Bio-Inspired Sparse Representation of Speech and Audio Using Psychoacoustic Adaptive Matching Pursuit

Al. Petrovsky, V. Herasimovich, A. Petrovsky

speaker: Vadzim Herasimovich

Department of Computer Engineering,

Belarusian State University of Informatics and Radioelectronics,

Minsk, Belarus

1. Introduction

Ideas of the Research:

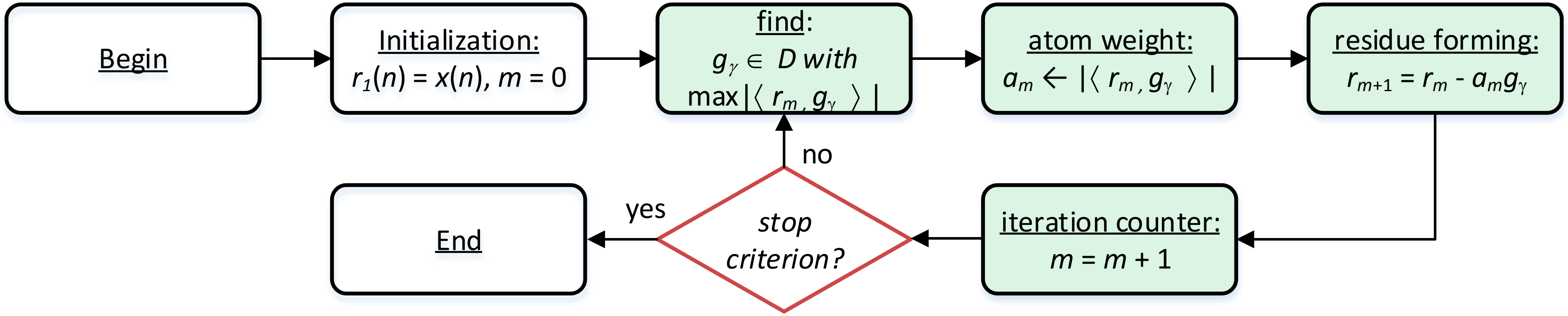
- Use sparse signal representation based on the matching pursuit (MP);
- Wavelet packet (WP) based dictionary;
- Dictionary adaptation using human auditory system properties;
- Psychoacoustically motivated parameters selection;

Application of the Research:

- Scalable audio/speech coding algorithm development;
- Single transform domain regardless of the nature of the input signal;
- High quality with low bitrates;
- Universality for all known types of audio content;
- Real-time processing.

2. MP using WP Dictionary

Common MP[1] Procedure:



WP Based Dictionary of Time-Frequency Functions:

WP-based dictionary ⟵ $g_\gamma \in D, \gamma = (l, n, k)$

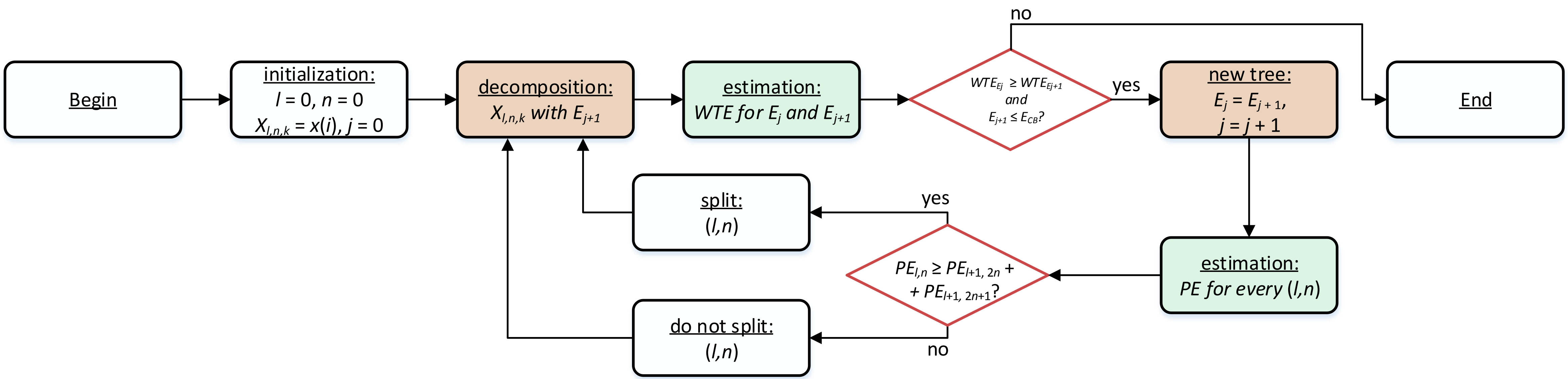
WP tree structure ⟵ $E \in \{(l, n): 0 \le l \le L, 0 \le n \le 2^l\}$

where l – WP tree level number, n – WP tree node number, k – coefficient index

[1] S. Mallat, Z. Zang, "Matching Pursuits with Time-Frequency Dictionaries", IEEE Transactions on signal processing, vol. 41, pp. 3397-3415 (1993 December).

3. Adaptive WP analysis

WP Decomposition (WPD) Tree Growth Algorithm:



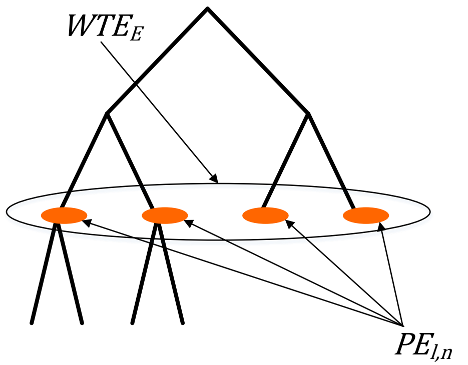
Adaptation Cost Functions:

Wavelet Time Entropy

$$WTE_{E_i} = -\sum_{\forall(l,n)\in E_i}\sum_{k}\frac{|X_{l,n,k}|}{\sum_{\forall(l,n)\in E_i}|X_{l,n,k}|}\ln\left(\frac{|X_{l,n,k}|}{\sum_{\forall(l,n)\in E_i}|X_{l,n,k}|}\right)$$

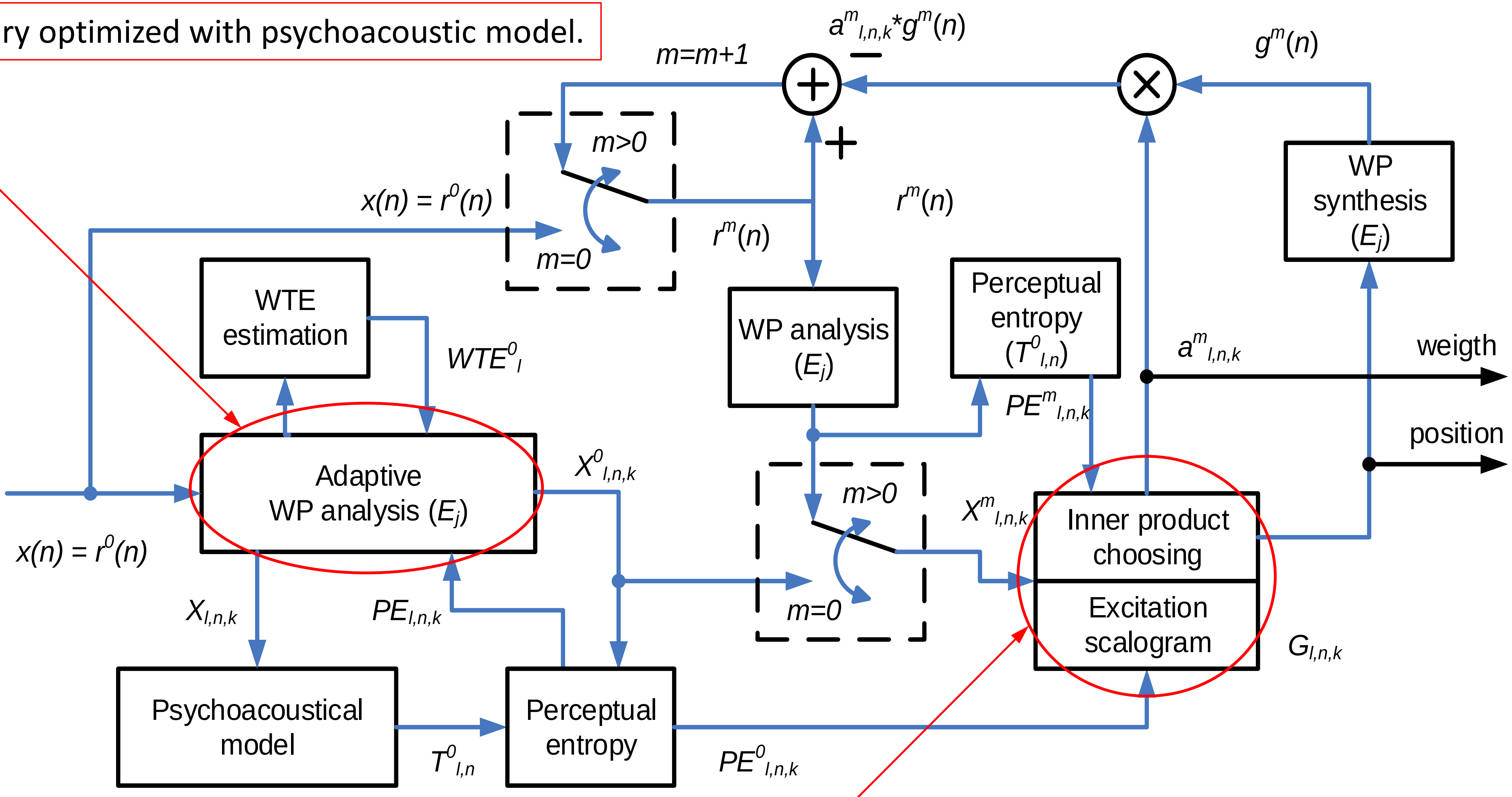
Perceptual Entropy

$$PE_{l,n} = \sum_{k=1}^{K_{l,n}-1}\log_2\left(2\left[nint\left(|X_{l,n,k}|/\Delta_{l,n}\right)\right]+1\right)$$



4. MP procedure with Optimized Dictionary

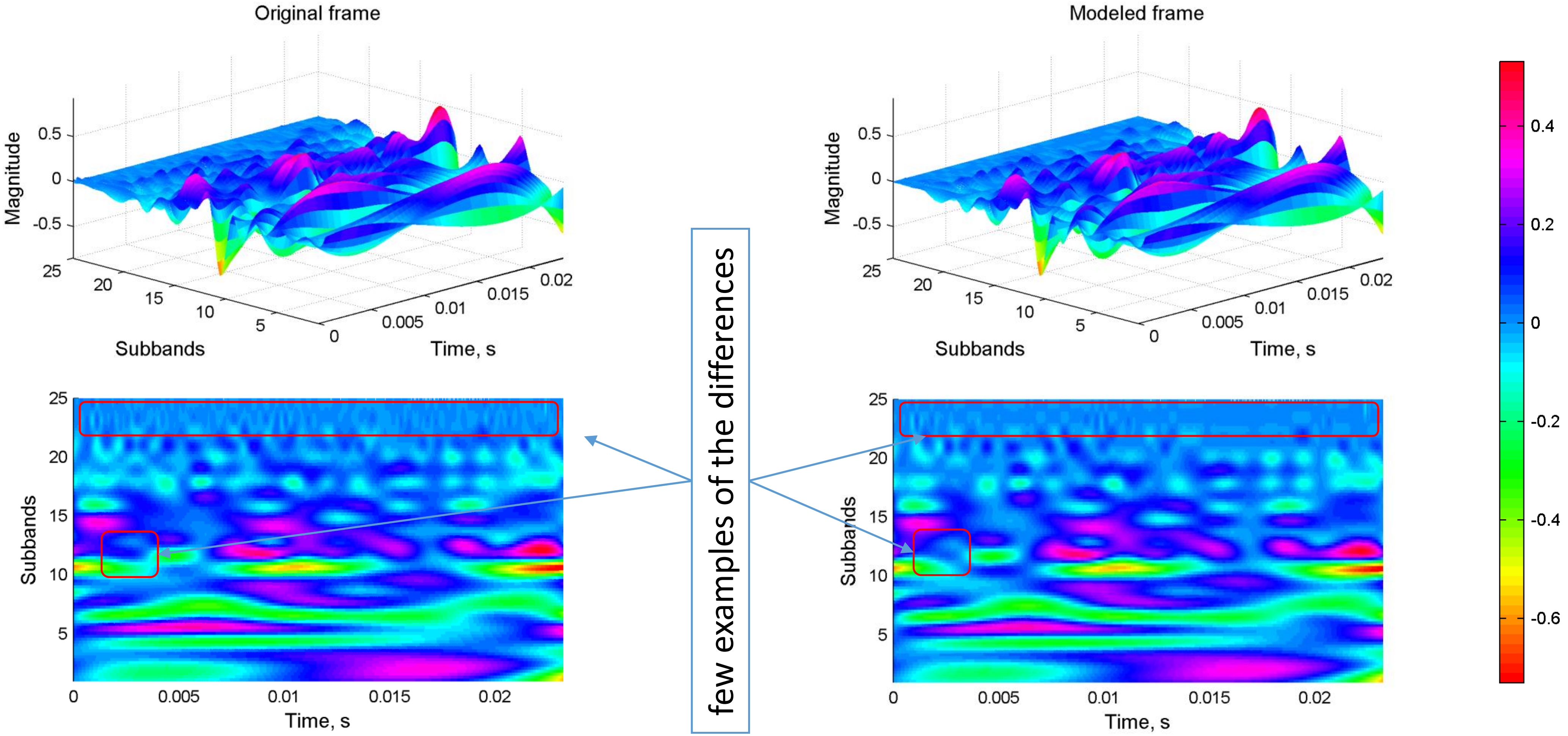
Dictionary optimized with psychoacoustic model.



Choosing and selecting most relevant coefficients, which has largest excitation weight.

5. Excitation Scalogram

Masking thresholds[3] $T_{l,n}$ and temporal maskers[4] $F_{l,n}$ are used for excitation scalogram estimation.

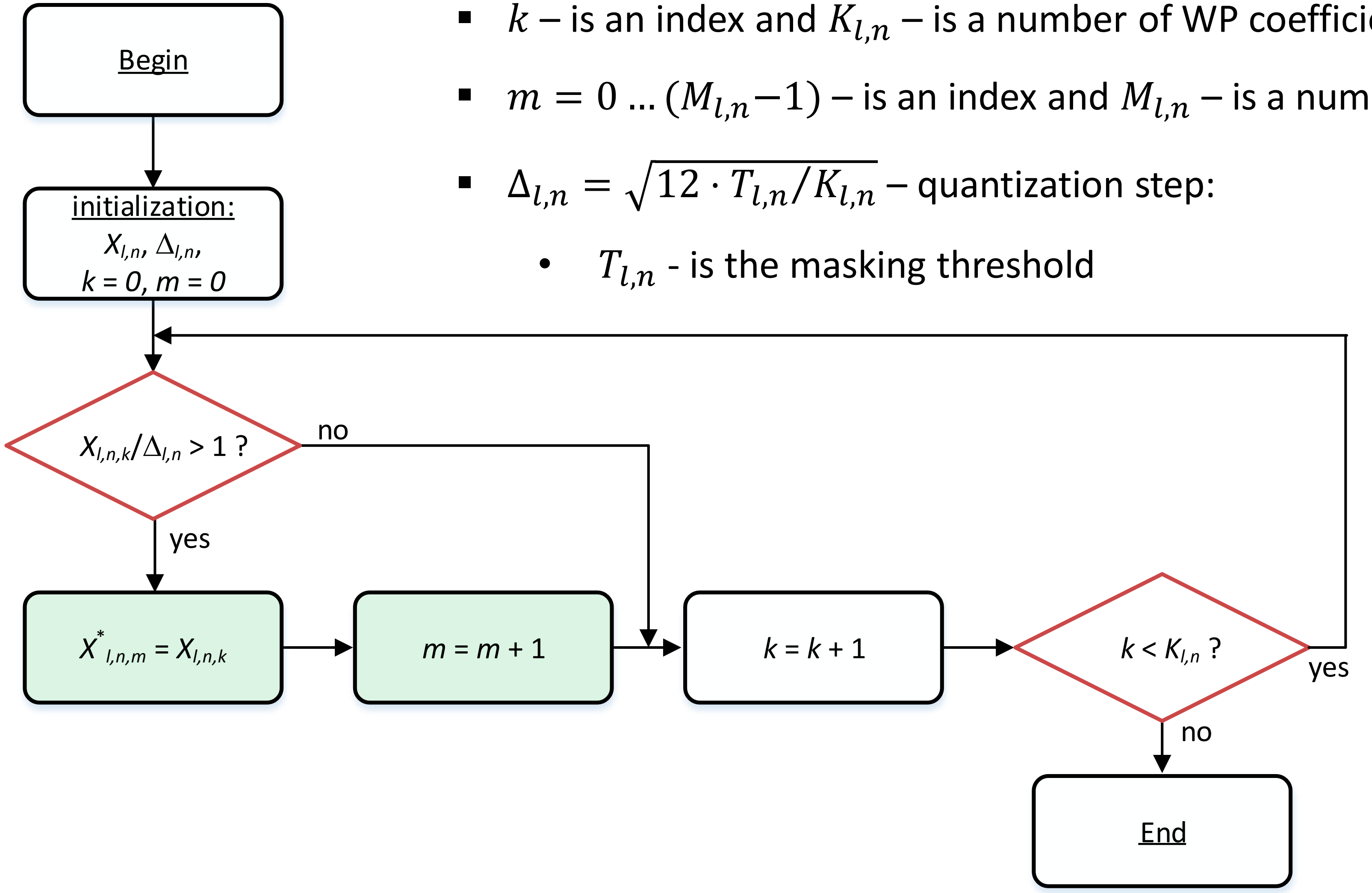


[3] A. Petrovsky, D. Krahe, A.A. Petrovsky, "Real-Time Wavelet Packet-based Low Bit Rate Audio Coding on a Dynamic Reconfigurable System", presented at the AES 114th Convention, Amsterdam, The Netherlands, 2003 March 22-25.

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

6. Time-Frequency (T-F) Plan Adaptation

- k – is an index and $K_{l,n}$ – is a number of WP coefficients,
- $m = 0 \ldots (M_{l,n}-1)$ – is an index and $M_{l,n}$ – is a number of the chosen coefficients $X^*_{l,n,m}$,
- $\Delta_{l,n} = \sqrt{12 \cdot T_{l,n} / K_{l,n}}$ – quantization step:
  - $T_{l,n}$ - is the masking threshold

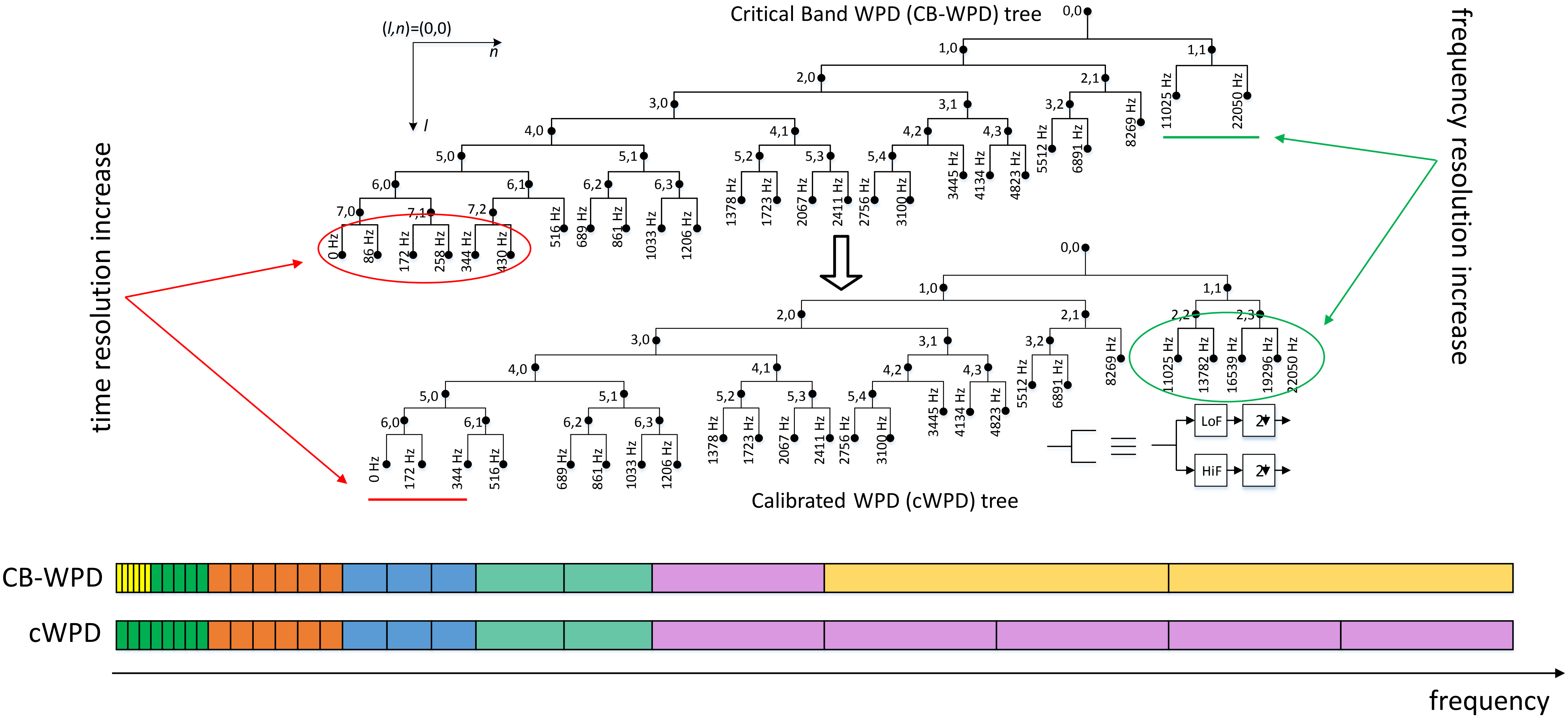


Procedure output

$X^*_{l,n,m}$, $M_{l,n}$ for PE estimation tuning

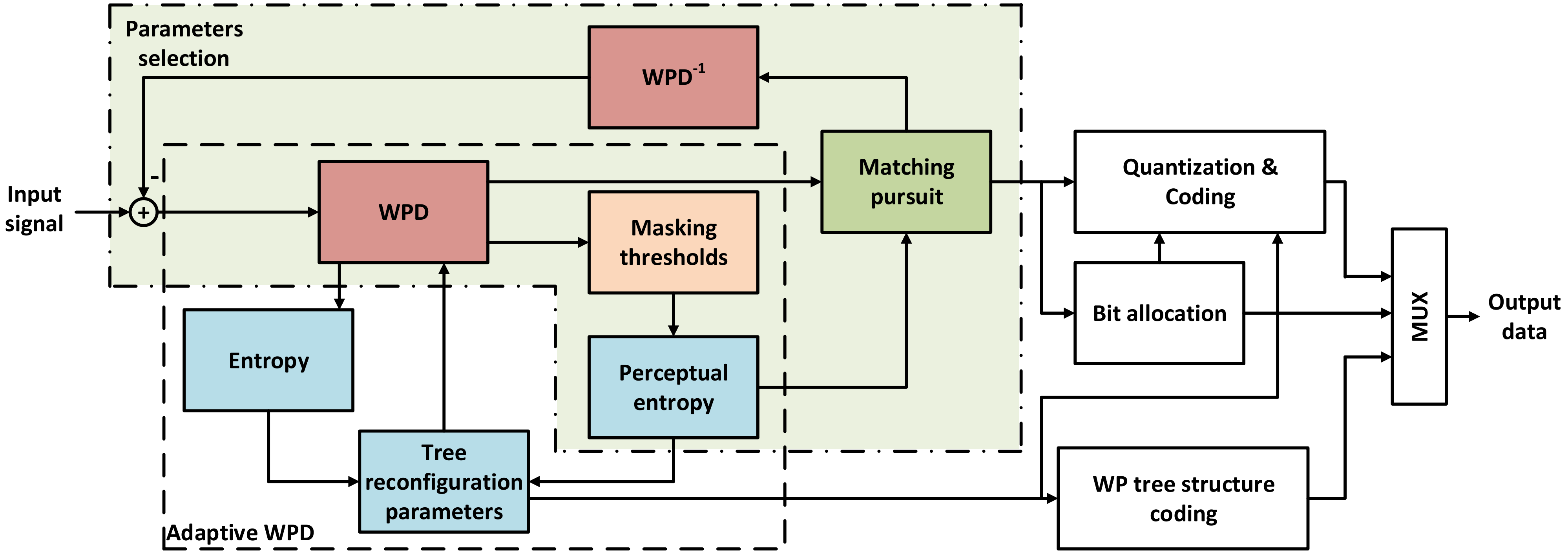
Calibrated WP tree

7. T-F Plan Adaptation



8. Speech and Audio MP Coding Scheme

Encoder Structure:



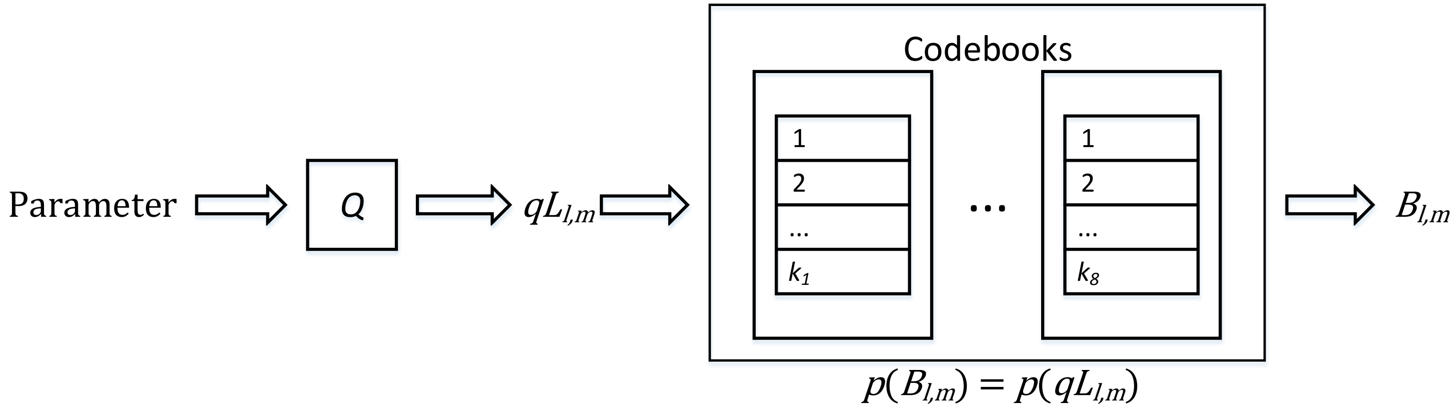
Decoder Structure:

9. Parameters Quantization & Coding

Parameter Quantization & Coding:

Quantized parameter ⟸ $$qL_{l,n,m} = 2\left|nint\left(\left|\frac{X^{*}_{l,n,m}}{\Delta_{l,n}}\right|\right) + 1\right|$$

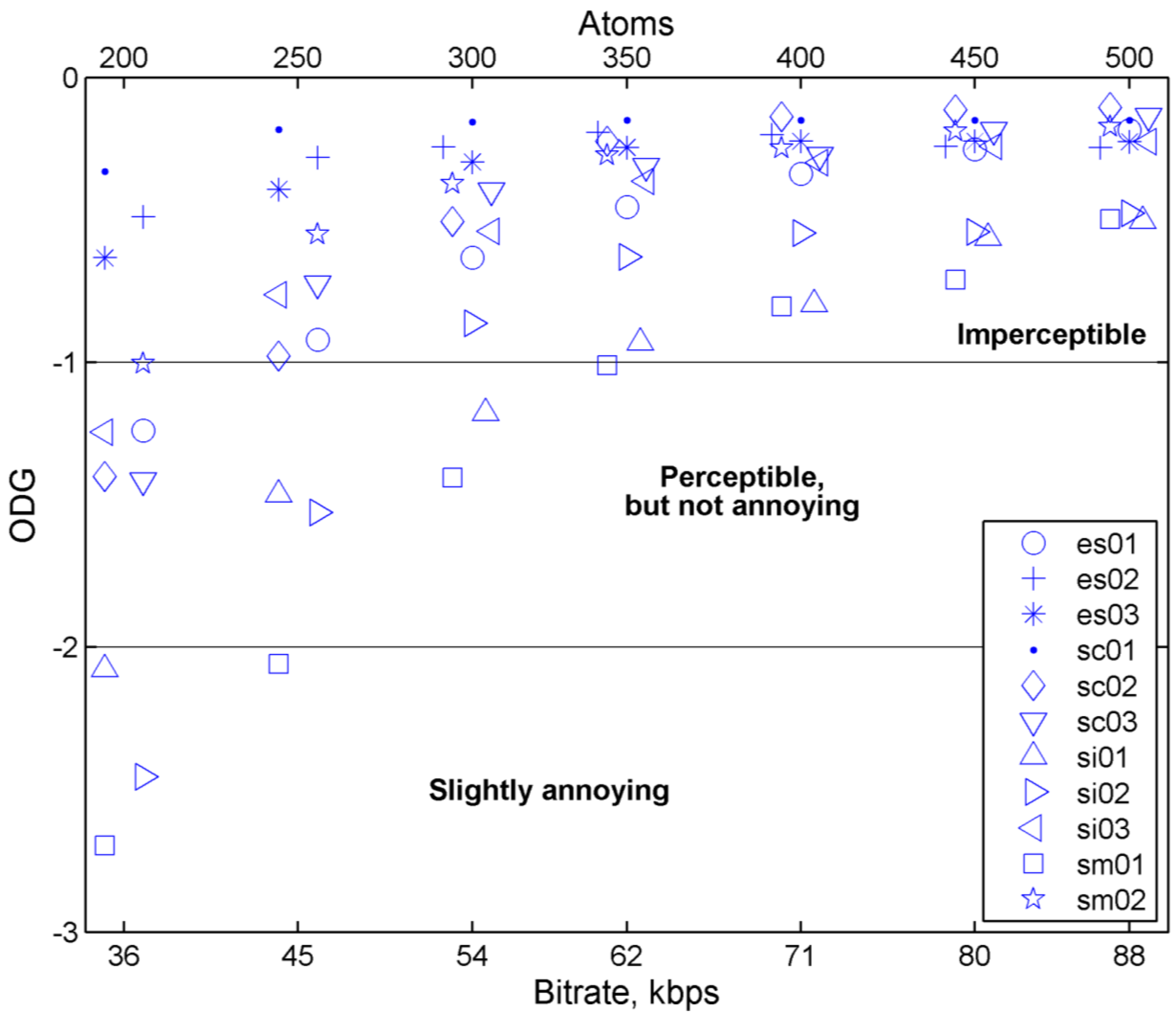
- $qL_{l,n,m}$ encoded using Huffman algorithm.
- $B_{l,m} = \left(b_{m,1}, b_{m,2}, b_{m,3}, \dots, b_{m,w_k}\right), b_{m,j} \in \{0,1\}, j = \overline{1, w_m}.$



WP Tree Coding:

#	Previous structure E_{i-1}	Code	New structure E_i
1	(l,n) (l+1,2n) (l+1,2n)	**00** *no changes*	(l,n) (l+1,2n) (l+1,2n)
2	(l,n) (l+1,2n) (l+1,2n)	**01** *delete*	(l,n)
3	(l,n)	**10** *grow*	(l,n) (l+1,2n) (l+1,2n)
4	**0000**	**11**	

10. An Objective Assessment of the Sound Quality



For the objective quality assessment *PEMO-Q*[5] model was used.

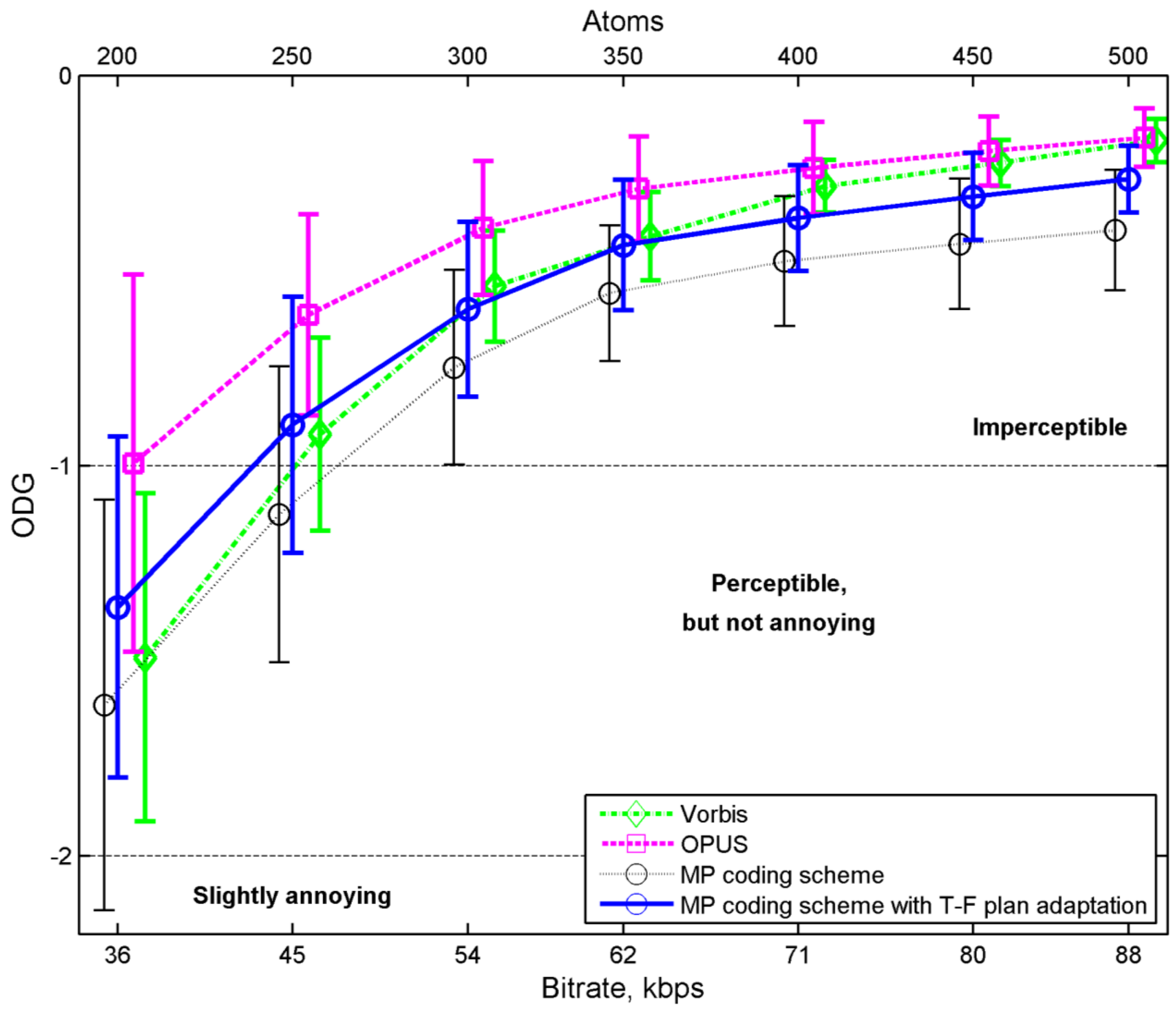
Impairment description	ODG
Imperceptible	0.0
Perceptible, but not annoying	-1.0
Slightly annoying	-2.0
Annoying	-3.0
Very annoying	-4.0

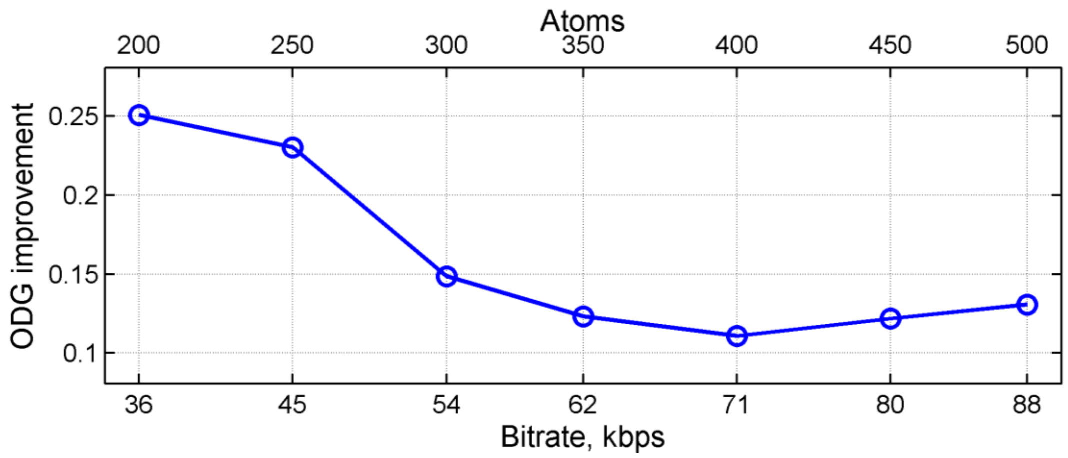
Test sequence: mono, 44.1 *kHz* sampling rate, 16-bit resolution

№	Test item	Description
1	es01	Vocal (Suzan Vega)
2	es02	German speech
3	es03	English speech
4	sc01	Trumpet solo and orchestra
5	sc02	Orchestra piece
6	sc03	Contemporary pop music
7	si01	Harpsichord
8	si02	Castanets
9	si03	Pitch pipe
10	sm01	Bagpipes
12	sm02	Plucked strings

[5] R. Huber, B. Kollmeier, “PEMO-Q– A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception”, IEEE Transactions on audio, speech, and language processing, vol. 14, pp. 1902-1911 (2006 November).

11. Overall quality comparison



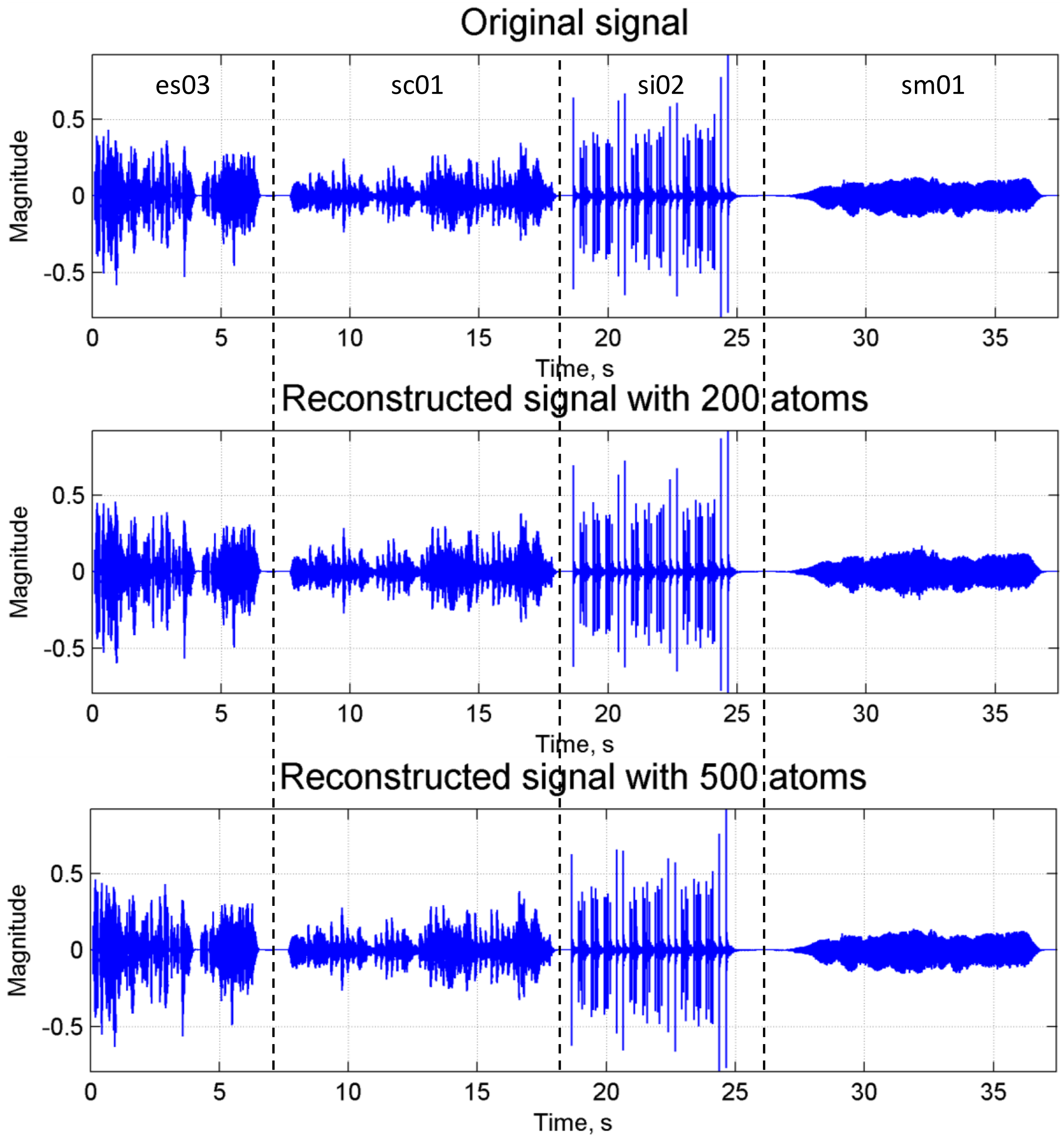


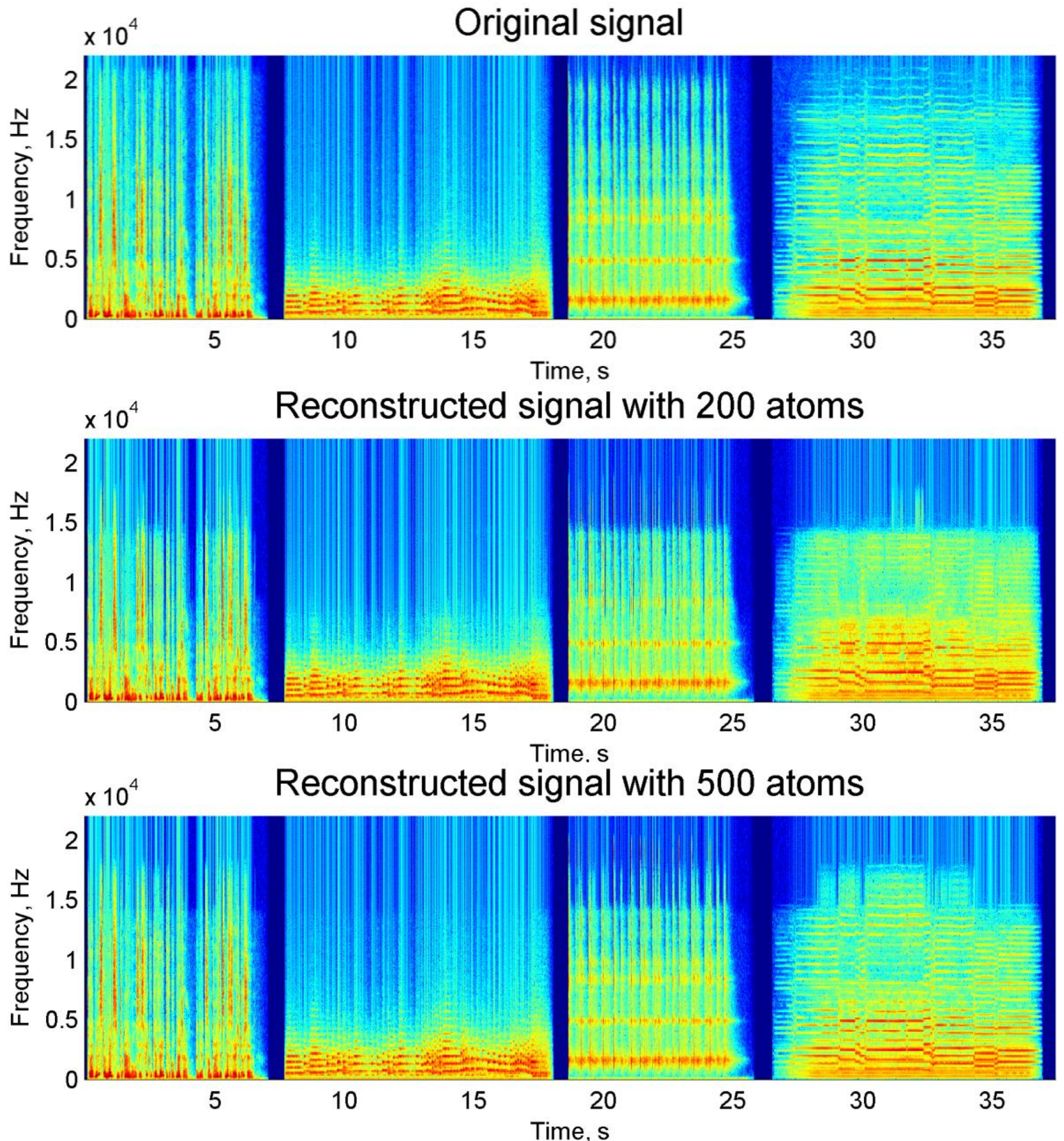
ODG improvement =

= MP coding scheme with T-F adaptation -

- MP coding scheme

12. MP Coding Scheme Results





13. Conclusions & Future Research

Conclusion:

- Bio-inspired sparse representation model of speech and audio signals has been proposed;
- Time-frequency plan adaptation and its impact to the signal modeling was shown;
- MP audio/speech encoding algorithm as an application of the proposed model was described.

Future Research:

- Optimization of the MP procedure to further improve of the model;
- Encoding algorithm coding and quantization scheme improvement;
- MP audio/speech coder implementation as a field programmable system-on-chip (FPSoC).

Acknowledgements

IT FOR YOU

MOST

Thank you for your attention!