but it is not suitable for high order harmonics as it is shown in Fig.3.

3.2 Synthesis of the frequency-modulated filter

Since we have the closed form impulse response we can easily adapt it to the fundamental frequency contour providing precise parameters estimate. Taking into account the fundamental frequency modulation the equation (3) can be written in the following form:

$$\bar{s}(n) = A(n)\cos(\frac{2\pi}{F_s}\varphi_k(n)) + B(n)\sin(\frac{2\pi}{F_s}\varphi_k(n)),$$

where

$$\varphi_k(n) = (\sum_{i=0}^{n} F_0(n) - \sum_{i=0}^{N/2} F_0(n))k,$$

$F_0(n)$ is the instantaneous fundamental frequency,

$$A(n) = \sum_{i=0}^{N-1} \frac{s(i)F_s}{(n-i)\pi}\sin(\frac{\pi}{F_s}F_\Delta(n-i))\cos(\frac{2\pi}{F_s}\varphi_k(n)),$$

$$B(n) = \sum_{i=0}^{N-1} \frac{s(i)F_s}{(n-i)\pi}\sin(\frac{\pi}{F_s}F_\Delta(n-i))\sin(\frac{2\pi}{F_s}\varphi_k(n)) \cdot$$

Instantaneous frequency F, magnitude MAG and phase φ can be presented in the following way:

$$F(n) = \frac{\alpha(n+1) - \alpha(n)}{2\pi}F_s + F_0 \cdot k, \qquad (4)$$

$$MAG(n) = C(n)\,,\ \varphi(n) = 2\pi F_0 kn + \alpha(n)\,,\text{ where}$$

$$C(n) = \sqrt{A^2(n) + B^2(n)}\,,\ \alpha(n) = \arctan\left(-\frac{B(n)}{A(n)}\right).$$

The frequency-modulated filter has frequency-modulated bandpass width aligned to the fundamental frequency contour. It provides accurate estimate of the high order harmonic parameters.

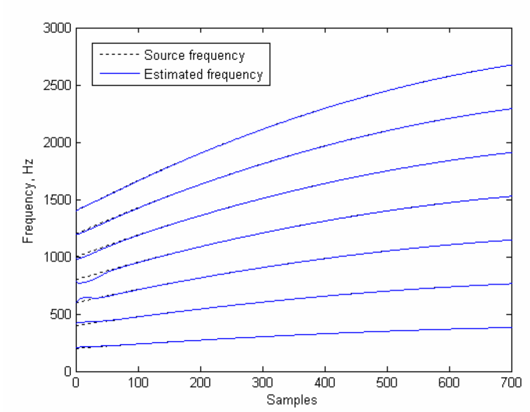


Figure 4 - Accurate estimation of the high order harmonics with frequency-modulated filter

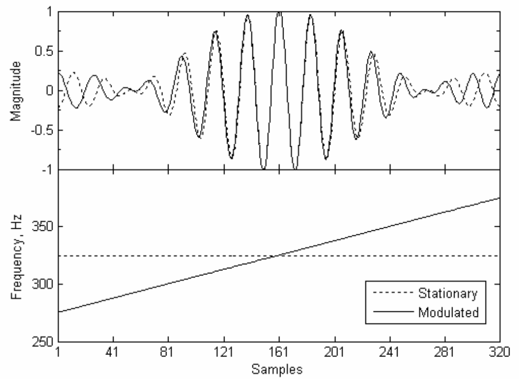


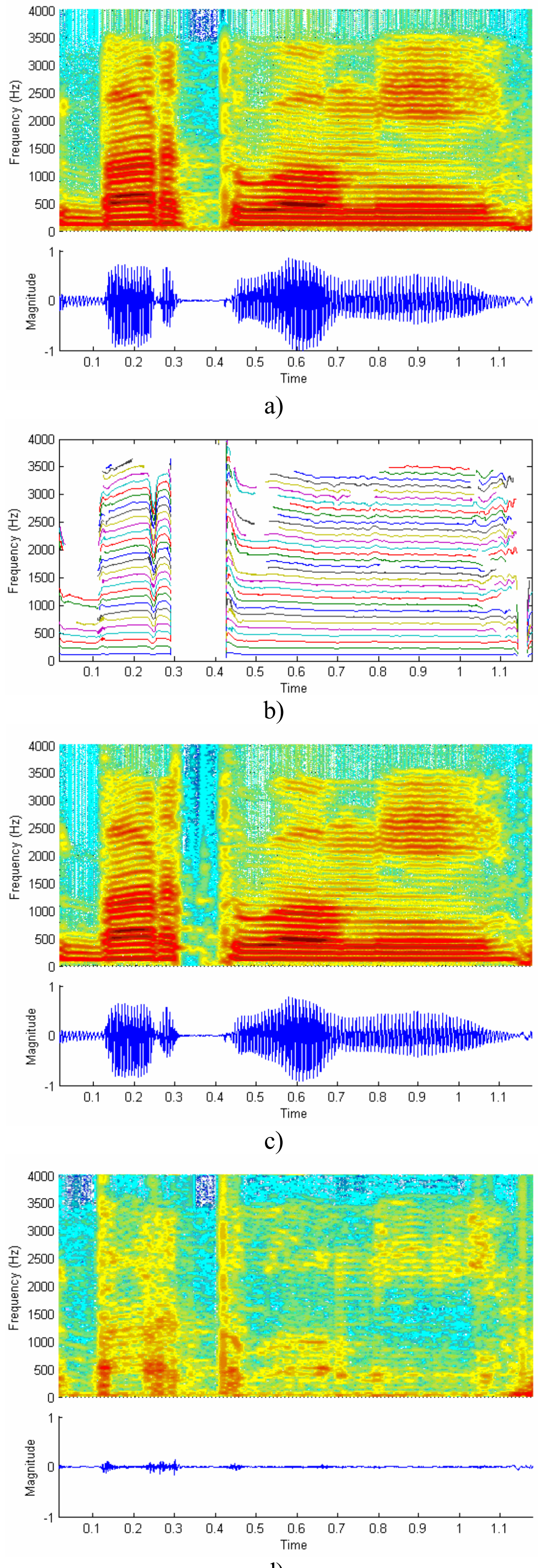
Figure 5 – Impulse responses of the stationary and frequency-modulated filters

The result of using frequency-modulated filter is shown in Fig.4. Two impulse responses are shown in Fig.5. The first one is the stationary filter impulse response (F_c=325 Hz, F_Δ=35 Hz), the second one is the frequency-modulated filter impulse response (F_c=[275,375] Hz, $F_\Delta = 35$ Hz).

4. SPEECH SIGNAL PERIODIC / NOISE DECOMPOSITION

4.1 Estimation algorithm of the harmonic parameters

For an accurate harmonic/noise separation of the speech signal it is quite necessary to know the fundamental frequency contour. It lets us to evaluate the number of filters, the bandpass locations of filters and to synthesize frequency- modulated impulse responses according to the fundamental frequency modulation. The starting point of the fundamental



a)

b)

c)

d)

Figure 6 - Speech signal decomposition (a – source signal, b – frequency trajectories estimated, c – synthesized periodic part, d – noise part

frequency can be taken in the beginning of the voiced segment by applying the filters with equidistant passbands to the low frequency range (from 65 to 470 Hz). Then the fundamental frequency contour can be estimated till the end of the voiced segment as described in section 3.1. It could be determined by a magnitude threshold (in the experiments we used 1% of maximum magnitude value) whether the current sample belongs to the voiced segment or not.

For the harmonic parameters estimation (magnitude, frequency, phase) we propose the following algorithm:

1) the fundamental frequency contour estimation;
2) synthesis of the current filter bank;
3) the evaluation of the harmonic parameters from (4) and going back to step 2).

The algorithm ends when the last sample of the signal is reached. Initially F_c^k of the k-th harmonic component is calculated as $F_c^k = F_0 k$, on the further steps F_c^k is equated to the evaluated frequency of the k-th harmonic. After the harmonic parameters estimation has been made, the periodic part of the signal can be synthesized by formula (1) and then subtracted from the original signal in order to obtain the noise part.

4.2 Experimental results

As an example of the speech signal decomposition we propose the result of separation of a phrase uttered by a male speaker. The result of the decomposition is demonstrated in Fig.6. In this example we used 161-order filters with 70 Hz passband width for the fundamental estimation and 50 Hz bandwidth for estimation of the harmonics. The resulting HNR value of the separation process is 22.54 dB. It can be easily seen that the periodic part of the signal has the same quantity of the harmonic components and their magnitude and frequency modulations are preserved. Moreover the periodic part contains some transient fragments, which can be observed at the beginning and at the end of the voiced segments. Trajectories of harmonic frequencies (Fid.6. b.) are smooth and exactly reflect frequency contours in the spectrogram of the source signal (Fid.6. a.) even in regions where the energy of the correspondent component is very low. The frequencies of high order harmonics are traced properly including the regions where the fundamental frequency changes rapidly.

5. CONCLUSIONS

In the present paper the method of the instantaneous harmonic parameters (magnitude, frequency and phase) estimation has been proposed. The parameters are calculated as the result of the narrow band filtering of the speech signal. We have proposed the method of synthesis of the frequency-modulated filters with the closed form impulse response. The filter frequency bounds can be determined during the components frequency tracking and can be adjusted according to the fundamental frequency modulations. The proposed method provides high accuracy of estimation and can be easily implemented in applications, requiring speech periodic/noise decomposition. We are currently working on establishing optimal filter parameters depending on the type of signal and estimating frequencies in order to achieve better performance characteristics.

6. ACKNOWLEDGMENT

This work was supported under the Grant MNiSW No 519 030 32/3775.

REFERENCES

[1] B. Yegnanarayana, C. d'Alessandro, V. Darsions, "An Iterative Algorithm for Decomposition of Speech Signals into Voiced and Noise Components", *IEEE Trans. on Speech and Audio Coding*, vol. 6, no. 1, pp. 1-11, 1998.

[2] A.S. Spanias „Speech coding: a tutorial review", *Proc. of the IEEE,* vol. 82, no. 10, pp. 1541-1582, 1994.

[3] Stylianou Y. "Applying the harmonic plus noise model in concatenative speech synthesis", *IEEE Trans. Speech, Audio Process.*, 2001, vol. 9, no. 1, pp.21-29.

[4] Zavarehei E., Vaseghi S., Yan Q. "Noisy speech enhancement using harmonic-noise model and codebook-based post-processing", *IEEE Trans. on Audio, Speech, and Language processing*, vol. 15, no. 4, pp. 1194-1203, July 2007.

[5] R.J. McAulay, T.F. Quatieri "Speech analysis/synthesis based on a sinusoidal representation" *IEEE Trans. On Acoustics, Speech and Signal Process.*, vol. 34, no. 4, pp.744-754, 1986.

[6] R.J. McAulay, T.F. Quatieri „Sinusoidal Coding" in *Speech Coding and Synthesis* (W. Klein and K. Palival, eds.), Amsterdam: Elsevier Science Publishers, pp. 121-176., 1995.

[7] George E.B., Smith M.J.T. „Speech Analysis/Synthesis and Modification Using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model", *IEEE Trans. on Speech and Audio Process.*, vol. 5, no. 5, pp. 389-406, 1997.

[8] L.B. Almeida, J.M. Tribolet "Nonstationary spectral modeling of voiced speech", *IEEE Trans. on Acoust., Speech and Sig. Proc.* ,Vol. ASSP-31. no. 3. pp. 664 – 678, 1983.

[9] T. Abe, T. Kobayashi, and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," in *Proc. ICASSP*, 1995, pp. 756–759.

[10] T. Abe, M. Honda, "Sinusoidal model based on instantaneous frequency attractors", *IEEE Trans. on Audio, Speech, and Language processing*, vol. 14, no. 4, pp. 1292-1300, July 2006

[11] Maragos P., Kaiser J. F., Quatieri T. F., "Energy Separation in Signal Modulations with Application to Speech Analysis", *IEEE Trans. on Signal Process.*, vol. 41, no. 10, pp. 3024-3051, 1993.

[12] Petrovsky A., Stankevich A., Balunowski J. "The order tracking front-end algorithms in the rotating machine monitoring systems based on the new digital low order tracking" // *in Proc. of the 6th Intern. congress "On sound and vibration", ICSV'99*, 1999, Copenhagen, Denmark, pp.2985-2992.

[13] P. Zubrycki, A. Petrovsky. "Accurate speech decomposition into periodic and aperiodic components based on discrete harmonic transform" // in *Proc. of the 15th European Signal Process. Conf., (EUSIPCO-2007)*, Poznan, 2007, pp.2336-2340.