COMBINING ADVANCED SINUSOIDAL AND WAVEFORM MATCHING MODELS FOR PARAMETRIC AUDIO/SPEECH CODING

Alexey Petrovsky[1], Elias Azarov[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 real-time systems, Bialystok Technical University,
Wiejska 45A, 15-351, Białystok, 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 presents two fundamental enhancements in a hybrid audio/speech signal model based on AM/FM and transient representation: sinusoidal, transient, and noise (STN) components. The first enhancement involves a method of instantaneous sinusoidal parameters estimation using an adaptive filtering of the speech signal along its harmonic components. The second and perhaps more significant STN enhancement is concerned with transient components modelling based on the matching pursuit with frame-based psychoacoustic optimized wavelet packet dictionary. It significantly reduces the number of coefficients required to achieve a given perceptual distortion.

1. INTRODUCTION

The approach to lossy audio coding on the basis of transform and subband coding techniques has matured and is believed to show no significant progress in the near future. Therefore, other techniques are considered, especially parametric audio coding [1]. In this case, the audio is modeled by a limited number of objects, for instance by transients (short-lasting events), sinusoidal and noise components. The sinusoidal approach was first introduced in speech coding in the early eighties [2]. Whereas concurrent and state-of-the-art stand-alone speech coders depend heavily on speech production models to realize low bit rates, modern applications ask for integrated audio and speech coding solutions. The sinusoidal model enables a unified approach to both audio and speech coding. In practice, sinusoidal model parameters are often considered as constants within an analysis frame (depending on the signal, the length of quasi-stationary segments can vary from a few milliseconds to several hundreds of ms). A fairly general model that is often used to represent speech and audio is based on AM/FM representation [3]. In this model, the signal is represented as a sum of sinusoidal components with time-varying amplitude, phase and frequency, which are only slowly time-varying functions of time. The sinusoidal modelling approach is effective to represent the harmonic structure of many speech and audio segments. However, speech and audio signals often contain noise-like segments and transient sounds that are not efficiently modelled by AM/FM representation. In particular, transient sounds can cause a type of distortion that is known as pre-echo. Pre-echo originates from the fact that in order to reduce the number of modeling parameters, sinusoidal coders normally produce signals with a fairly high degree of pseudo-stationarity. Thus, the sharpness of transient attack segments in fact is building up gradually before the attack. One of the most recent enhancements of the sinusoidal model is the introduction of a new method that handles not only the harmonic aspects of the signal but also its broadband and transient components. This new form of adaptive signal representation is called the sines+transients+noise (STN) model [3,4]. The time-scale and pitch-scale modifications become possible due to signal separation. The sinusoidal part can be stretched or shrunk in time domain without losing its pitch. The phase and amplitude values are easily interpolated at any given moments of time. The noise can be easily transformed in time domain with good results. The transients can also be time-rescaled while preserving their original temporal envelopes. The SNT model is widely used in speech/audio processing applications because of these powerful features [1].

However, the crucial point in SNT systems design is analysis accuracy since it defines the overall performance of the system. Every analysis technique that implemented in the system should provide high accurate parameters estimation. The coordination of all analysis techniques should be carefully organized in order to get appropriate signal separation.

The focus of this paper is application of new methods for sinusoids and transients selection in hybrid (STN) modeling of audio/speech.

2. GENERAL STRUCTURE OF HYBRID STN ANALYSIS SYSTEM

The approach to hybrid audio/speech modeling is based on a combination of three different signal processing techniques: sinusoidal, matching pursuit with frame-based psychoacoustic optimized wavelet packet dictionary and bark-scaled adapted wavelet packet noise analysis. Sinusoidal part is represented as sums of sinusoids with instantaneous parameters (amplitude, frequency and phase), transients are modeled by matching pursuit with frame-based psychoacoustic optimized wavelet packet dictionary and finally noise is

processed by bark-scaled adapted wavelet packet analysis. The general structure of the analysis system is presented in figure 1.

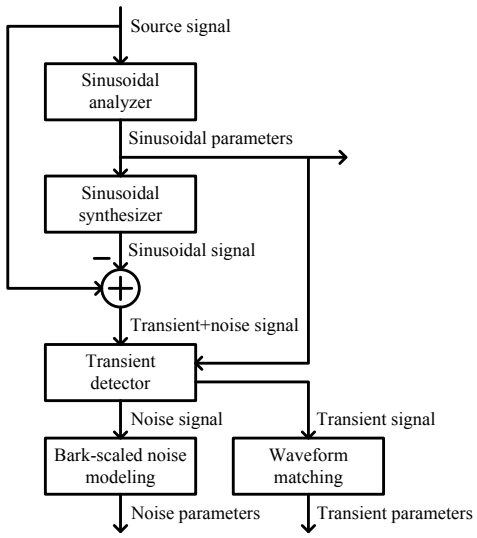

Figure 1 – General structure of hybrid SNT analysis system.

In the given sinusoidal + transients + noise (STN) model, sinusoidal modeling is directly applied to the input signal. Then, transients are detected via an energy threshold combined with a partial loudness edge detection scheme that operates on the sinusoidal modeling residual. Once the sinusoidal and transient components have been analyzed, the residual of the sinusoidal + transients modeling procedure is captured by the bark-scaled adapted wavelet packet noise model. The transient signal is parameterized by matching pursuit with frame-based psychoacoustic optimized wavelet packet dictionary. Thus the proposed system produces SNT separation and parameterization of each separated part.
This analysis scheme provides good coordination of the used analysis techniques and allows efficient processing of any speech/audio signal.

3. ADVANCED SINUSOIDAL MODEL

3.1 Sinusoidal analysis

The sinusoidal part of the signal $s(n)$ can be expressed by the following formula:

$$s(n) = \sum_{k=1}^{K} A_k(n)\cos\left(\varphi_k(n)\right), \quad (1)$$

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

$$\varphi_k(\mathrm{n}) = \sum_{i=0}^{n} \frac{2\pi f_k(i)}{F_s} + \varphi_k(0), \quad (2)$$

where F_s is sampling frequency and $\varphi_k(0)$ the initial phase of k-th harmonic. The implemented sinusoidal analysis system extracts the periodic part of the signal. This part is represented by sinusoidal parameters that are instantaneous frequency, amplitude, phase and frequency gradient. The scheme of the sinusoidal analysis operates as follows: first, the source signal is processed through windowing procedure in order to form analysis frames. Then the following instantaneous harmonic parameters estimation technique is applied. The signals bandwidth is separated into overlapping bands and instantaneous sinusoidal parameters are estimated in each band by analysis filter that is described in [5]. The values of instantaneous amplitude, frequency, and phase are evaluated as [5]:

$$A(n) = \sqrt{A^2(n) + B^2(n)}, \quad (3)$$

$$f(n) = \frac{\alpha(n+1) - \alpha(n)}{2\pi} F_s + F_c, \quad (4)$$

$$\varphi(n) = 2\pi F_c n + \alpha(n), \quad (5)$$

where

$$A(n) = \sum_{i=0}^{N-1} \frac{s(i)F_s}{(n-i)\pi} \sin\left(\frac{2\pi}{F_s} F_\Delta (n-i)\right) \cos\left(\frac{2\pi}{F_s} F_c i\right), \quad (6)$$

$$B(n) = \sum_{i=0}^{N-1} \frac{s(i)F_s}{(n-i)\pi} \sin\left(\frac{2\pi}{F_s} F_\Delta (n-i)\right) \sin\left(\frac{2\pi}{F_s} F_c i\right), \quad (7)$$

$$\alpha(n) = \arctan\left(-\frac{B(n)}{A(n)}\right), \quad F_c = \frac{F_2 + F_1}{2}, F_\Delta = \frac{F_2 - F_1}{2}, \quad (8)$$

F_1 and F_2 are frequency values that specify frequency band of the filter in Hz and N is the length of the analysis frame.
The estimation procedure involves iterative frequency recalculation with a predefined number of iterations. At every step the bandwidth of the filter is adjusted in accordance with the calculated frequency value in order to position energy peak in the centre of the band (see figure 2). At the initial stage the frequency range of the signal frame is covered by overlapping bandwidths $B_1, \ldots, B_h$ (where h is the number of frequency bands) with central frequencies $F_C^{B_1}, \ldots, F_C^{B_h}$ respectively. At every step the respective instantaneous frequencies $f^{B_1}(n_0), \ldots, f^{B_h}(n_0)$ are estimated at the instant n_0 that corresponds to the centre of the frame. Then the central bandwidth frequencies are reset $F_C^{B_x} = f^{B_x}(n_0)$ before the next iteration. After energy peaks localization (figure 2b) the final sinusoidal parameters (amplitude, frequency and phase) are estimated. Additional instantaneous frequency values are calculated with a specified time offset in order to estimate frequency gradient. During adjustment of the filter bands some of them may locate the same sinusoid. Duplicated parameters are discarded by comparison of estimated frequency values. To avoid estimation of transients by sinusoidal modelling evaluated parameters are tracked from frame to frame. The frequency and amplitude values of adjacent frames are compared in order to eliminate short sinusoidal components that apparently model the transient part of the signal.

3.2 Sinusoidal synthesis

The major steps of sinusoidal synthesis are the following. The phase values are matched between frames using cubic polynomial interpolation function. Exact phase matching obviously guarantees exact frequency matching. Having in-

stantaneous phase functions for every sinusoidal component the sinusoidal part of the frame can be synthesized using (1). Synthesized frames are concatenated with a specified window function and overlap.

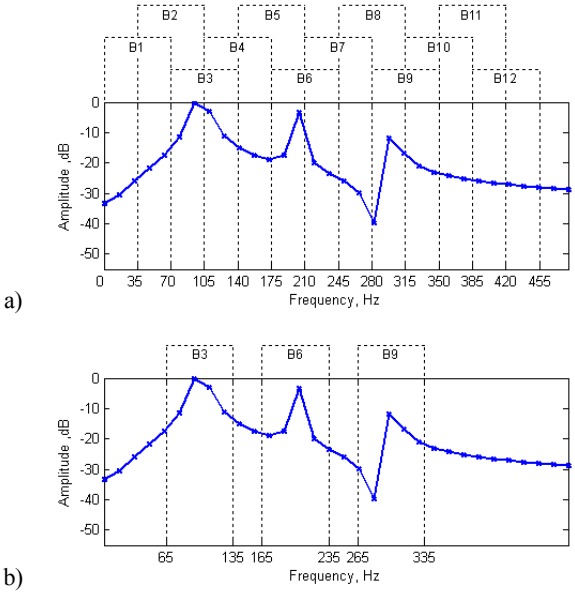

Figure 2 – Iterative filters adjustment (B1-B12 frequency bandwidths): a) – initial frequency band separation; b) – frequency band separation after the second iteration.

4. WAVEFORM MATCHING MODELS

4.1 Transients modelling using matching pursuit

Matching pursuit (MP) algorithms for compact representation of the transient part of the signal are used in several parametric audio encoding techniques [6,7]. The main task of MP procedure in application is to find a method for ranking and choosing most relevant component in the signal and selecting the function from the dictionary for compact input signal representation with minimal error. The optimization process of MP procedure can be based on the knowledge of psychoacoustic properties and human perception of a signal. It allows scaling the dictionary size according to auditory perception. The psychoacoustic adaptive criterion is used for assigning the dictionary elements to the individual segments in a rate-distortion optimal manner. Such techniques are successfully applied for damped sinusoid and wavelet packet (WP) [7,8].

4.2 MP with frame-based psychoacoustic optimized WP dictionary

From WP retrospective let's assume that $\{\varphi_n(t): n \in Z\}$ are scores of WP and $E \in \{(l,n): 0 \le l \le L, 0 \le n \le 2^l\}$ are the nodes of the WP tree structure. Then the interval $[0,1)$ is divided into dyadic intervals $I_{l,n} = [n2^{-l}, (n+1)2^{-l}]$ that correspond to the specific scores of nodes $(l,n) \in E$. Particularly $\{\varphi_{l,n,k}(t): (l,n) \in E, k \in Z\}$, where $\varphi_{l,n,k}(t) \triangleq 2^{-l/2} X_n(2^{-l}t - k)$ is a basic form in a signal space $span\{\varphi_0(t-k): k \in Z\}$. The node $(l,n) \in E$ of the WP tree is associated with the frequency band. According to the dyadic tree structure WP the signal is decomposed nearly into critical bands [9]: $(l,n) \in E_{CB}, l = \{0,b\}$, where E_{CB} describes limit WP tree structure, b is a maximum number of WP decomposition levels and depends on frequency range. E.g., for audio processing b is equal to 8. According to E_{CB} the spectral band [0–44.1 kHz] is divided into 25 subbands for audio. The root node $(l,n) = (0,0)$ of that tree corresponds to the full frequency range of audio signal.

The general MP algorithm can be described as an approximation of the analyzed signal $x(n)$ by linear expansion with atoms g_γ chosen from a WP-based dictionary D [7]. Each vector $g_\gamma \in D$ is indexed by $\gamma = (l,n,k)$, with $0 < l < log_2(N)$, $0 < n < 2^l$, $0 < k < 2^{-l}N$, where N is the signal frame length. Such vectors have a similar time-frequency localization properties as a discrete window function, dilated by 2^l centred at the $2^l(k+1/2)$.

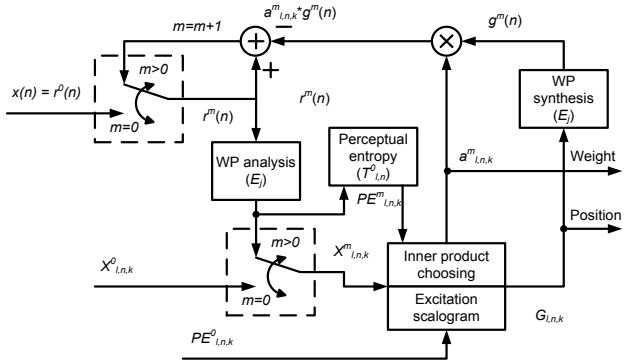

Figure 3 – The block diagram of the MP with frame-based psychoacoustic optimized WP dictionary.

The transients modelling method using the MP with frame-based psychoacoustic optimized WP dictionary consists of two stages. The first one is a frame-based auditory WP optimization based on the entropy cost function for the input signal $x(n)$ [9] and the second one is MP algorithm with perceptual criteria. At the first stage the results of the transient modelling are: the frame-based optimized WP tree E_j of the input signal $x(n)$; computed masking threshold $T_{l,n}$, temporal masker $F_{l,n}$ in nodes of WP tree structure E_j [9]; created auditory excitation scalogram associated with input signal $x(n)$ using $T_{l,n}$ and $F_{l,n}$ for all nodes. At the first MP procedure iteration (see figure 3), the input signal $x(n)$ is decomposed with the filter bank which implements the frame-based psychoacoustic adaptive WP tree. Each wavelet coefficient corresponds to the inner product of the input signal and an atom g_γ of the dictionary. The most relevant components can be found via selected perceptually relevant WP coefficients ranking [4]. Selecting the coefficients in the way that each new coefficient added provide maximum incremental gain in matching between the auditory excitation scalograms $G_{l,n,k}$ associated with the original and the modeled signals. The auditory excitation scalograms of original and modelled signals are constructed the knowledge of masking thresholds $T_{l,n}$ in wavelet domain. The selected WP coefficient with the maximum absolute value is chosen. The contribution of this vector $\alpha_\gamma^m \cdot g_\gamma^m(n)$ is then subtracted from

the signal $x(n)$ and the process is repeated on the residue $r(n)$. At the m-th iteration, the residue $r^m(n)$ is:

$$r^m(n) = \begin{cases} x(n) & m = 0 \\ r^{m+1}(n) + \alpha_\gamma^m \cdot g_\gamma^m(n) & m \neq 0' \end{cases} \quad (9)$$

where α_γ^m is the weight associated with the optimum vector $g_\gamma^m(n)$ at the m-th iteration, and γ^m is the WP-dictionary index at the m-th iteration. The optimum vector is the vector with the highest inner product and with the residual signal $\langle r^m, g_\gamma^m \rangle$. Each WP coefficient which has largest excitation weight is added to the modelled representation. The excitation weight is associated with difference between the reference WP coefficients excitation scalogram and the modelled excitation scalogram.

MP algorithm can be realized according to the following steps:

Input data: frame-based optimized WP tree structure E_j to the input signal $x(n)$; masking threshold $T_{l,n}$; temporal masker $F_{l,n}$ in nodes of E_j; auditory excitation scalogram $G_{l,n,k}$ associated with input signal $x(n)$.

- set the iteration number $m = 0$;

NEXT:

- allocate $G_{l,n,k}$ and set $G_{l,n,k} = 0$ for all l, n, k in correspondence with WP tree structure E_j;
- calculate $PE_{l,n}^m$ for all nodes (l, n), using $T_{l,n}$ [9];
- if $PE_{l,n}^m == 0 \; \forall \; (l, n, k) \in E_j$ then STOP
- if $PE_{l,n}^m == 0$, then $X_{l,n,k}^m = 0$ for $k = \{0, K_{l,n} - 1\}$ of node (l, n);
- select from $X_{l,n,k}^m$ the relevant coefficients $X_{l,n,k}^{*m}$ which has largest excitation weight;
- create auditory excitation scalogram associated with modeled signal using $T_{l,n}$ and $F_{l,n}^{m-1}$ for passed iteration and each new relevant coefficients $X_{l,n,k}^{*m}$;
- choose the weight $\alpha_{l,n,k}^m = X_{l,n,k}^{*m}$ which improve the matching between the reference excitation scalogram and the modelled excitation scalogram;
- get the position of chosen WP coefficient: $l^* = l$, $n^* = n$, $k^* = k$;
- set 1 at position (l^*, n^*, k^*): $G_{l^*,n^*,k^*} = 1$;
- synthesis of the atom $g^m(n)$ from G_{l^*,n^*,k^*} using inverse WP with the corresponding tree structure E_j associated with WP-dictionary;
- compute the residual signal $r^m(n)$ from $g^m(n)$ and $\alpha_{l,n,k}^m$ according to (9);
- apply the frame-based optimized WP with corresponding tree structure E_j to the residual signal $r^m(n)$;
- increase the iteration number $m = m + 1$;
- GO to NEXT.

The main advantage of the algorithm is perceptual distortion measure minimization defined in the frame-based perceptually optimized time-frequency tilling map of corresponding WP decomposition to select the optimum atom for each iteration of the pursuits. Furthermore, a psychoacoustic stopping criterion for the given procedure is presented. The number of MP algorithm iterations on the analysis frame is determined by quantity of the perceptually relevant WP coefficients in corresponding WP decomposition. A comparison of convergence behaviour between three different MP algorithms is shown in figure 4. The transient part is modelled by MP procedure using frame-based psychoacoustic optimized WP dictionary (dick solid line) has lower Mean-Square-Error (MSE) then another one based on the MP with over-complete WP dictionary (think solid line) and the MP with damped sinusoids (dashed line).

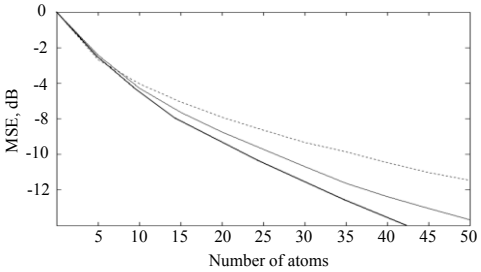

Figure 4 – A comparison of different MP algorithms.

4.3 Transient detection

The transient detection schema is based on the idea that energy of the residual signal (transient + noise) increases rapidly in the presence of a transient [7]. These changes may correspond with energy variations or energy redistribution among different frequency bands. The residual signal is transferred to the wavelet domain using 2 level WP decomposition. The algorithm computes the energy of the wavelet coefficients in each subband. The energy in each subband of frame i is divided by the energy of neighboring frames $(i-1)$ and $(i+1)$ and compared with a threshold. The threshold value depends on amplitude parameters, extracted at the sinusoidal analysis stage, in order to ignore masked transients.

5. EXPERIMENTS

An audio sound is used in order to show analysis system's performance. It is a bell tune that was sampled at 44100 Hz (figure 5(a),(b)). Each stage of the separation process is provided with the corresponding estimated part of the signal (as a spectrogram and a waveform) to give explicit presentation of the whole technique.

The sinusoidal analysis was carried out using the following features: analysis frame length – 48 ms, analysis step – 14 ms, filter bandwidth – 35Hz, windowing function – Hamming window. The synthesized periodic part is shown in figure 5(c),(d). As can be seen from the spectrogram, the periodic part contains only long sinusoidal components with high energy localization. The transients are left untouched in the residual signal that is presented in figure 5(e),(f). The periodic/residual ratio is rather high – 14.77 dB, that indicates that the most of the source signal's energy was represented by sinusoidal parameters.

Figure 5(g),(h) shows the transients components which was detected from residual part (figure 5(e),(f)), and modelled by proposed MP with frame-based auditory optimized dictionary algorithm. The input samples of residual signal (figure 5(e),(f)) were partitioned into frames of length 1024. In the experiments filters from Daubechies family with 40 coefficients were used.

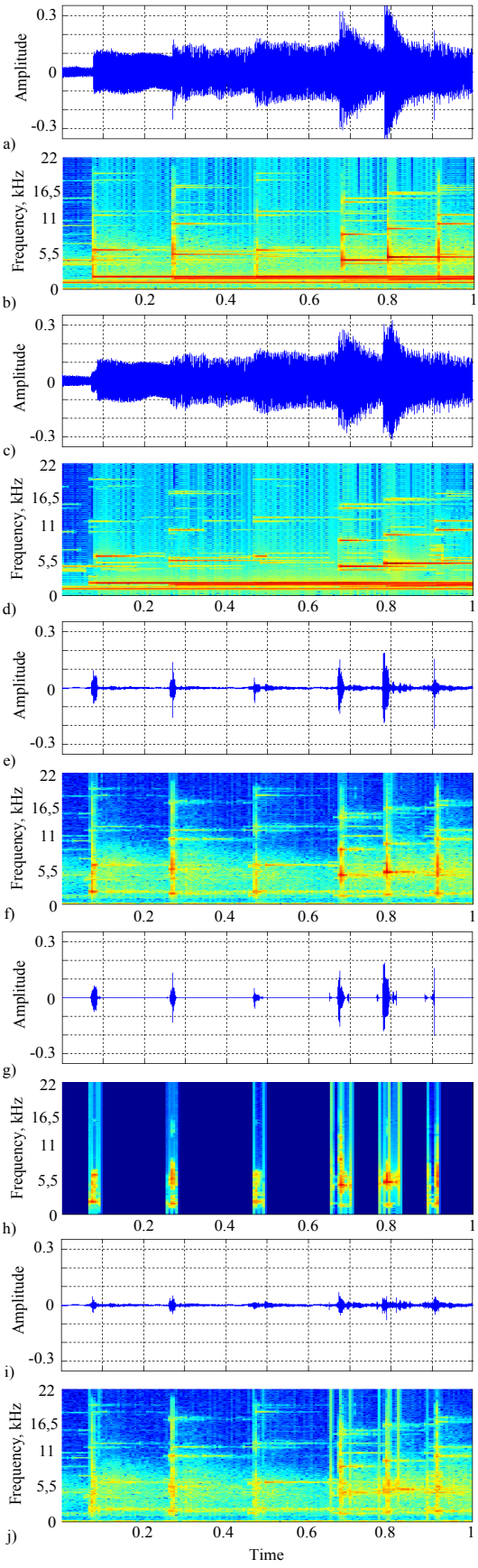

Figure 5 – Experimental results.

The reconstructed transients shown in figure 5(g) required 20, 23, 18, 32, 36, 25, 27 and 20 atoms correspondingly. The noise component is illustrated in figure 5(i),(j). The summation of the sines + transient + noise portions yields a signal that is perceptually indistinguishable from the original.

6. CONCLUSIONS

The advanced sinusoidal analysis with parameters tracking can properly process a signal without any prior detection; can accurately separate periodical part, saving noise and original transients in the residual. Making periodic separation first significantly simplifies further processing (especially transient detection). The proposed methodology for selecting most relevant wavelet coefficients is based on maximizing the matching between the auditory excitation scalograms associated with original and modeled signal correspondingly. The major advantage of this method is that the wavelet packet dictionary is perceptually optimized for each signal segment. It significantly reduces the number of coefficients required to achieve a given perceptual distortion.

7. ACKNOWLEDGMENT

This work was supported by the Polish Ministry of Science and Higher Education (MNiSzW) in years 2009-2011 (Grant no. N N516 388836).

REFERENSES

[1] A. Spanias, T. Painter, V. Atti, *Audio Signal Processing and Coding*. John Wiley & Sons, Inc., New Jersey, 2007.

[2] T. Quatieri R. McAulay, “Speech analysis/synthesis based on a sinusoidal representation", *IEEE Trans. on ASSP*, vol. 34(4), pp. 744-754, August 1986.

[3] Levine S., Smith J., “A Sines+Transients+Noise Audio Representation for Data Compression and Time/Pitch Scale Modifications”, *AES 105th Convention* (San Francisco, CA, USA), Preprint 4781, September 1998.

[4] Painter T., Spanias A., “Sinusoidal Analysis-Synthesis of Audio Using Perceptual Criteria”, *EURASIP Journal on Applied Signal Processing*, N1, pp. 15-20, 2003.

[5] E. Azarov, A. Petrovsky, M. Parfieniuk. “Estimation of the instantaneous harmonic parameters of speech” in *Proc. EUSIPCO-2008*, Lausanne, August 2008, (CD ROM)

[6] S. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, *IEEE Trans. on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[7] P. Vera-Candeas, and etc., “Transient modeling by Matching-Pursuits with a wavelet dictionary for parametric audio coding”, *IEEE SP Letters*, Vol. 11, No. 3, pp. 349-352, March 2004.

[8] T. S. Verma, *A perceptually based audio signal model with application to scalable audio compression*, PhD thesis, Standford University, 1999.

[9] A. Petrovsky, D. Krahe, A. A. Petrovsky, “Real-Time Wavelet Packet-based Low Bit Rate Audio Coding on a Dynamic Reconfigurable System”, *AES 114th Convention,* Amsterdam, preprint 5778, 22p., May, 2003.