Lecture 14. Design of audio WM

Model of digital audio CO (WAV-format):

- dependent samples with frequency 44,1κHz,
- amplitude of samples is quantized on 2^{16}=65536 levels (16 bits)

Remark 1. There exists a lossy compression of WAV to mp3-format.

Remark 2. Another model of audio CO is commercial speech signals.

All methods of WM embedding and extraction considered before can be applied to audio CO however a design of audio WM systems has some peculiarities :

- CO (music first of all) lose their value even after small corruptions (Gaussian noise, filtering, pulse noise ctr.)
- In design of audio WM it is necessary to take into account so called HAS (*Human Auditory System* [20])

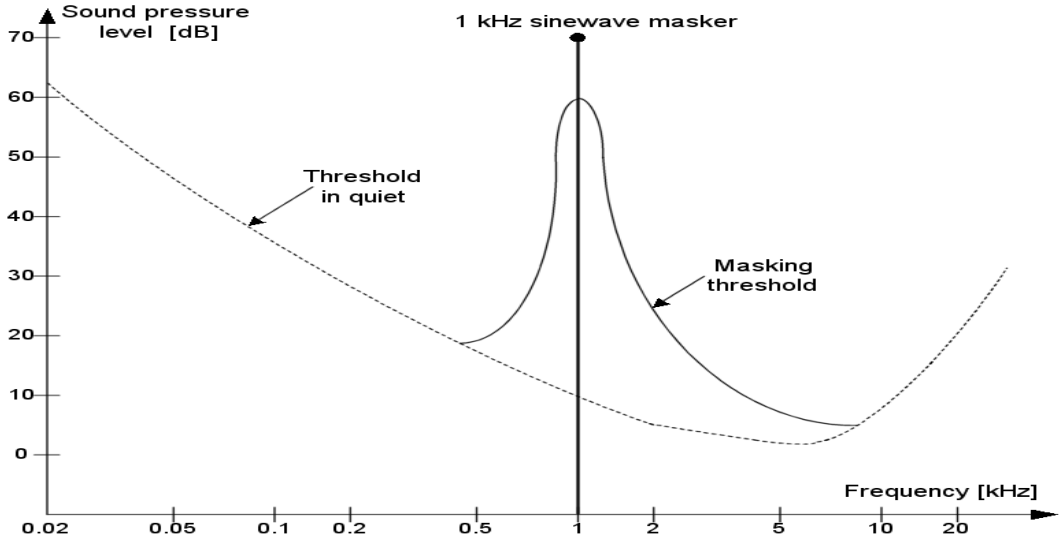
Remark 3. For video WM it is also necessary to take into account so called HVS (*Human Visual System* [20]), but its impact on such design is not so large because audio signal has large frequency range .

Some properties of HAS

1. HAS is modeled usually by bank of filter overlapping on frequency with band width of about 200Hz for central frequencies lower 500Hz and with band width of about 5000 Hz for central frequencies within the range 500-24000 Hz.
2. Sensitivity of HAS to additive Gaussian noise is very large (it can be heard even with level 70dB lower than the level of some other sound)
3. HAS has *masking property on frequency.*
4. HAS has *masking property in time.*
5. HAS is tolerable to *phase spectrum* of audio signal
6. HAS perceives addition of some *echo signals* as a changing of acoustic environment.

Consider some properties more detail.

3. Frequency masking Signal with small amplitude on the frequency f_m is not heard (masked) in the presence of another signal with larger amplitude on the frequency f_o, that is close to the first frequency.



Here masking frequency is 1κHz

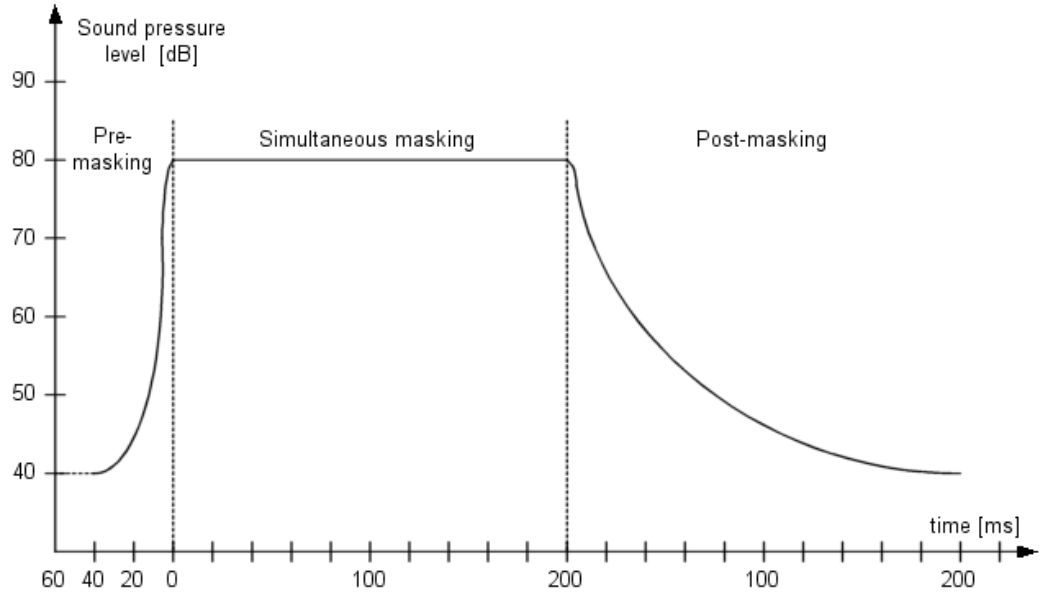
Threshold quiet-is such level of audio signal (depending on its frequency) that this signal is not heard with lower amplitude.

Masking threshold – is the level of masked signal (depending on the frequencies of both masked and masking signals) that signal is not heard with amplitude below this threshold.

Conclusion: Audio signal can be masked even if its amplitude is larger than threshold quiet , moreover the signals on high frequencies are masked better than on lower frequencies.

4. Time masking

Additional signal (ahead or delayed with respect to the main signal) is not heard if these differences in time are not so large.



Conclusion: Delayed signal is masked worse than ahead signal.

The main embedding methods for audio WM:

1. Embedding in LSB.
2. Modified embedding in LSB.
3. Embedding in phase domain.
4. Embedding with the use of echo signals.
5. The use of SS in frequency or in time domain.
6. The use of ISS in frequency or in time domain.
7. The use of QPD in frequency or in time domain.
8. Embedding based on a modulation of some characteristics of audio signal.

Consider further some of these methods.

1. Embedding in LSB:

The samples with LSB embedding are chosen by stegokey.

Under the attack these samples are unknown but a randomization of all samples can result in large distortions of CO

Remark 1: Trade off between embedding rate , quality of CO and efficiency of attack are commonly for this method.

Remark 2: It is commonly for audio CO to embed messages not in one least significant bit of samples but in several least significant bits.

2. Modified embedding in LSB:

The idea is : change LSB in additional samples which have been not chosen by stegokey if it results in reducing of CO distortions .

Let us denote :

C(n) – the amplitude of CO sample before embedding,

$C'_w(n)$ –the amplitude of CO sample after embedding of k LSB

$C''_w(n)$ –the amplitude of CO sample after embedding of k+1 LSB.

$C_w(n)$ – final amplitude after embedding of k LSB

$e'(n) = |C(n) - C'_w(n)|$ *-the error under embedding in k LSB*

$e''(n) = |C(n) - C''_w(n)|$ *-the error under embedding in k+1 LSB*

Embedding:

$$C_w(n) = \begin{cases} C'(n)_w, \text{if } e'(n) \le e''(n) \\ C''(n)_w, \text{if } e'(n) > e''(n) \end{cases}$$

3. Embedding in phase of CO

Justification: HAS is tolerable to phase shift of audio signal.

a) *WM embedding:*

1. Samples of audio signal are splited on the blocks of fixed length.
2. To each of blocks is applied DFT.
3. WM embedding is provided by phase shifting of DFT.
4. It is provided «a smoothing» of transitions between blocks of DFT because there may have jumps of phases just after WM embedding.
5. It is performed IDFT in order to form the WM-ed signal finally.

b) WM extraction:

1. Samples of audio stegosignal are divided on the blocks of fixed length , moreover the bounds of these blocks have to coincide exactly with the bounds of blocks under embedding.
2. DFT is applied to each of blocks.
3. WM is extracted by digital phase detector.

Embedding rate: 8-32 bits/s under a good quality of CO after embedding.

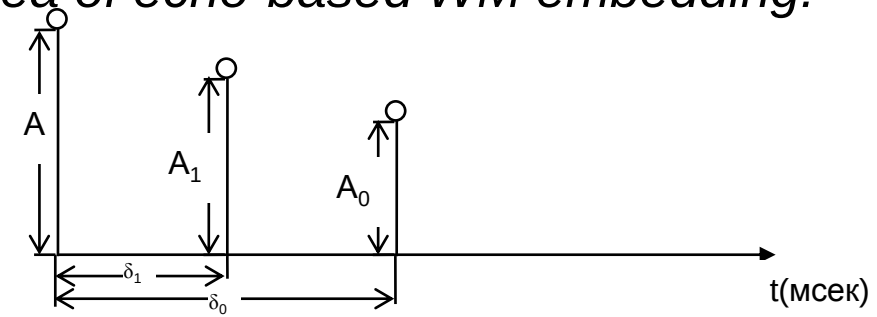
Improved method: Divide signal into frequency domains and embed in DCT phases for each of these domains separately .Experiment shows that date embedding rate can be increased then up to 1200 bit/s.

Our experiments shown that embedding in phase domain can not be taken for granted against WM removal attack.

4. Echo watermarking

Justification: Addition of «echo» (original signal shifted in time) to the main CO is considered by HAS not as additive noise but like to appearance of additional *resonances* (similar to changing of acoustic environment under creation of audio signal).

The idea of echo-based WM embedding:



A – amplitude of original audio signal,

A_1 –amplitude of echo signal corresponding to embedding of the bit «1».

A_0 –amplitude of echo signal corresponding to embedding of the bit «0».

δ_1 – delay of echo corresponding to the bit «1».

δ_0 – delay of echo corresponding to the bit «0».

The parameters which affect on the CO quality after embdding:

$$\eta_1 = \frac{A_1}{A}, \eta_0 = \frac{A_0}{A}, \delta_1, \delta_0.$$

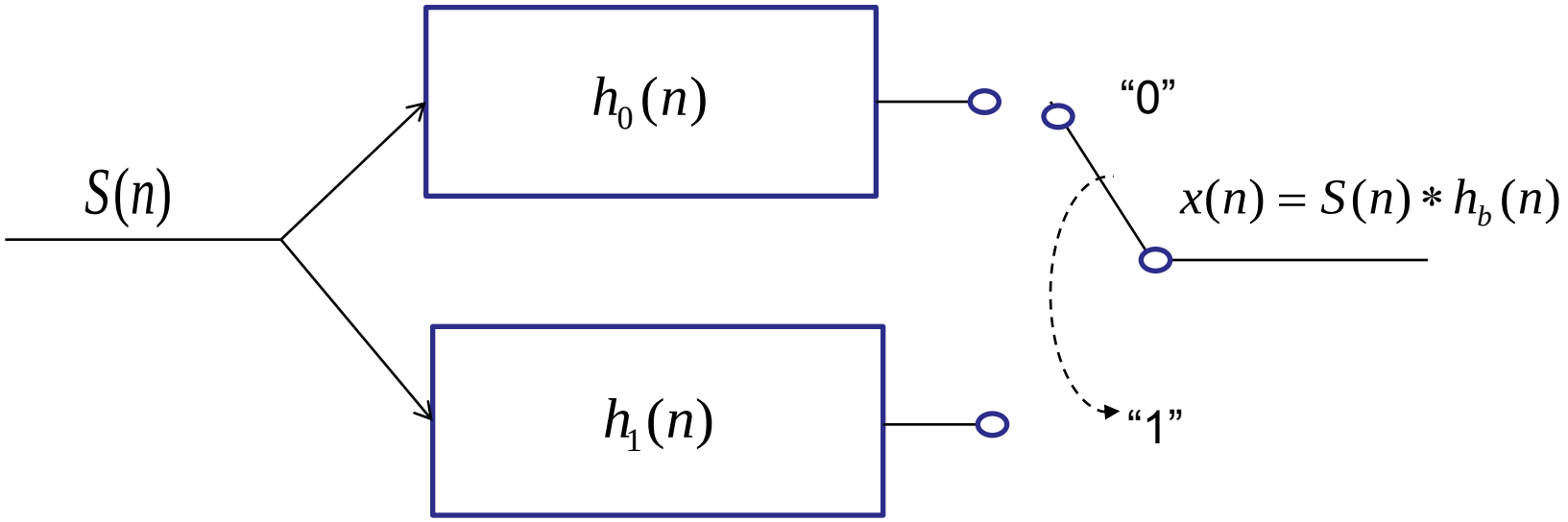
a) WM embedding:

1. Split samples of audio signal into blocks of the fixed length.
2. Embed WM on each of block using echo signals.
3. Smooth transition areas (as a consequence of echo signal embedding) between blocks .

b) WM extraction :

1. Split samples of audio signal into blocks of fixed length for which the bounds have to be coincide with the bounds of blocks after embedding.
2. Calculate autocorrelation function(ACF) for each of blocks.
3. Compare the side peaks of ACF with some threshold .Take decision about embedding of the corresponding bit depending on a disposition of peaks exceeding to the threshold.

WM embedding based on echo hiding



$$x(n) = S(n) * h_b(n) \ , \mathrm{n} = 1,2...\mathrm{N}, \qquad (1)$$

where

$S(n)$ - input audio signal

$h_b(n)$ - filter pulse response corresponding to, embedding bit

$\mathrm{b} = (0,1)$

$*$ - convolution

Complex cepstrum

Complex cepstrum $\tilde{S}(n)$ of signal $S(n), \mathrm{n} = 0,1,..., \mathrm{N}\text{-}1$ is determined as follows :

$$\tilde{S}(n) = \frac{1}{N}\sum_{k=0}^{N-1}[\log|S(k)| + j\Theta(k)]e^{-\frac{2\pi jnk}{N}} \qquad (2)$$

where $S(k) = \sum_{n=0}^{N-1} S(n) \cdot e^{-\frac{2\pi jnk}{N}}$, $|S(k)|$ - signal amplitude

$\Theta(k)$ - signal phase

Correlation receiver based on cepstrum

$$\tilde{x}(n) = \tilde{S}(n) + \tilde{h}_b(n),\ \mathrm{n} = 1,2,... \tag{3}$$

$$\sum_n \tilde{x}(n) \cdot \tilde{h}_0(n) \underset{b1}{\overset{b0}{\gtrless}} \sum_n \tilde{x}(n) \cdot \tilde{h}_1(n) \tag{4}$$

Under the condition

$$\sum_n \tilde{h}_0(n) \cdot \tilde{h}_1(n) \approx 0, \sum_n \tilde{h}_0^{\,2}(n) \approx \sum_n \tilde{h}_1^{\,2}(n) \tag{5}$$

The following relation holds for gaussian model:

$$p = 1 - F\left(\frac{\sum_n \tilde{h}_0^{\,2}(n)}{\sqrt{D}}\right) \tag{6}$$

where

$$D = Var\left\{\sum_n \tilde{S}(n) \cdot \tilde{h}_\Delta(n)\right\}, \tilde{h}_\Delta(n) = \tilde{h}_0(n) - \tilde{h}_1(n),\ F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \tag{7}$$

Correlation receiver on subintervals

$$\sum_{k=1}^{L}\sum_{n}\tilde{x}_k(n)\cdot\tilde{h}_0(n) \underset{b1}{\overset{b0}{\gtrless}} \sum_{k=1}^{L}\sum_{n}\tilde{x}_k(n)\cdot\tilde{h}_1(n) \qquad (8)$$

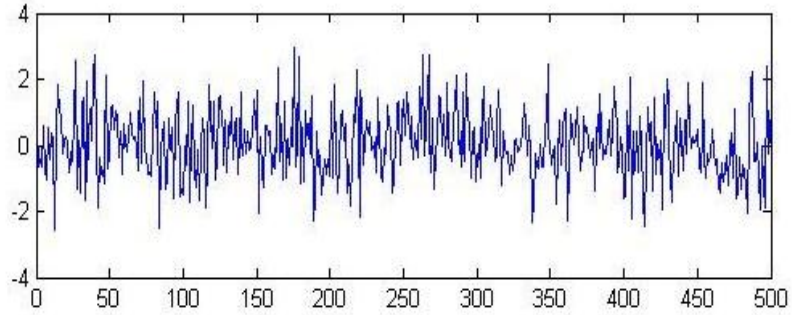
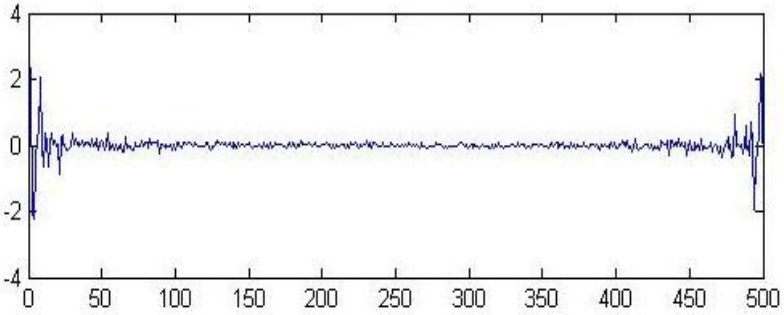
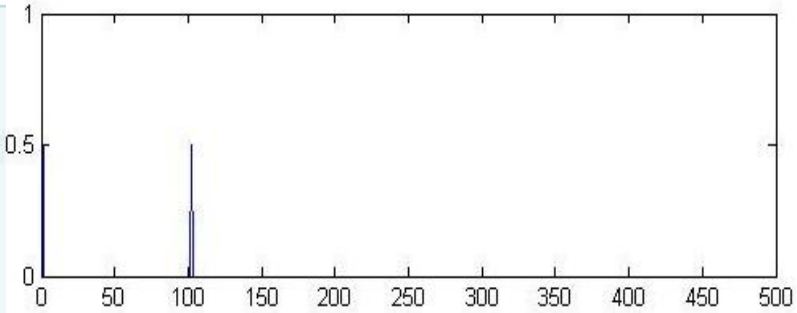
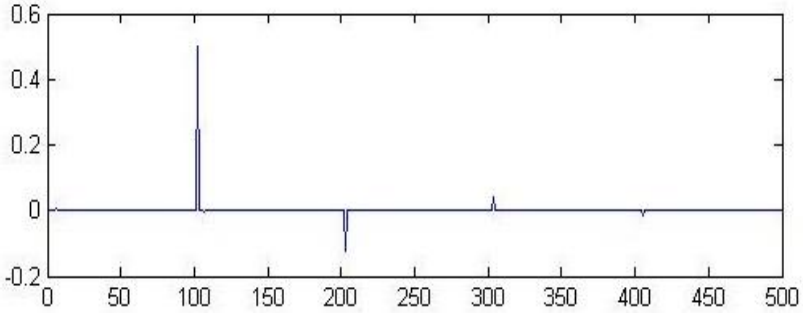
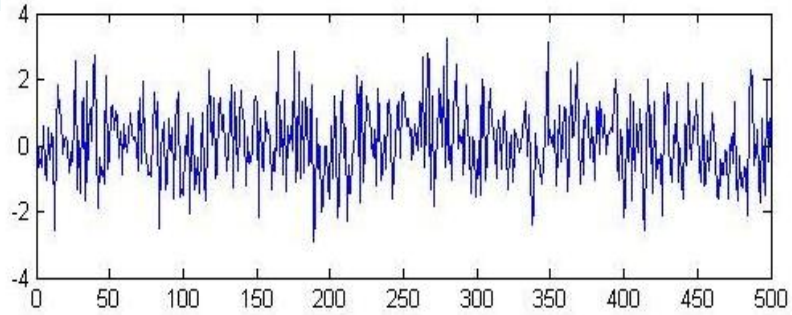
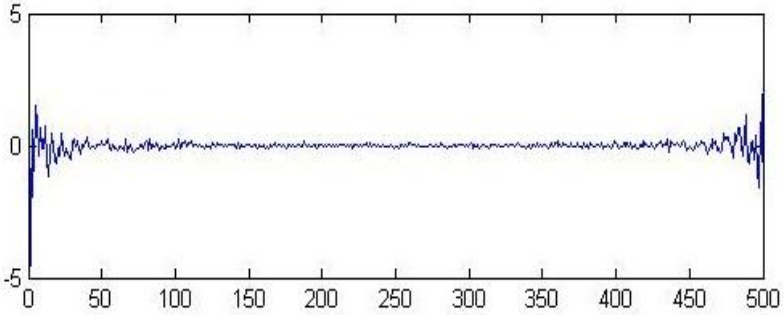
where $\tilde{x}_k(n)$ - Cepstrum of WM-ed signal computed on the *k*-th subinterval,

L – the number of subintervals on „bit" interval N.

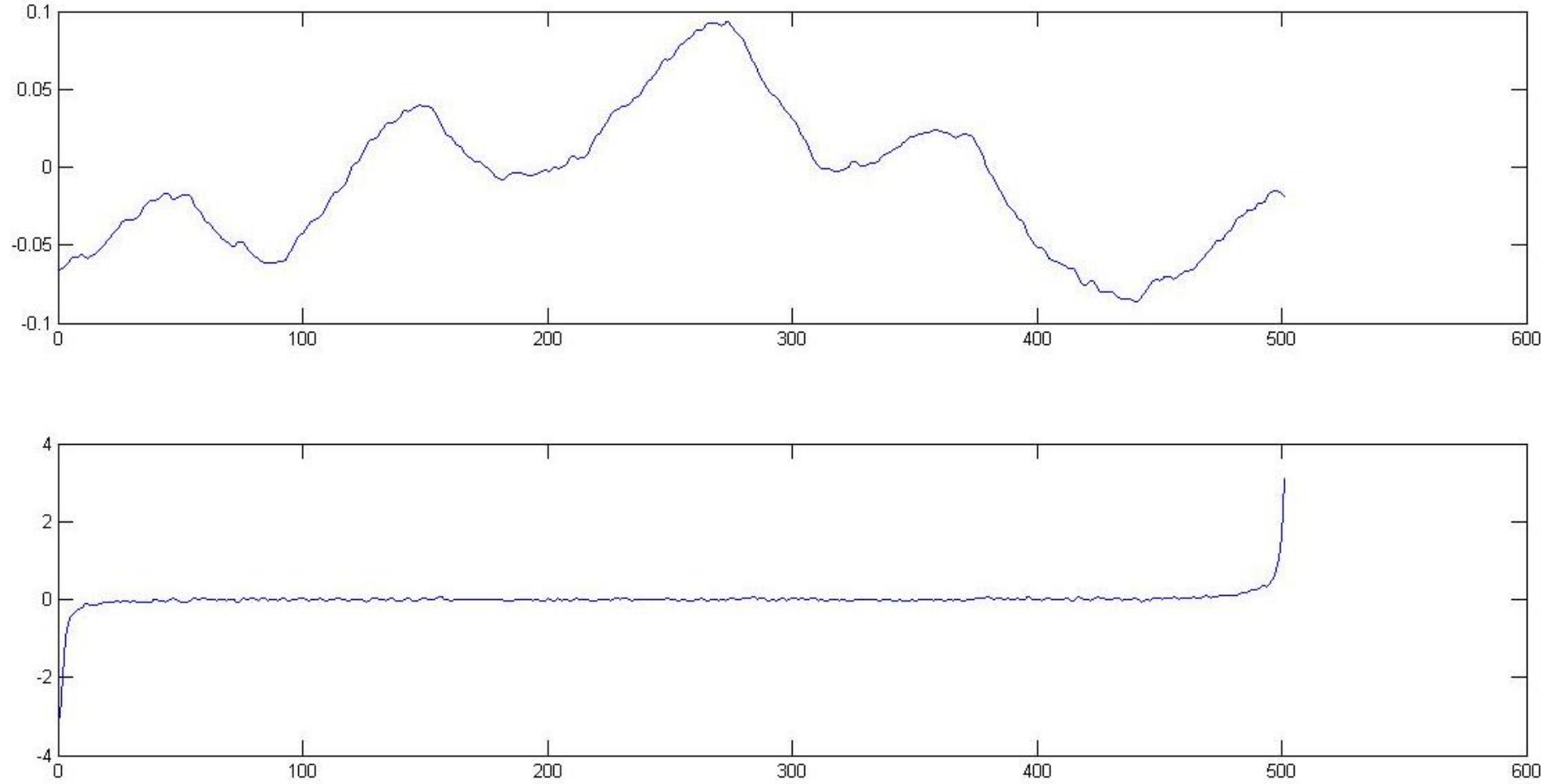
Then we get the following formula for the probability of error :

$$p = 1 - F\left(\sqrt{\frac{L\cdot\sum_{n}\tilde{h}_0^2(n)}{2\sigma^2}}\right) \qquad (9)$$

Experimental results

	Signal	Complex cepstrum
GN		
PR		
Conv.		

Waveform of music and its cepstrum

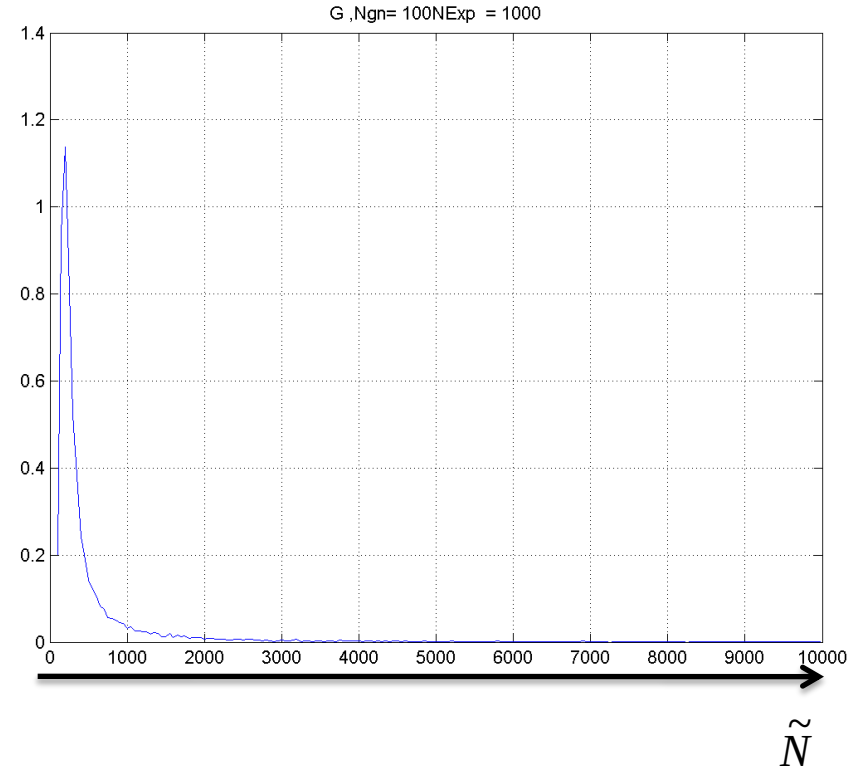


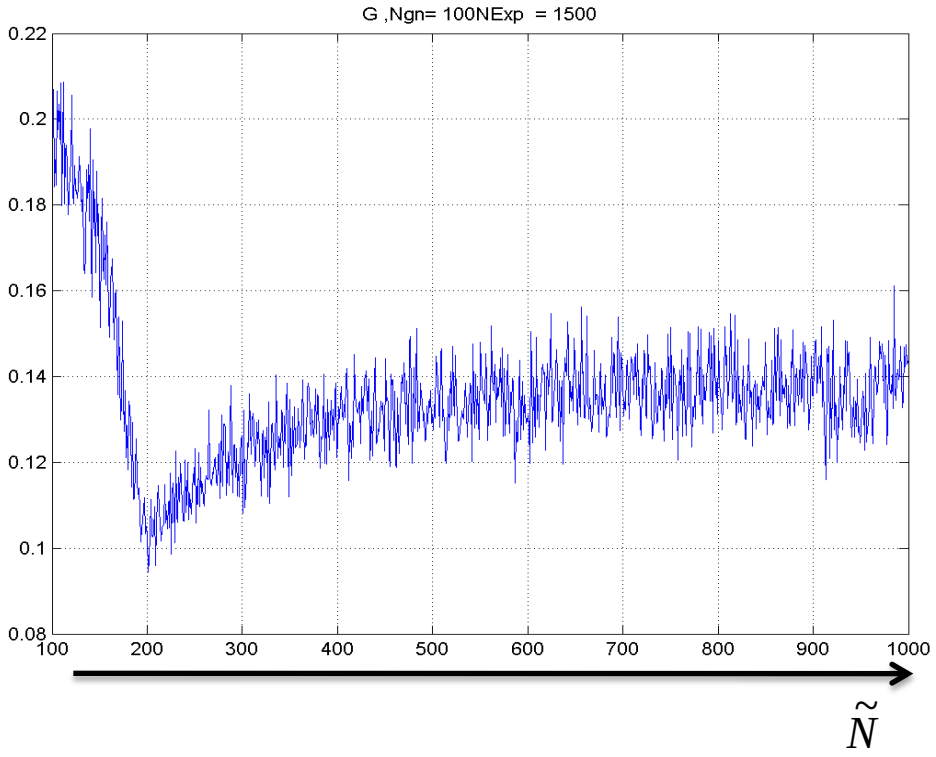
Whether equality given below holds?

$$\tilde{x}(n) \stackrel{?}{=} \tilde{S}(n) + \tilde{h}_b(n)$$

Let us denote by

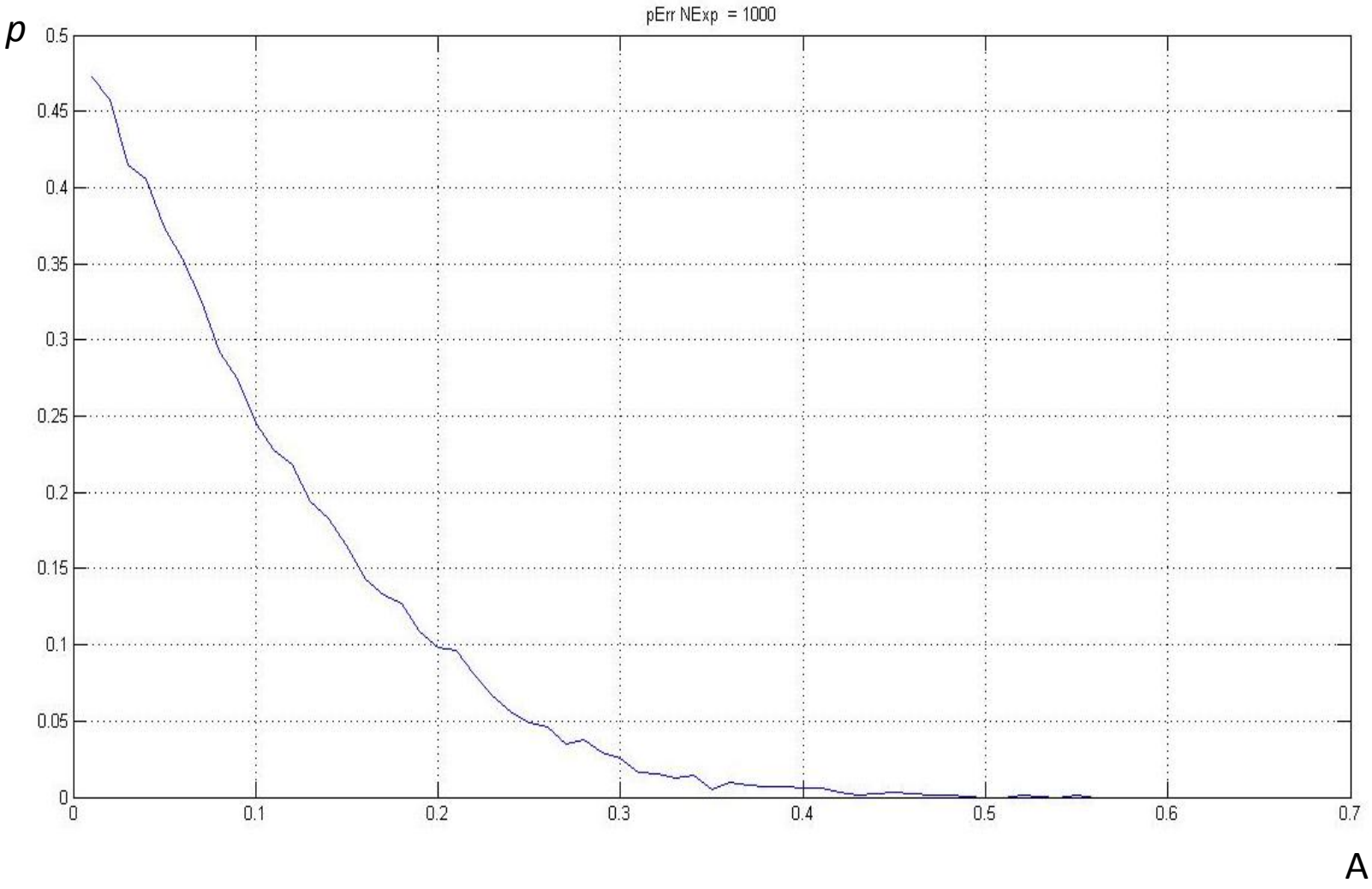
$$\delta = \frac{\sum_n (\tilde{x}(n) - (\tilde{S}(n) + \tilde{h}_b(n)))^2}{\sum_n (\tilde{S}(n))^2}$$





$\tilde{N}$ - The length of Gaussian noise appended to zeros

The probability of errors *p* versus echo amplitude *A*



The number of subintervals	The probability of errors computed by simulation			
	Rectangular window	Exponential window	Hamming window	Hanna window
1	0.38200	0.29775	0.39825	0.37950
2	0.28175	0.14750	0.30425	0.28425
4	0.13925	0.05975	0.16600	0.13250
5	0.09350	0.04575	0.11575	0.08975
6	0.05975	0.04525	0.08675	0.06225
8	0.03100	0.04925	0.06050	0.03225
9	0.02675	0.05725	0.05350	0.02700
11	0.01650	0.06700	0.04550	0.01500
13	0.01225	0.08500	0.05275	0.01175
15	0.01375	0.10025	0.07450	0.01325
16	0.01275	0.11075	0.08450	0.01375
17	0.01225	0.12500	0.11175	0.01225
18	0.01375	0.13000	0.11775	0.01500
19	0.01875	0.15075	0.16125	0.01650
22	0.02525	0.18850	0.25950	0.02225
25	0.03225	0.23575	0.37600	0.04000
32	0.12450	0.35550	0.49625	0.13600
40	0.48750	0.50375	0.49825	0.49175

Remark. The length of one-bit interval is 4000 samples

Conclusion on WM systems for audio signals based on echo hiding

- Qualitatively , the results of simulation coincide with the theory , but quantitatively there are significant differences which can be explain by the fact that “The additive property of cepstrum for convolution”, does not hold correctly on limited intervals.

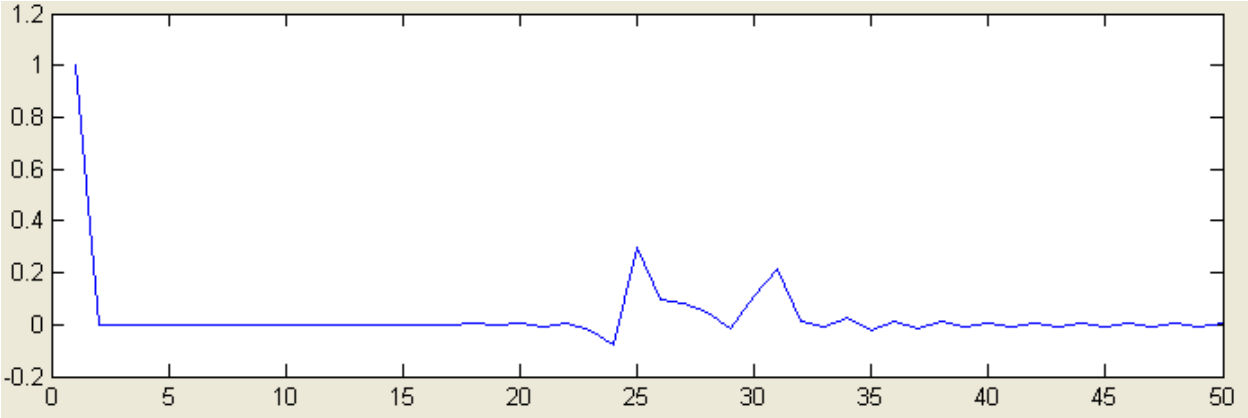
- Receiving on subintervals results in a significant decreasing of the probability of errors.

- Changing of Gaussian signal to real audio (musical) signals results in sever degradation that can be explained by the existence of strong correlation peaks in audio cepstrum.

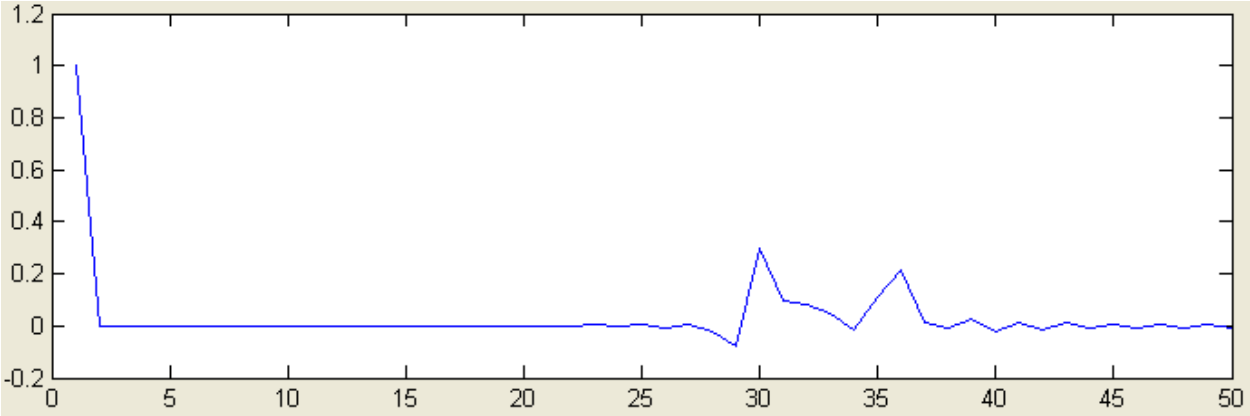
“Error free extraction method based on the use of wet paper codes(WPC)”

Examples of filter pulse responses corresponding to reverberation.[49]

Filter 1.



Filter 2.



Amplitude of reverberation	The number of subintervals	Total number of embedded bits	Errors
1	1	144	1
0,5	1	144	1
0,3	1	144	4
0,15	1	144	20
0,15	2	144	16
0,15	10	144	5
0,15	15	144	3
0,1	15	144	11
0,05	15	144	35

Conclusion on WM systems for audio signals based on reverberation [49].

1.The use of “non-trivial” window allows to decrease the amplitude of pulse response given the probability of errors and a consequence to improve the original signal quality after embedding.. The best results gives “Hanna Window”

2.The use of multi-interval receiver decreases the probability of errors.

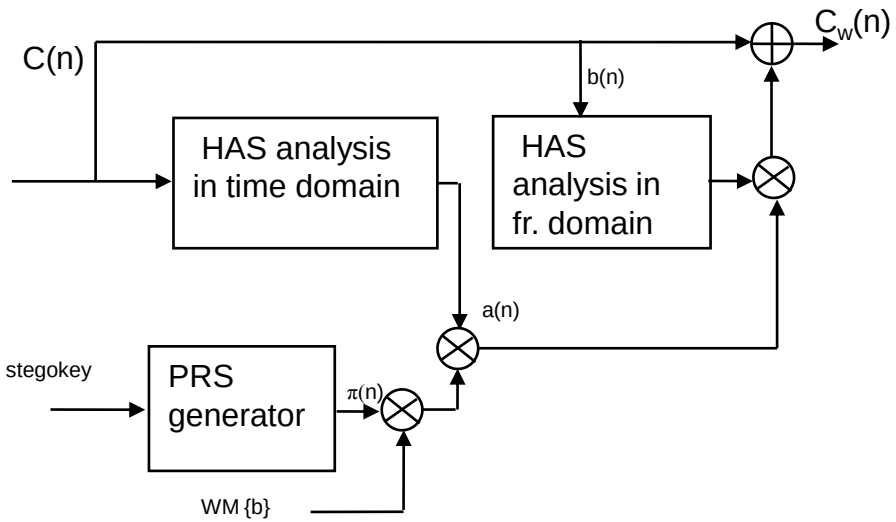
3.The choice of filter pulse responses simulating reverberation should satisfy to the number of conditions :
- small “ruining” of the original sounds (music) ,
-impossibility of pulse response guessing (like stego-key),
-small correlation of pulse response cepstrum and audio signal cepstrums,
-resistance to de-reverberation attack,
-acceptable time of signal processing in embedding procedure,
-acceptable message embedding rate.

4.Design of audio WM based on reverberation requires further investigations.

6. SS-based WM

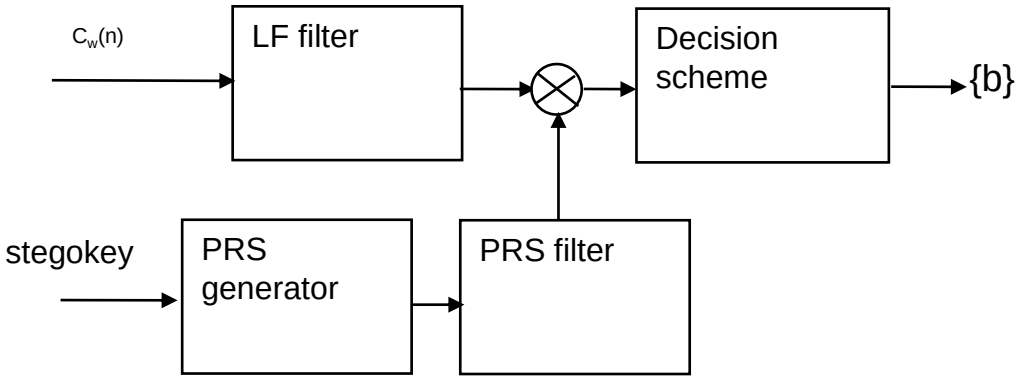
One can use all methods of embedding and extraction considered before in Lecture 9 but taking into account HAS :

a) WM embedding:



a(n) и b(n) – amplification coefficients, that have to be obtained with the use of time and frequency analysis in line with HAS.

b) WM extraction:



Low frequency filter is used in order to reduce interference from CO.

PRS filter results in such PRS that occurs the same as after filtering of WM-ed signal.

Remark 1. It is commonly to use blind decoder for audio WM .

Remark 2. In order to increase the embedding date rate can be used orthogonal PRS or error correction codes with soft decoding algorithm.

Improvement of WM scheme for audio signals [37].

1. Adaptive embedding based on possible attack analysis for each block of samples (attack characterization).

2. The use of «WM signal diversity» (similar to those methods which are used in communication technique) taking into account that mp3 transform can be considered as *selective fading* of WM carrier .

3. The use of adaptive ISS or QPD in line with HAS and possible attacks.

4.The use of error correcting codes with soft decoding algorithm (turbo codes, LPDC-codes).

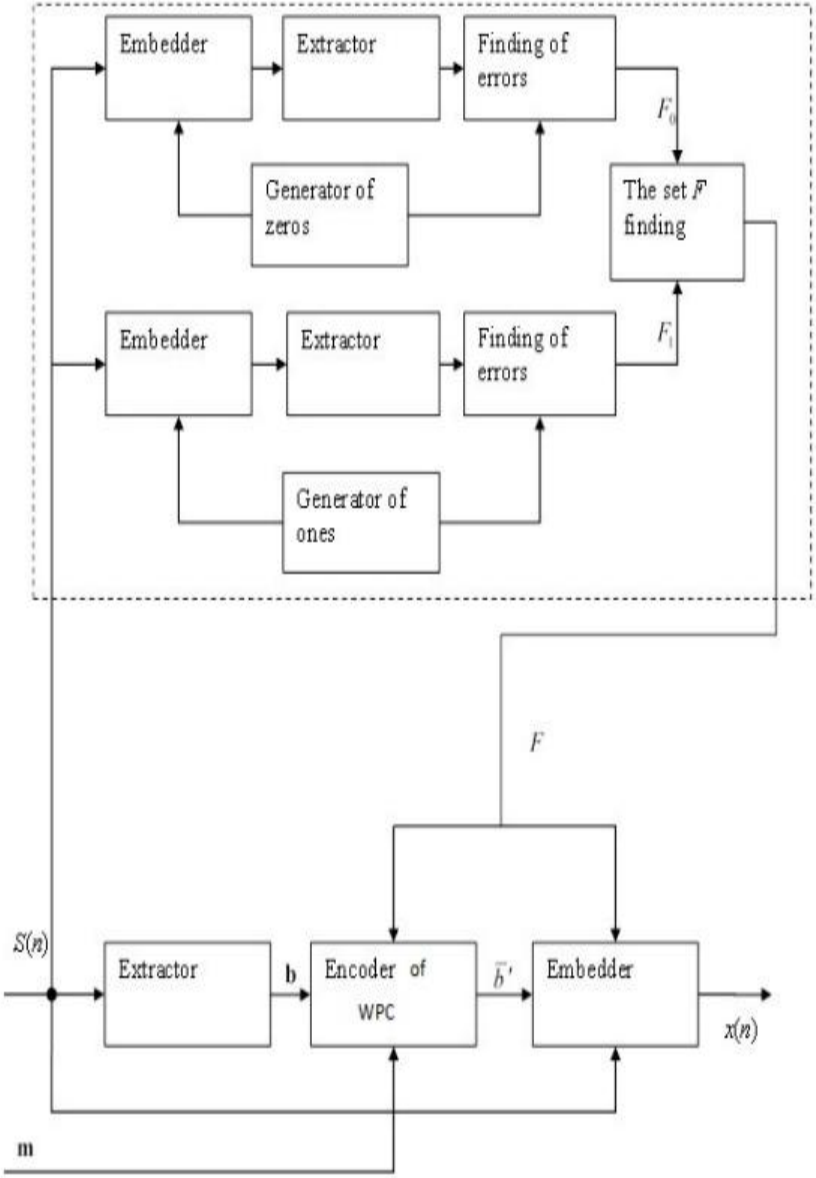
Application of wet paper codes(WPC) for a decreasing of errors within WM extraction.

In order to decrease the bit error probability after extraction(moreover, even to provide error free extraction) has been proposed to use WPC [69].

WPC were described in Lecture 5 jointly with PQS stegosystem. The general idea is to embed bits only in such N-blocks where the interference does not result in error after extraction.

The embedding algorithm is presented in the next slide:

- N_0-blocks in which extractors procedure during an embedding of both zeros and ones are XXXXX
- let F be the positions of blocks where embedding is possible.
- the message sequence $\boldsymbol{m}$ is encoded with WPC and the echo embedding is performed for those *N*-blocks (belonging to the set F) where the errors are absent
- the extractor outputs at the input of embedder the bit string $\mathbf{b}$ decoded from audio file before embedding
- The sequence $\mathbf{b}$ is embedded in the file with the rule of WPC



The embedding process with the use of WPC

It is easy to see that such embedding procedure results in error free extraction.

The sacrifice within this approach is a decreasing of embedding rate because some *N*-blocks are removed from the embedding process.

Simulation results for this approach given different music files are shown in Table below:

Name of file	t	d	T	
			$\tilde{N} = 500$	$\tilde{N} = 1000$
music1.wav	10000	9989	9563	9781
music2.wav	12179	12156	11490	11727
music3.wav	12244	12185	11430	11685
music4.wav	8135	8093	7613	7786
music5.wav	9625	9512	8990	8994

In this table are used the following notations:

- *T* - the total number of the embedded bits which were embedded in given musical file,
- *t* – is the potential number of embedded bits,
- *d* – is the number of changeable bits in audio file.