ASRSV – curs 6

Analiza Semnalului Vocal -

Analiza perceptuala

http://www.animations.physics.unsw.edu.au/jw/dB.htm

http://mirlab.org/jang/books/audioSignalProcessing/index.asp

http://www.sengpielaudio.com/calculator-soundlevel.htm

Paul R. Hill - Audio And Speech Processing With MATLAB-CRC Press (2018).

http://www.practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

Analiza semnalului vocal

Domeniul timp

- Average and maximum amplitude
- Amplitude density
- Average energy
- TEAGER means energy
- Zero crossing rate
- Fundamental frequency (F0)
- TESPAR coding

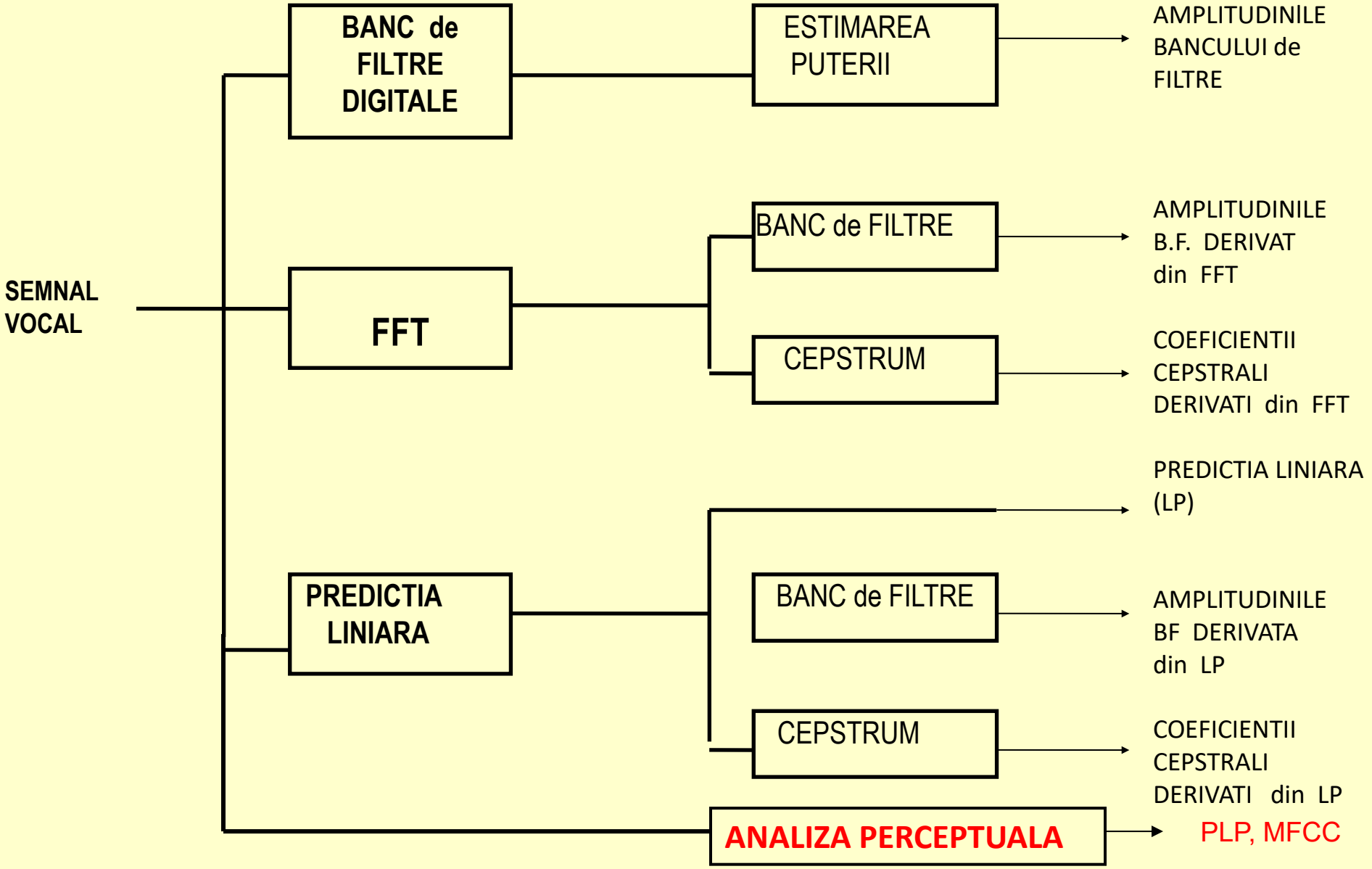
Domeniul frecventa

- DFT (FFT)
- LPC analysis
- Digital filter bank
- Cepstral analysis
- Perceptual analysis

Domeniul timp-frecventa

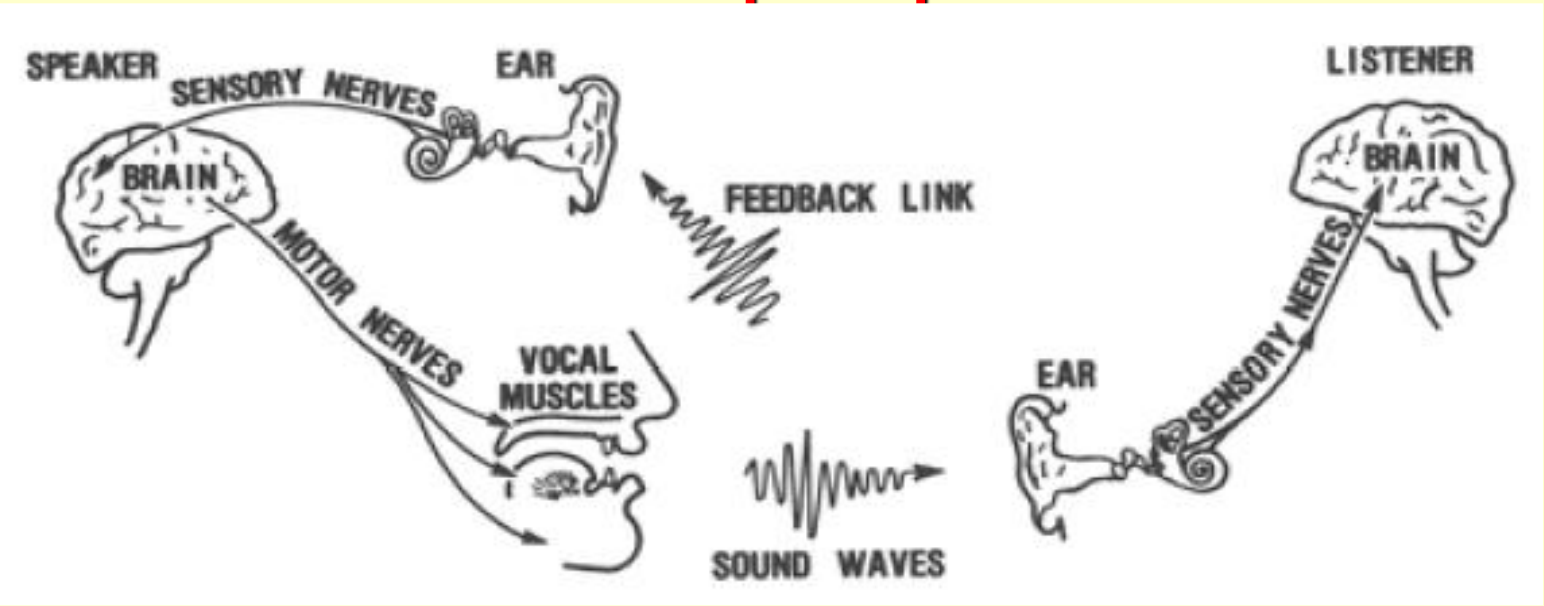
- Short-time Fourier transform (STFT)
- Discrete wavelet transform (Haar) (DWT)
- Continuous wavelet transform (Morlet) (CWT)
- Pseudo-Wigner distribution

https://www.clear.rice.edu/elec631/Projects99/mit/index2.htm



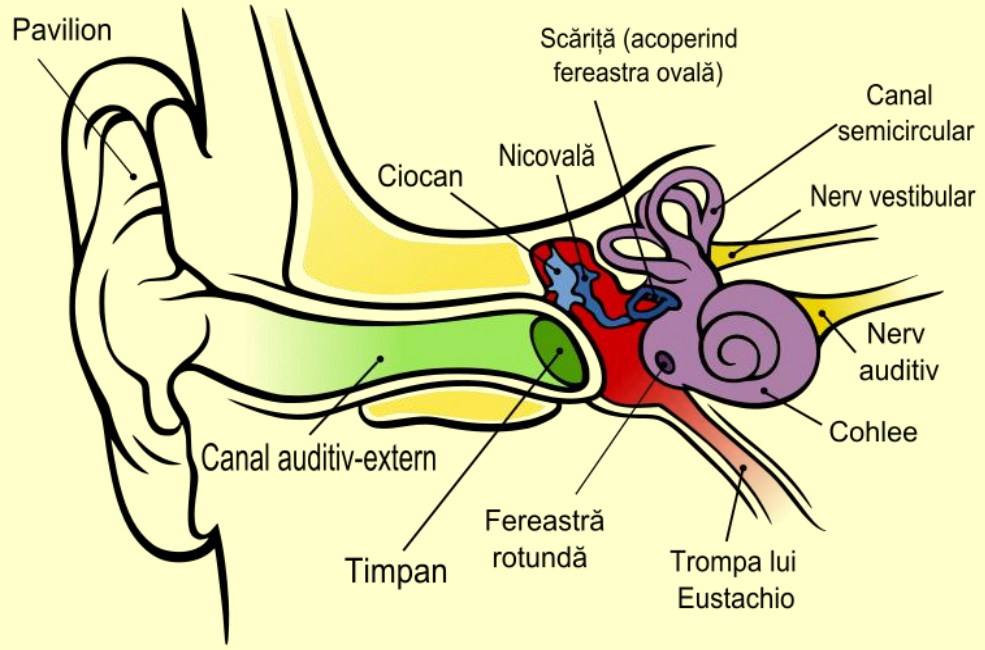
Algoritmi folosiţi în analiza spectrală [Picone]

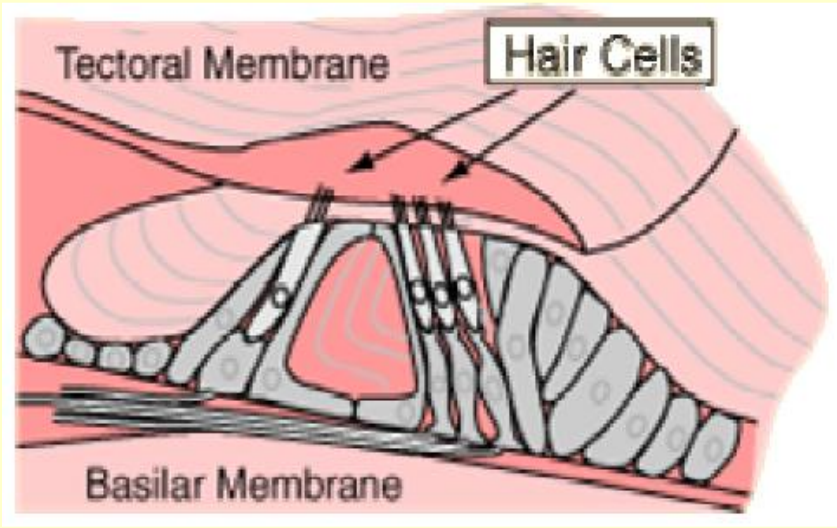
Analiza perceptuală

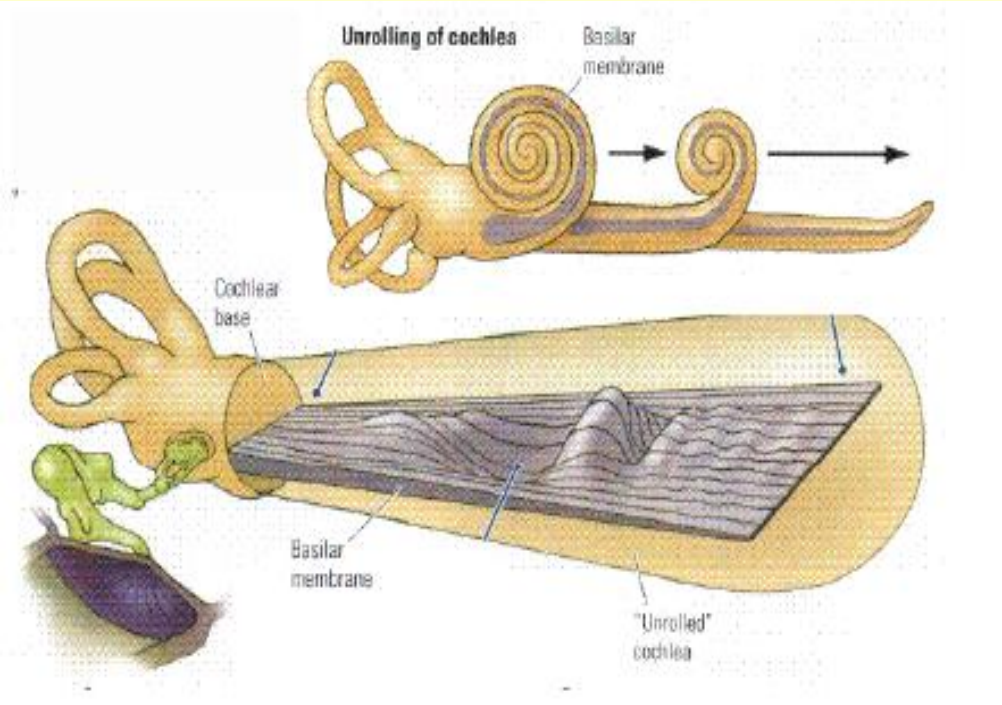


<u>Mesajul</u> pentru a fi transmis prin voce trece prin cinci niveluri de reprezentare între vorbitor şi ascultător, şi anume:

- **Nivel lingvistic** (în care sunetele de bază ale comunicării sunt alese pentru a exprima anumite gânduri)
- **Nivel fiziologic** (componentele tractului vocal produc sunete asociate cu unităţile lingvistice ale rostirii)
- **Nivelul acustic** (sunetul eliberat din buze şi de nări este transmis atât la vorbitor (feedback-ul de sunet) cat si la ascultător
- **Nivel fiziologic** (sunetul este analizat de către ureche şi nervii auditivi)
- **Nivel lingvistic** (în cazul în care vocea este perceputa ca o secvenţă de unităţi lingvistice şi înţeles în termen de idei comunicate)

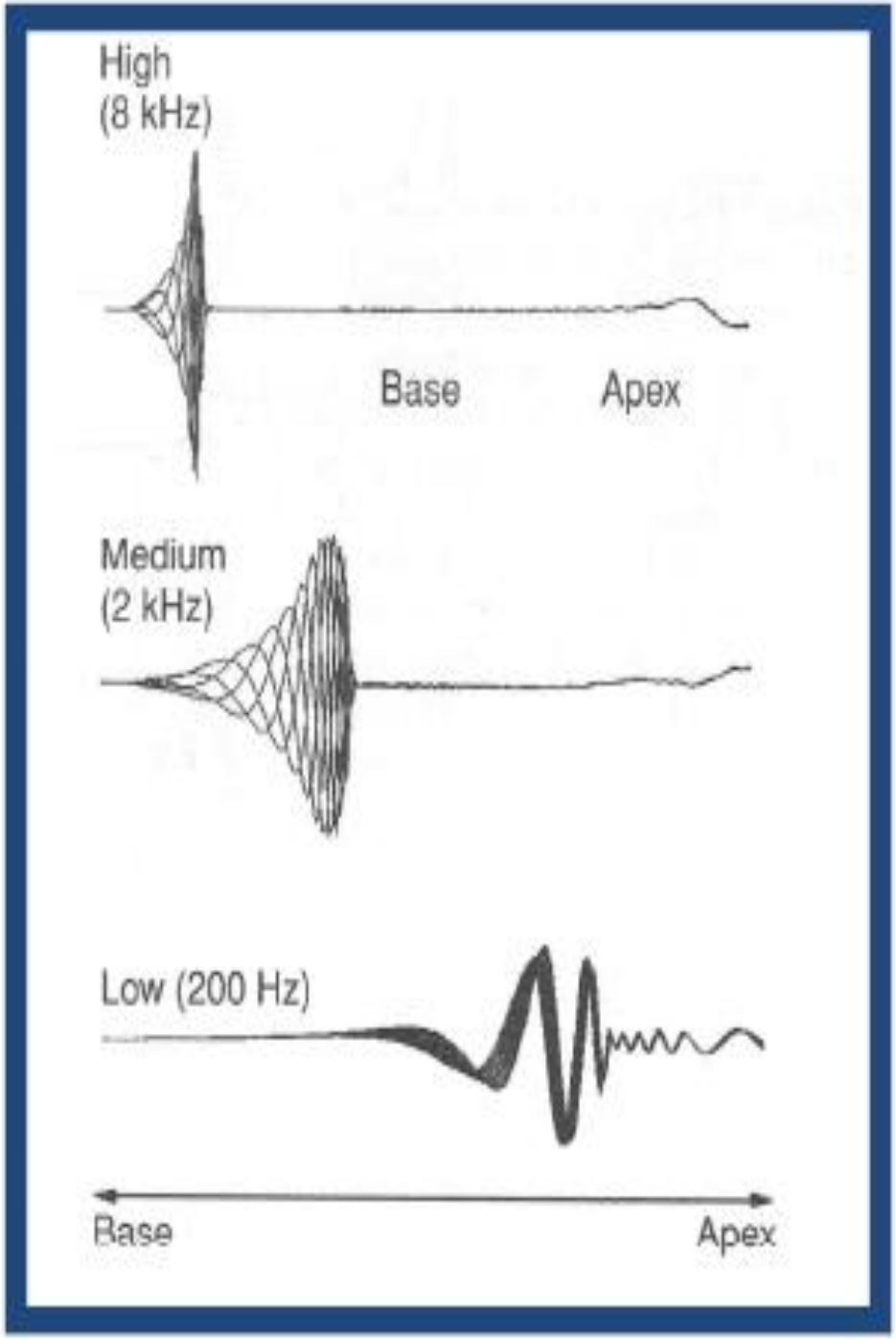


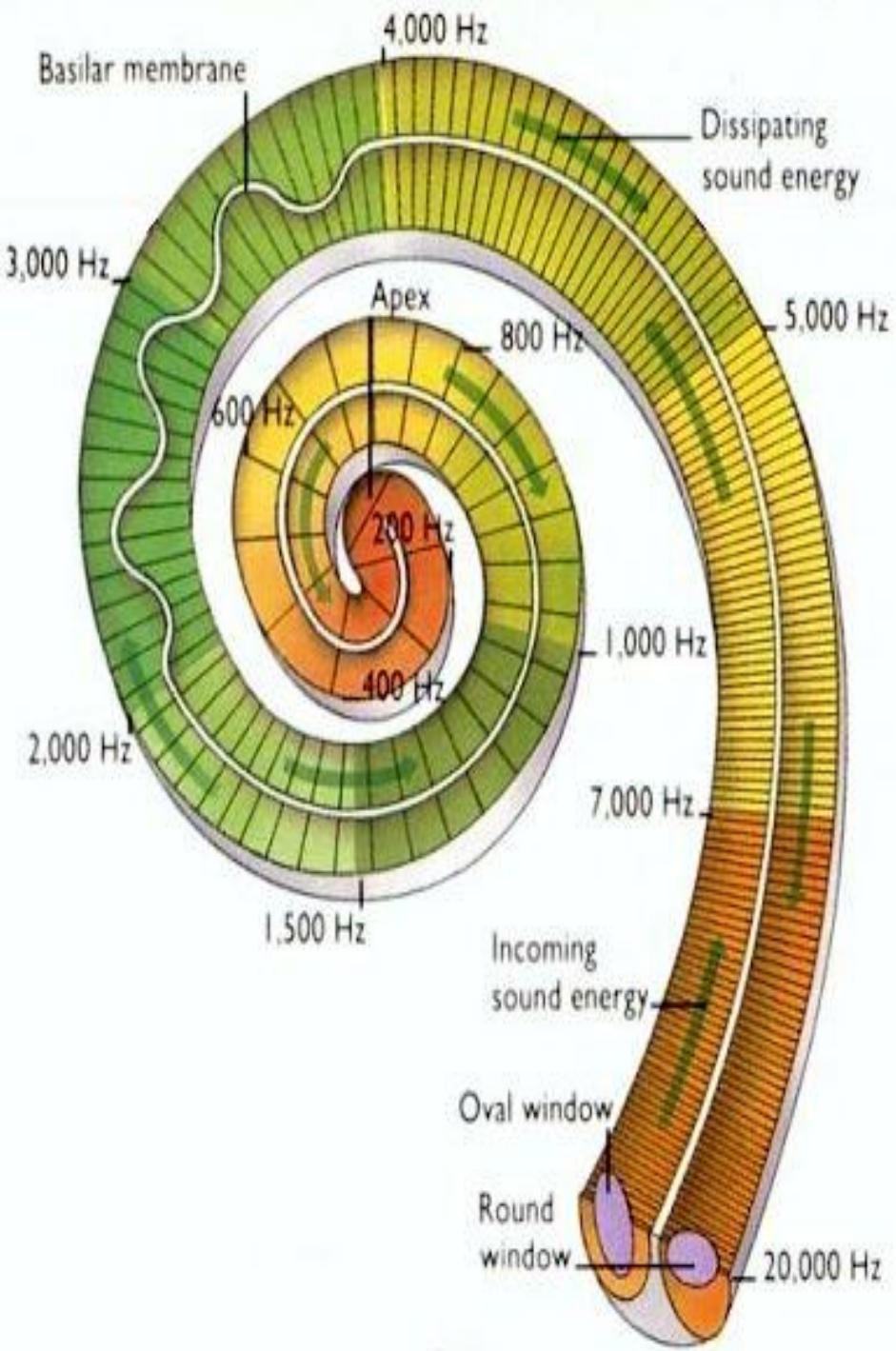




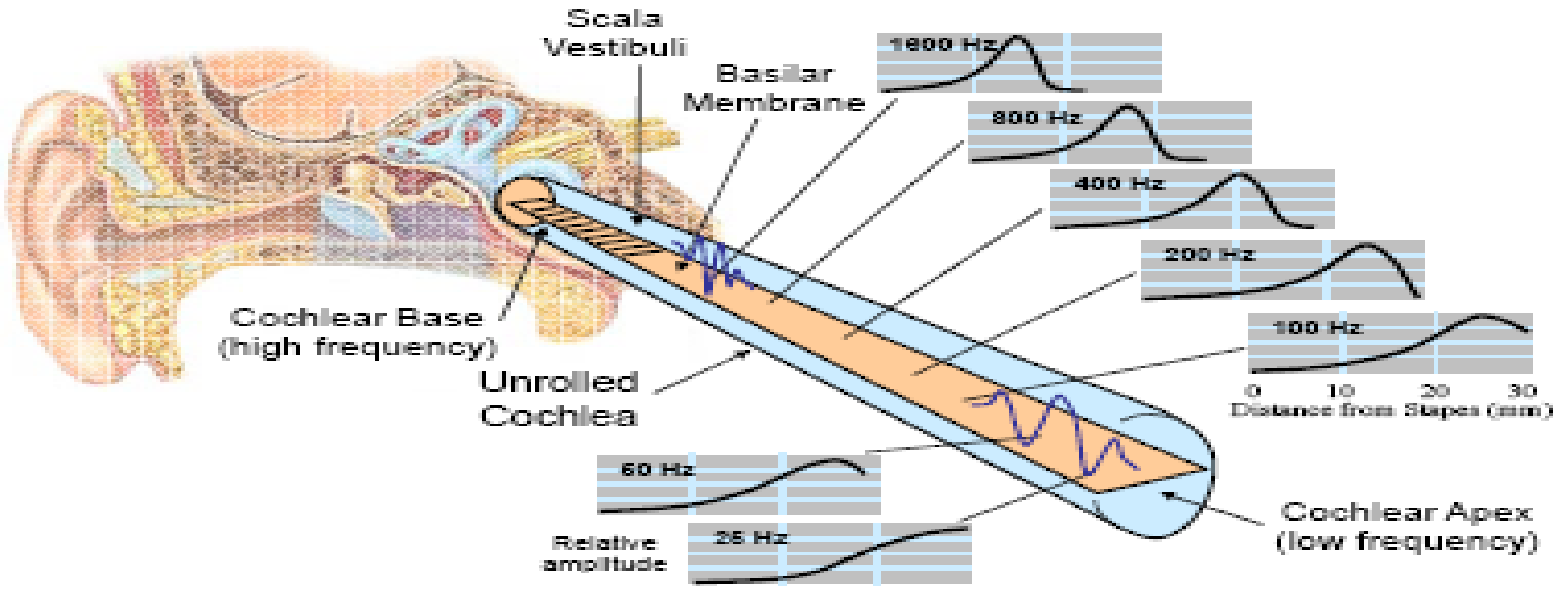
De ce avem 2 urechi?

- **Localizare Sunet** – localizeaza surse de sunet în 3D
- **Anulare sunet** - atenţia se poate concentra pe o sursa de sunet dintr-o serie de surse de sunet “efect coctail party"
- **Efectul de ascultare peste căşti** => localizare sunete din interiorul capului (mai degrabă decât spaţial în afara capului)


High
(8 kHz)
Base
Apex
Medium
(2 kHz)
Low (200 Hz)
Base
Apex


4,000 Hz
Basilar membrane
Dissipating
sound energy
3,000 Hz
5,000 Hz
Apex
800 Hz
600 Hz
200 Hz
400 Hz
1,000 Hz
2,000 Hz
7,000 Hz
1,500 Hz
Incoming
sound energy
Oval window
Round
window
20,000 Hz

Stretched Cochlea & Basilar Membrane



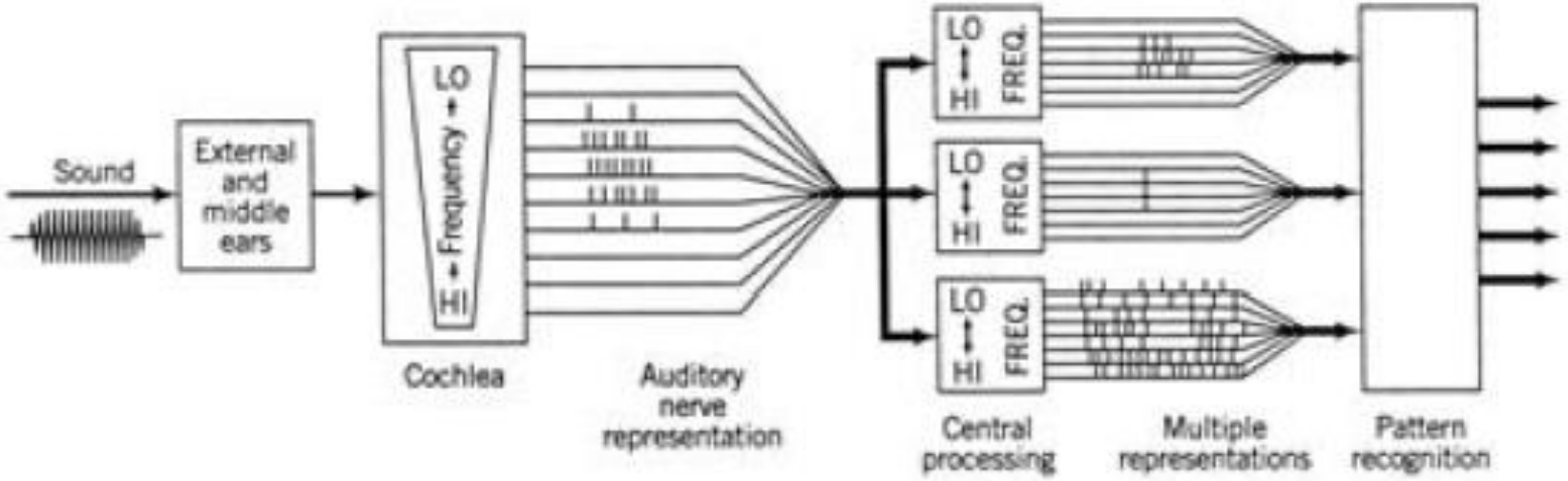
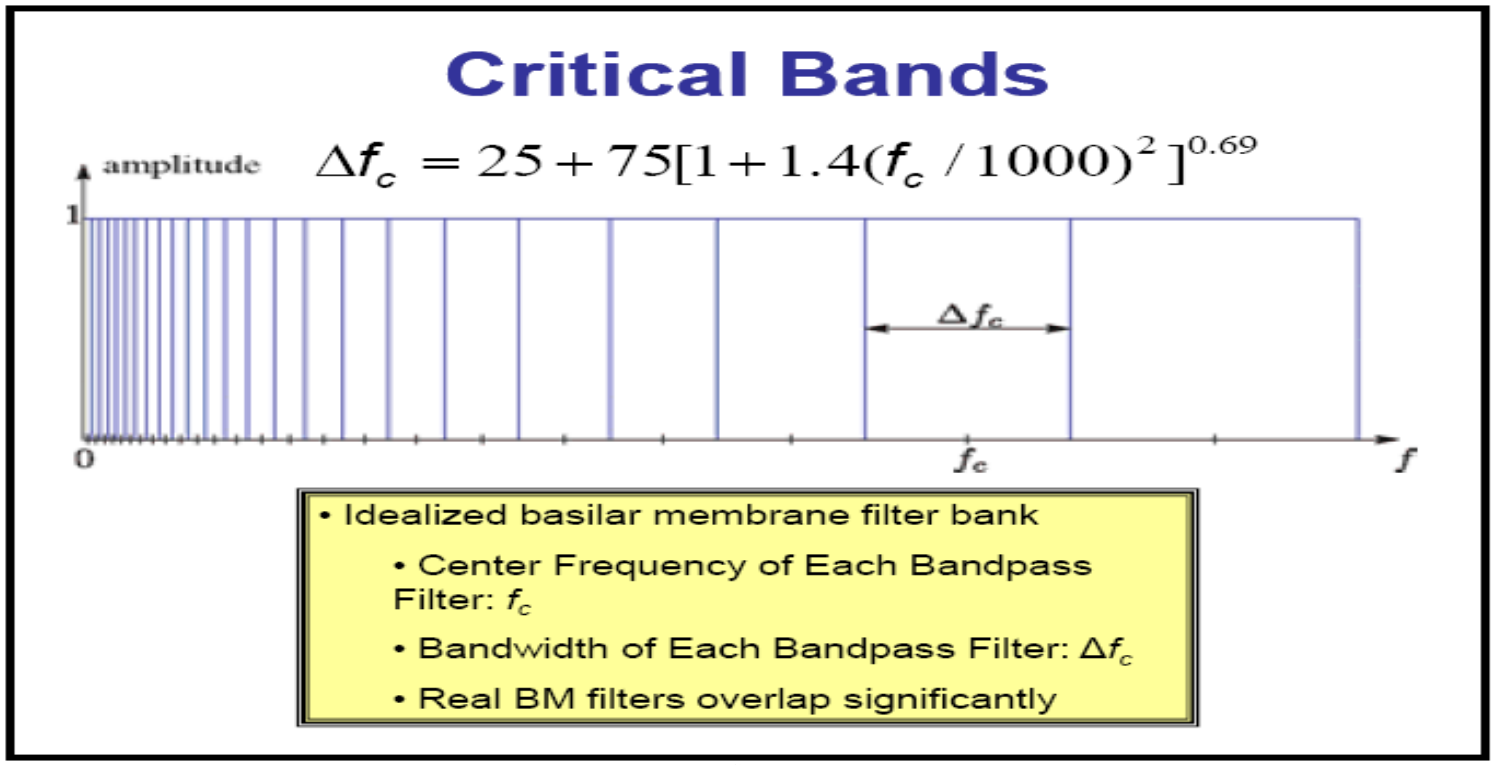
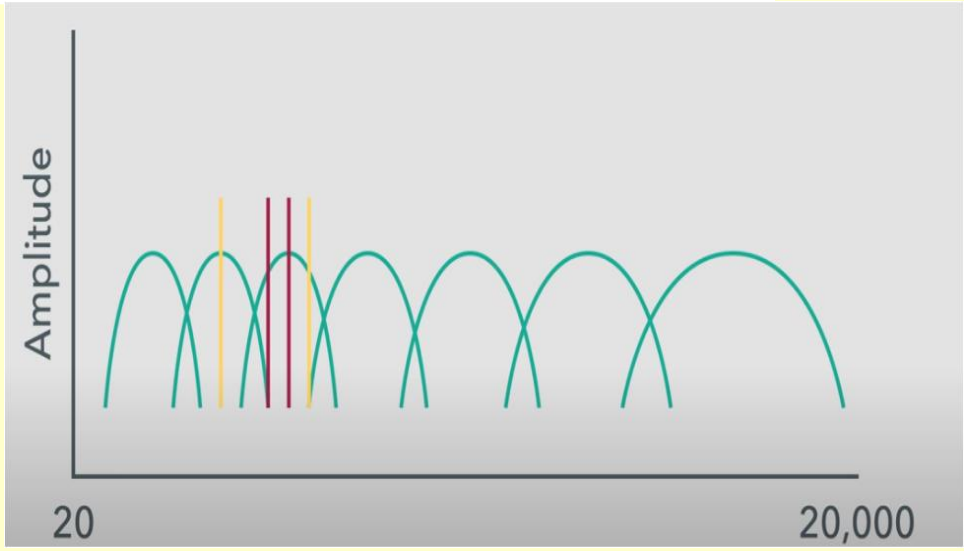


FIGURE 17.1 Block diagram of sound representation in the auditory system. From [16].



- Membrana bazilara => este realizarea mecanică a unui banc de filtre
  - filtrele sunt aproximativ constante Q (frecvenţa centrala fc/ lăţime de bandă Δfc), cu lăţime de bandă crescătoare logaritmic

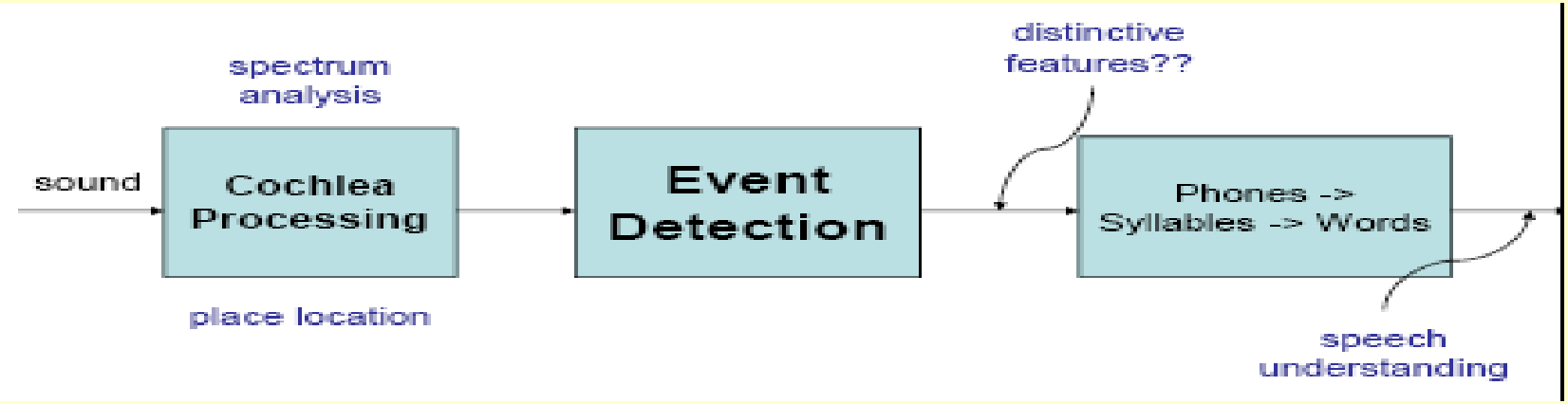


•**Banda critică** este banda frecvențelor audio în cadrul căreia un al doilea ton va interfera cu percepția primului ton prin mascarea auditivă

Ce rezulta din modelele auditive (perceptuale)?

– Analiza segmentelor de vorbire atât pe *timp scurt* (20 ms ptr. foneme) cat si pe *termen lung* (200 ms pentru silabe); este de mare importantă :

- structura temporală a vorbirii
- structura spectrala a sunetelor (formanti)
- caracteristicile dinamice (delta) - evolutia

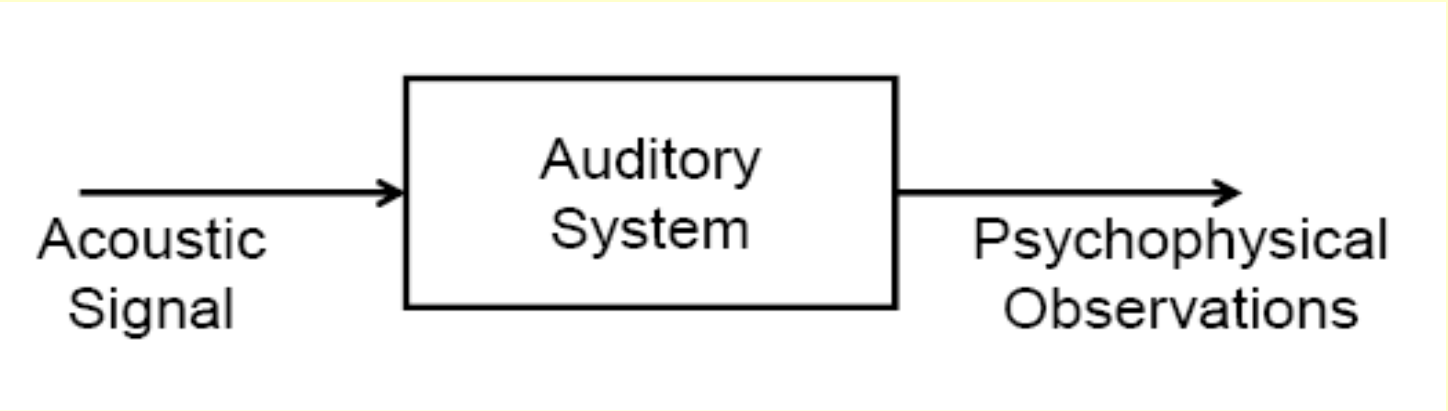


Potrivirea cu perceptia umana a vorbirii

- scară neliniară de frecvenţă - Mel, Bark
- compresia amplitudinii spectrale (gama dinamica) - taria (compresie log)
- curbele de egala tarie – scaderea sensibilitatii la frecvenţele joase
- integrarea spectrala pe termen lung - caracteristici “temporale"
- mascarea auditiva - tonuri puternice (sau zgomot) tind sa mascheze semnale adiacente în benzile critice de frecvenţă

- *Aceste efecte sunt, în general, continute direct în modelele auditive*

Modelul “BLACK-BOX” al sistemului auditiv



Atribut fizic (cantitativ masurabil)	Observatie Psihoacustica (cantitate perceputa)
Intensitate	Tarie
Inaltime	F0 (pitch)

Aspecte interesante ale perceptiei:

- Gama sunetelor audibile este: ~ 20Hz - 20kHz
- Urechea nu este sensibila in mod egal la toate frecventele
- Taria perceputa este functie de frecventa si amplitudinea sunetului

- Corespondentele intre marimi sunt complexe si departe de a fi liniare

Modele auditive propuse

- Perceptual Linear Prediction (PLP) – H. Hermansky
- Mean-Synchrony Auditory Model – S. Seneff
- Cochlea Model – D. Lyon
- Ensemble Interval Histogram – O. Ghitza
- MFCC

http://www.sengpielaudio.com/Calculations03.htm

- **Sound Studio and Audio Calculations Audio and Acoustics Conversions 1**
  Assistance and explanations of special issues around the recording technology

Intensitatea sunetului

- Intensitatea unui sunet este o marime fizică care poate fi măsurata şi cuantificată
- **intensitatea acustica (I),** definita puterea medie măsurată printr-o unitate de suprafata [W/m^2]
- Gama de intensităţi între 10^{-12} W/m^2 - 10 W/m^2, corespunde intervalului de la pragul de audibilitate la pragul de durere
- pragul de audibilitate definit la $I_0 = 10^{-12}$ W/m^2, iar **nivelul de intensitate (IL)** al unui sunet (IL):

$$IL = 10 \log_{10} \left(\frac{I}{I_0} \right) \quad \text{dB}$$

- Pentru un ton pur, de amplitudine I, $I \sim P^2$, iar **nivelul de presiune** al sunetului (SPL): unde $P_0 = 2*10^{-5}$ N/m^2

$I = p^2 / Z_0$

$$SPL = 10 \log_{10} \left(\frac{P^2}{P_0^2} \right) = 20 \log_{10} \left(\frac{P}{P_0} \right) \text{dB}$$

Prag de audibilitate = nivelul intensitatii acustice (I_0) a unui sunet pur, care poate fi abia auzit la o anumita frecvenţă

- **Nivelul pragului de audibilitate (IL_0)** ≈ 0 dB la 1000 Hz

-**prag de senzaţie** ≈ 120 dB

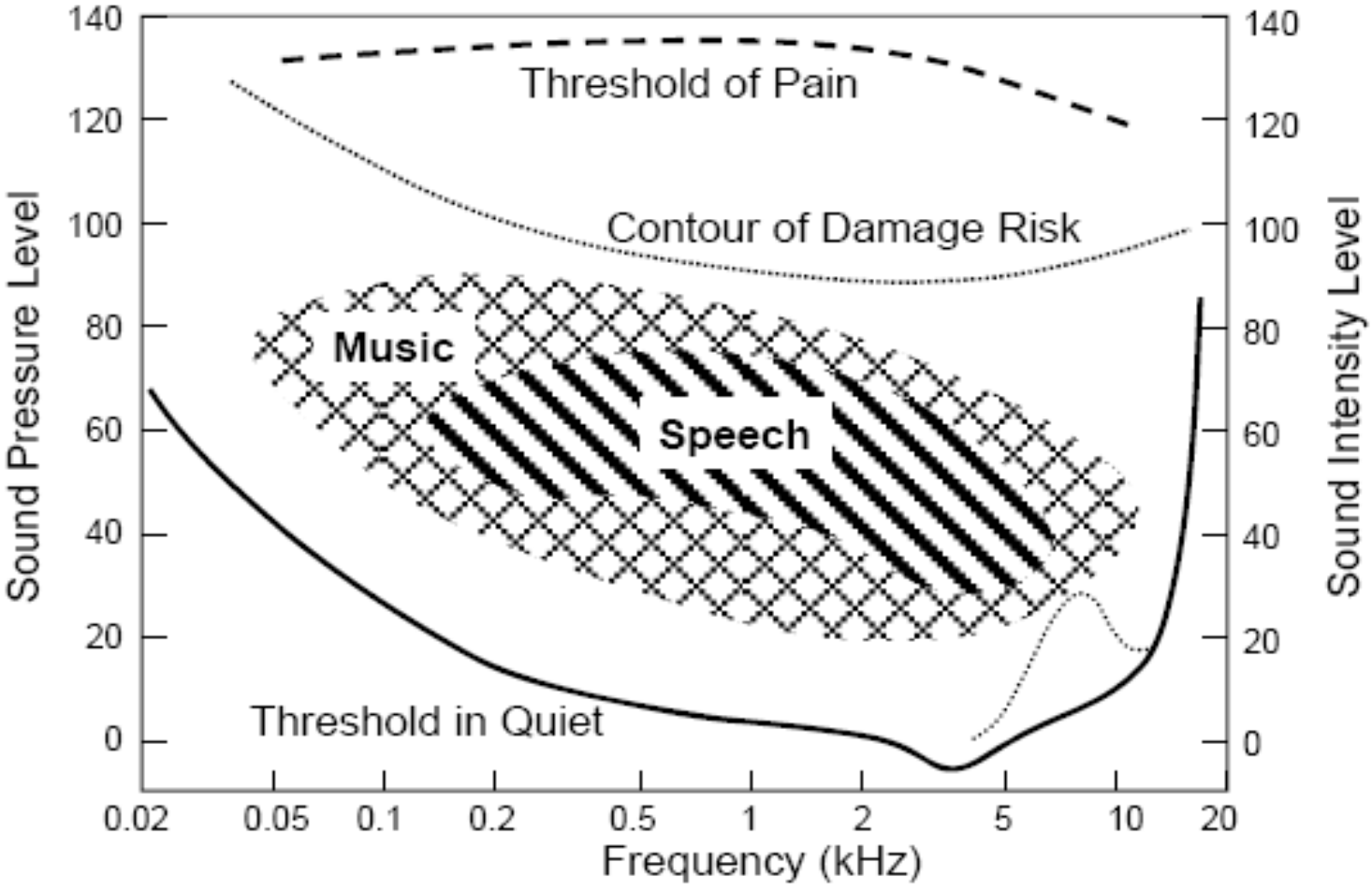
-**Prag de durere** ≈ 140 dB

-**Daune imediate** ≈ 160 dB

- Pragurile - variaza în funcţie de frecvenţă şi de la persoană la persoană
- sensibilitatea maximă este la ~ 3 kHz

Sursa de sunet	Puterea sunetului P_{ac} watts	Nivelul de putere al sunetului L_w dB re 10^{-12} W
Racheta	**1,000,000 W**	180 dB
Motor turbojet avioane	**10,000 W**	160 dB
Sirena	**1,000 W**	150 dB
Concert rock, camioane	**100 W**	140 dB
Mitraliera	**10 W**	130 dB
Pic-hammer	**1 W**	120 dB
Trompeta, escavator	**0.3 W**	115 dB
Latratul cainelui	**0.1 W**	110 dB
Elicopter	**0.01 W**	100 dB
Voce tare, plans copil	**0.001 W**	90 dB
Vorbirea, masina de scris	**10^{-5} W**	70 dB
frigider	**10^{-7} W**	50 dB
Pragul audibil	**10^{-12} W**	0 dB

Range of Human Hearing

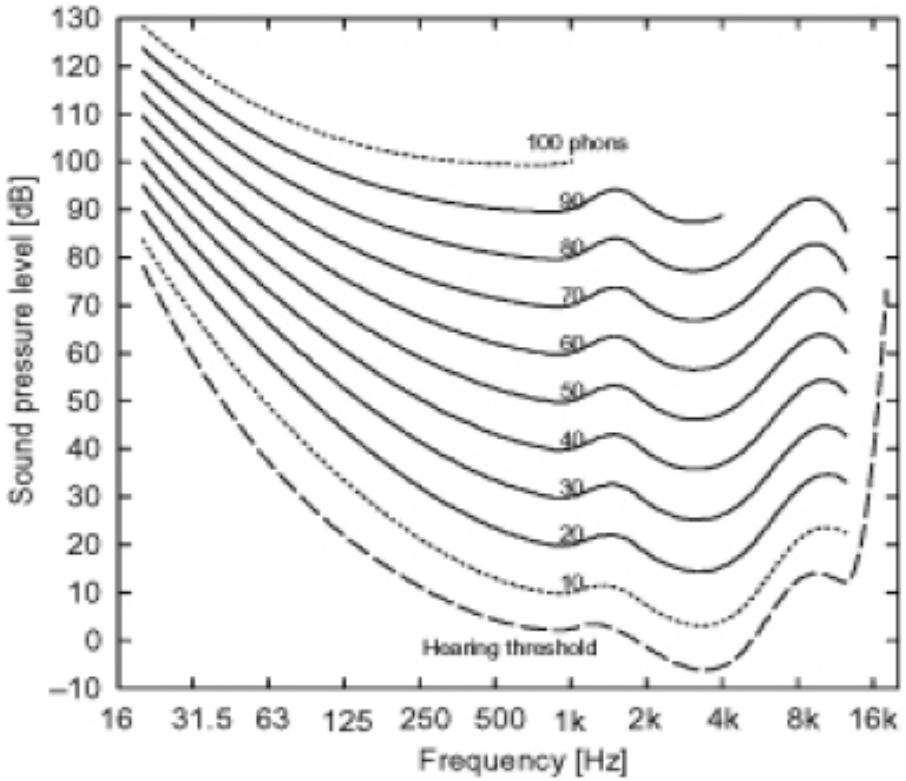


NIVELUL TARIEI ACUSTICE (LL)

- **Nivelul tariei (LL)** = cu IL al unui ton de 1KHz, care este apreciat de majoritatea observatorilor ca fiind de egala tarie cu tonul (phons)

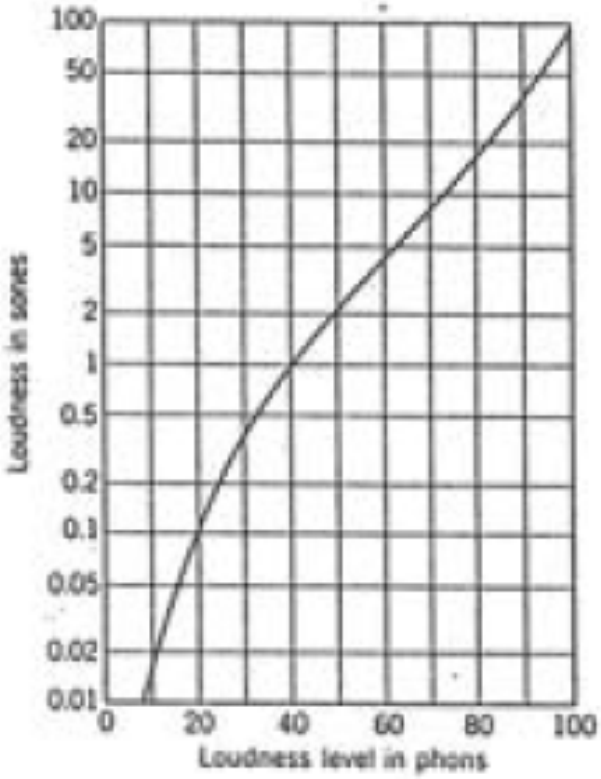
Loudness Level

- ***Loudness Level (LL)*** is equal to the *IL* of a 1000 Hz tone that is judged by the average observer to be equally loud as the tone



TARIA SUNETULUI (L)

- **Taria (sones)** este o scara care se dubleaza cand *taria perceputa* se dubleaza
- ***Loudness (L)*** (in sones) is a scale that doubles whenever the *perceived* loudness doubles



$$\log L = 0.033\,(LL - 40)$$
$$= 0.033LL - 1.32$$

• for a frequency of 1000 Hz, the loudness level, LL, in phons is, by definition, numerically equal to the intensity level IL in decibels, so that the equation may be rewritten as

$$LL = 10\log(I / I_0)$$

or since $I_0 = 10^{-12}$ watts/m^2

$$LL = 10\log I + 120$$

Substitution of this value of LL in the equation gives

$$\log L = 0.033(10\log I + 120) - 1.32$$
$$= 0.33\log I + 2.64$$

which reduces to

$$L = 445I^{0.33}$$

Perceptia frecventei - Pitch

Pitch-ul este o cantitate perceputa, in timp ce frecventa este una fizica (Hz)

- ***Pitch-ul*** si ***frecventa fundamentala(FF) sunt*** diferite !!!!

- suntem sensibili la schimbari ale pitch-ului
  - $F < 500$ Hz, $\Delta F \approx 3$ Hz
  - $F > 500$ Hz, $\Delta F \approx 0.003*F$

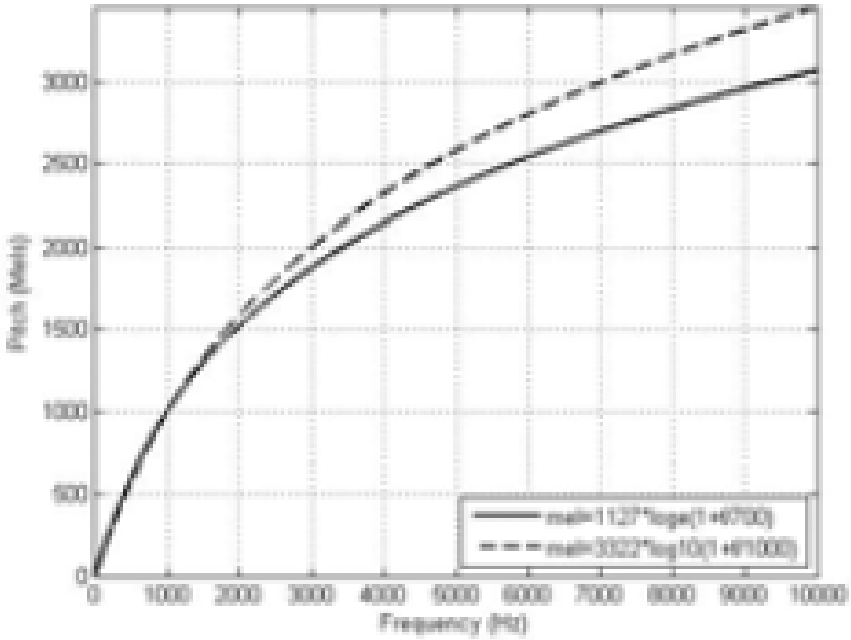
- relatia intre pitch si frecventa fundamentala nu este simpla, nici chiar ptr tonuri pure

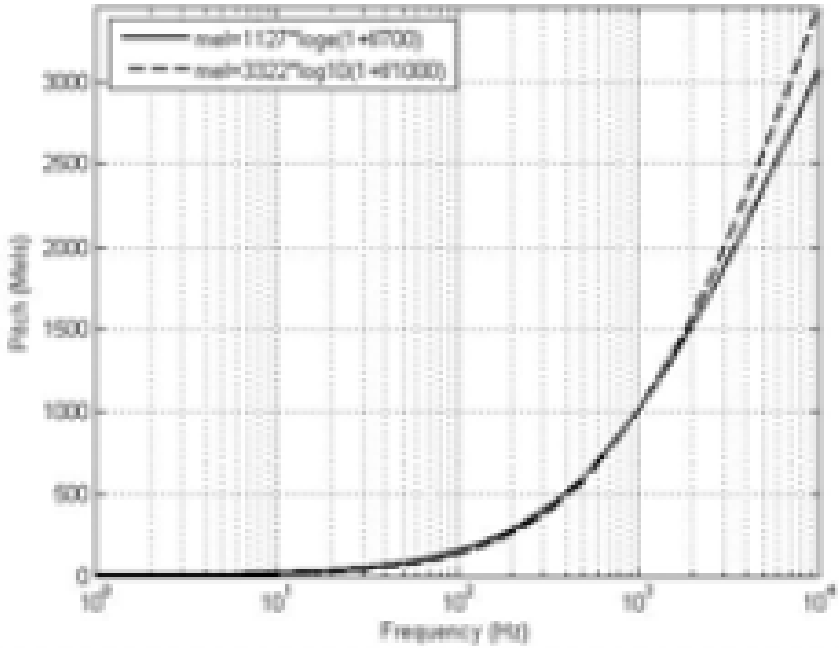
Ex. – un ton care are pitch ½ din pitch-ul unui ton de 200 Hz are FF ~ 100 Hz

– un ton care are pitch-ul ½ din pitch-ul unui ton de 5000 Hz are FF ~2000 Hz

- ***La sunete complexe – vocea – Pitch-ul*** este corelat cu **FF, (**dar nu la fel), relatia fiind mai complexa decat la tonurile pure

Pitch-The Mel Scale



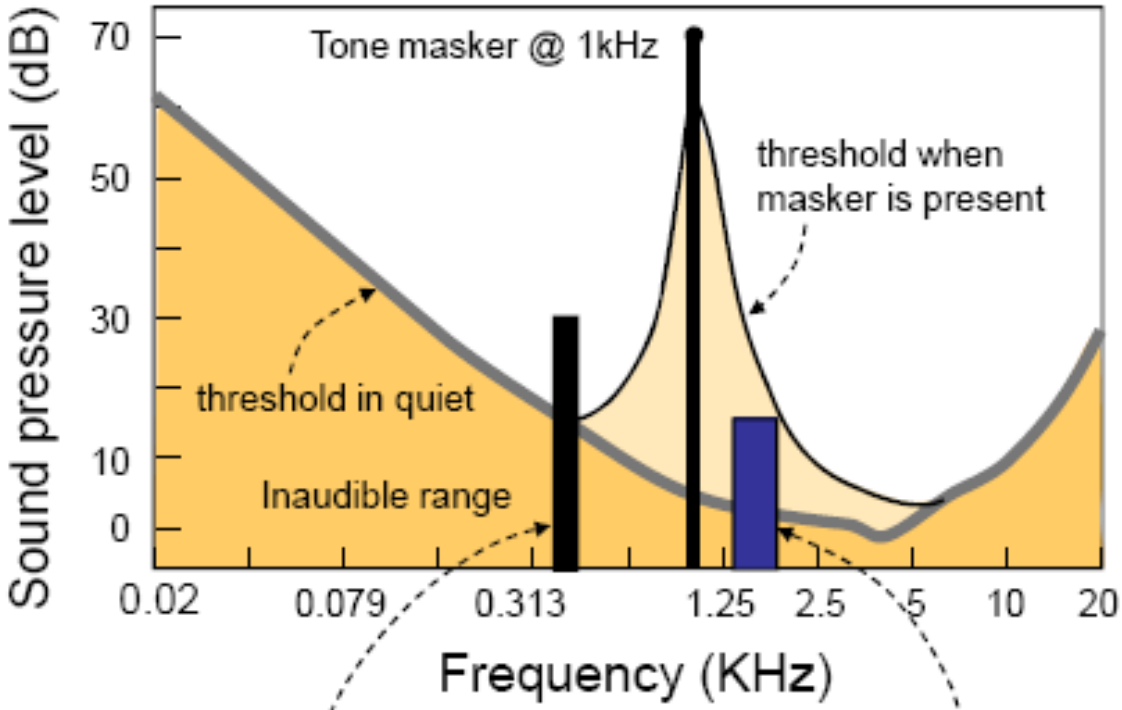


$$\text{Pitch } (mels) = 3322 \log_{10}(1 + f / 1000)$$

Alternatively, we can approximate curve as:

$$\text{Pitch } (mels) = 1127 \log_{e}(1 + f / 700)$$

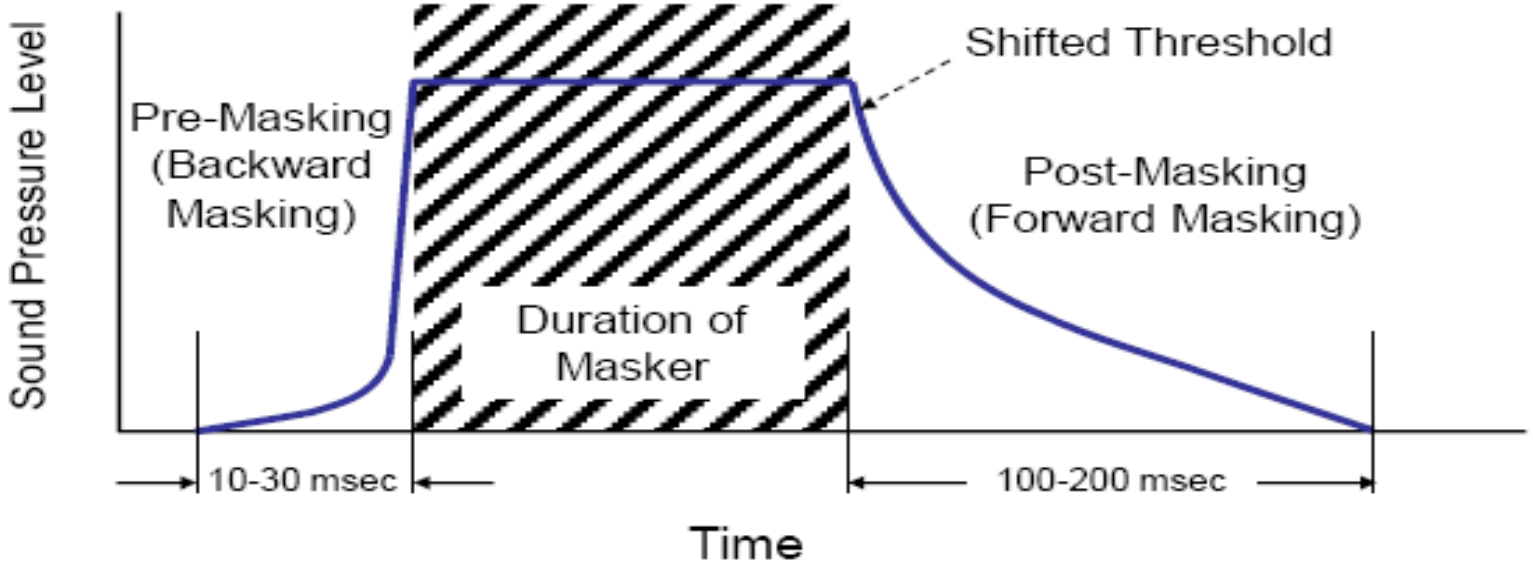
Auditory Masking



Signal perceptible even in the presence of the tone masker

Signal not perceptible due to the presence of the tone masker

Temporal Masking



• mascarea temporală - apare atunci când un stimul sonor brusc face alte sunete neauzite, care preced imediat stimulul sau după stimul

1. **Mascarea inversă (pre-mascare)**

- Ce este: Un sunet ulterior (mascatorul) poate masca un sunet anterior (ținta).
  Fereastra de timp: Acest efect este foarte scurt, de obicei până la 20 ms înainte de apariția mascherului.
- De ce se întâmplă (teorii): Teoria dominantă este că răspunsul neuronal puternic al mascarului puternic „ajunge din urmă” și copleșește răspunsul neuronal încă în curs de procesare al sunetului țintă mai slab din calea auditivă.
- În esență, acesta întrerupe procesarea de către creier a primului sunet înainte ca acesta să poată fi perceput pe deplin.
- Ex. Un clic foarte scurt și silențios urmat imediat de un zgomot puternic. Nu vei percepe clic-ul. Zgomotul puternic a „mascat” sunetul care a venit înainte de el.

- **2. Mascarea înainte (post-mascare)**
- Ce este: Un sunet anterior (mascatorul) poate masca un sunet ulterior (ținta).

Fereastra de timp: Acest efect durează mai mult, de obicei între 50 -200 de ms după terminarea mascarului. Este ca „imaginea reziduală” a unui sunet puternic, care face dificilă vederea (sau, în acest caz, auzirea) a ceea ce urmează.

- De ce se întâmplă (teorii): Celulele ciliate din cohleea și neuronii asociați devin „obosiți” sau se adaptează după ce răspund la mască. Se află într-o perioadă refractară temporară și sunt mai puțin sensibili la un sunet nou, mai slab, care urmează imediat după. Le ia timp să-și recupereze sensibilitatea.
- Ex. După un foc de armă puternic, s-ar putea să aveți probleme în a auzi șoapta cuiva pentru o fracțiune de secundă. Focul de armă maschează șoapta.

Cum utilizează MP3 mascarea temporală?

Scopul compresiei audio este de a reduce dimensiunea fișierului prin eliminarea datelor audio pe care urechea nu le percepe. Mascarea temporală este o bună oportunitate pentru a face acest lucru.

1. Codificatorul analizează semnalul audio și identifică sunetele puternice și tranzitorii (cum ar fi o lovitură de tobă sau o lovitură de cinel) care acționează ca mascare puternică.

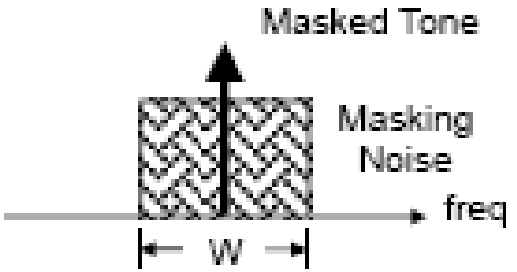
2. Calculează „pragul de mascare” pentru perioadele scurte imediat înainte și imediat după aceste evenimente puternice.

3. Orice sunete mai slabe care se încadrează sub acest prag de mascar în timpul acestor intervale de timp sunt considerate inaudibile și sunt pur și simplu eliminate din fișierul final.

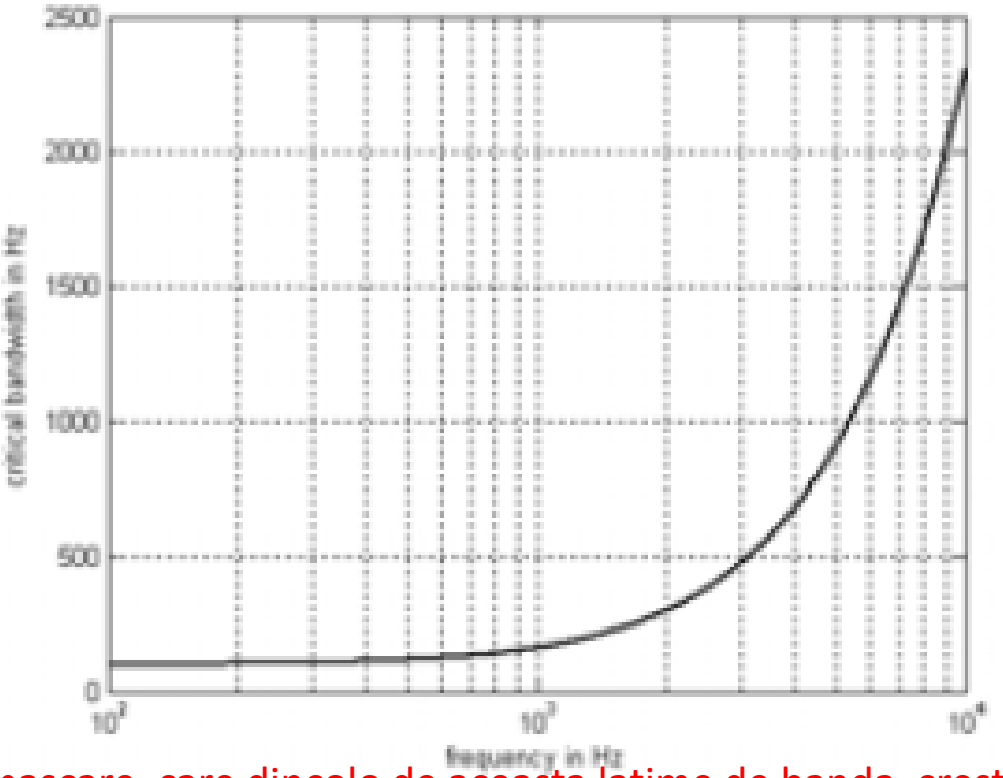
4. Rezultatul este un fișier mult mai mic, iar pentru urechea umană este in continuare foarte similar cu originalul, deoarece sunetele eliminate ar fi fost oricum „mascate”.

Masking & Critical Bandwidth

- ***Critical Bandwidth*** is the bandwidth of masking noise beyond which further increase in bandwidth has little or no effect on the amount of masking of a pure tone at the center of the band



The noise spectrum used is essentially rectangular, thus the notion of equivalent rectangular bandwidth (ERB)



Banda critica = banda unui zgomot de mascare, care dincolo de aceasta latime de banda, cresterea benzii nu are efect sau are efect redus asupra mascarii unui ton pur aflat la centrul benzii

Scari utilizate

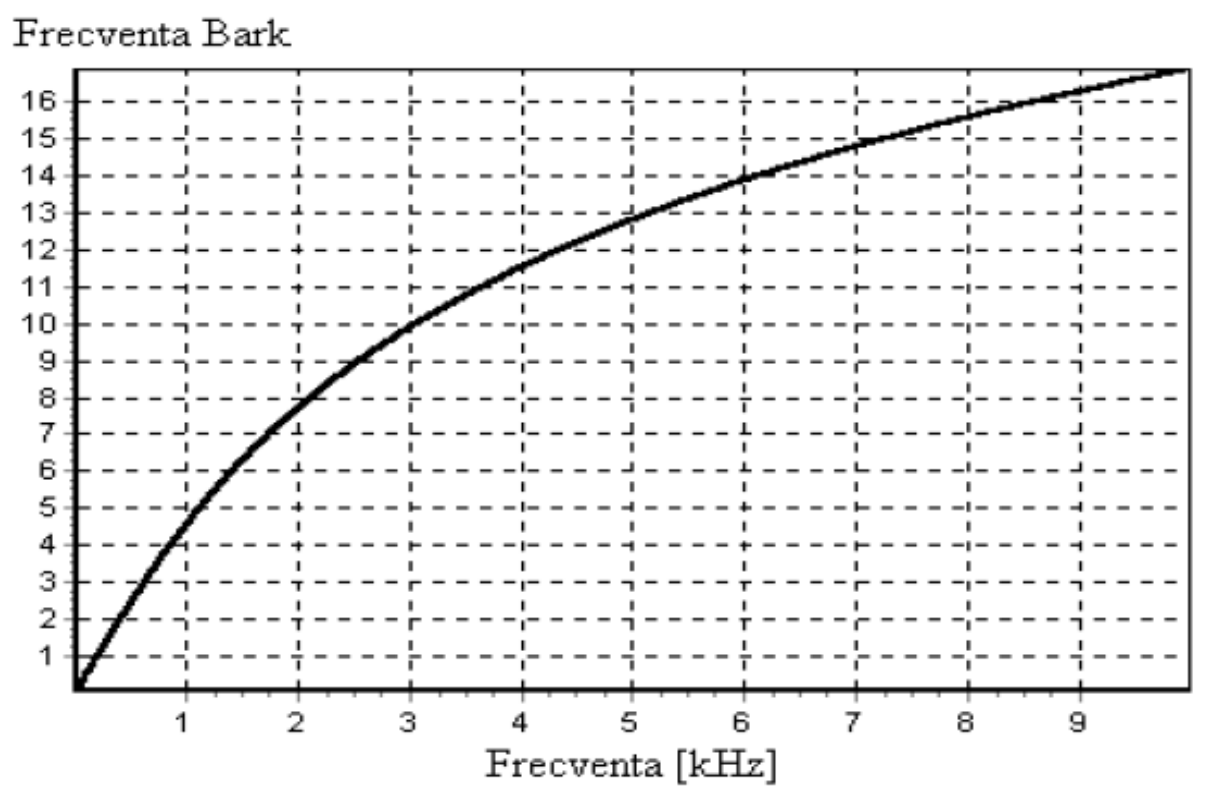
- **scara constanta Q**, unde Q este raportul dintre latimea de banda a filtrului si frecventa centrala, rezultand o forma/ scara exponentiala

- scara Bark, derivata din experimente de perceptie

- scara membranei bazilare, care se bazeaza pe distanta masurata de-a lungul acestei membrane si frecventa perceputa

- **scara Mel**, adoptata de ingineri (melodica)

Scara Bark

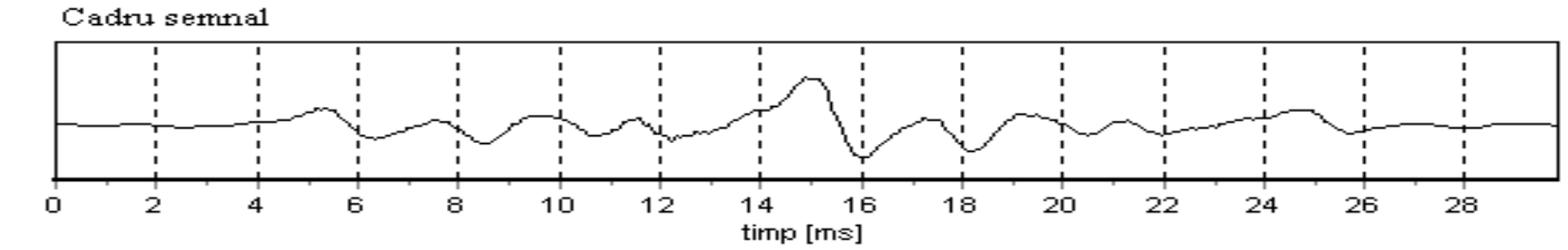


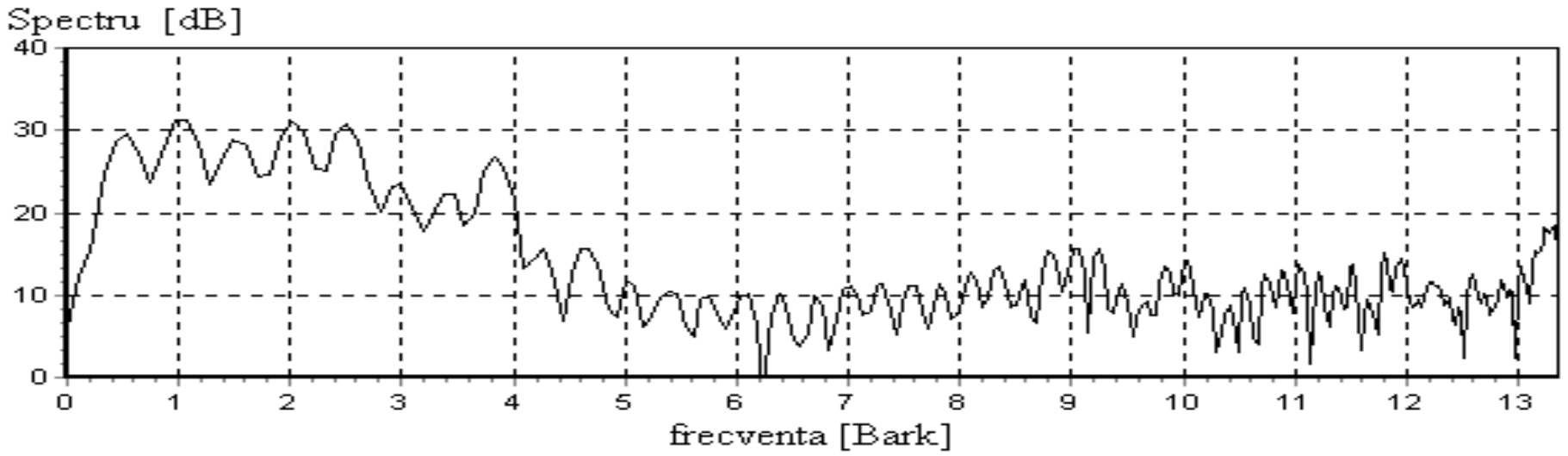
approx:

$$Bark(f) = \begin{cases} .01f, & 0 \le f < 500 \\ .007f + 1.5, & 500 \le f < 1220 \\ 6\ln(f) - 32.6, & 1220 \le f \end{cases}$$

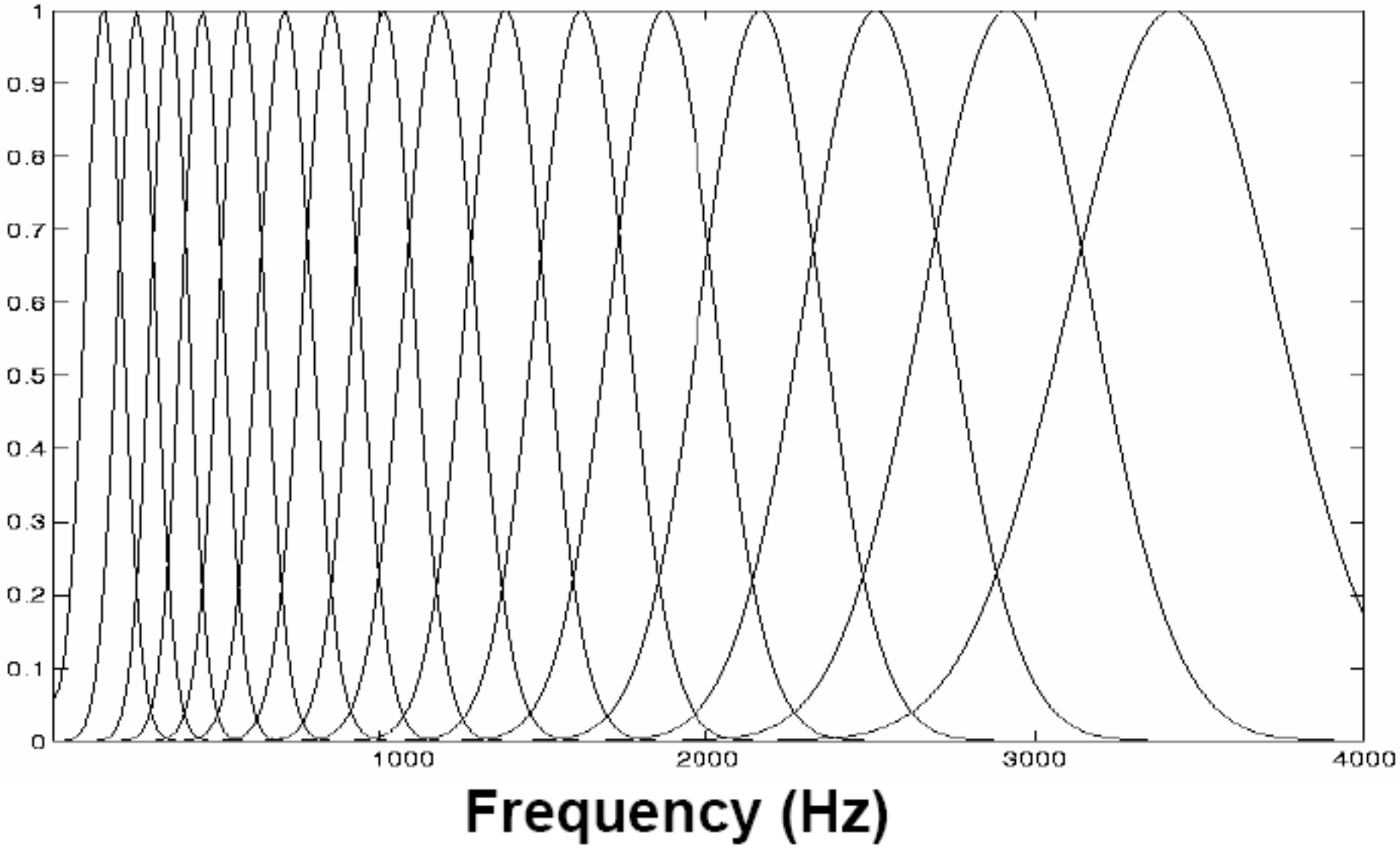
or,

$$Bark(f) = \left\{ 13\text{atan}\left(\frac{0.76\text{f}}{1000} \right) + 3.5\text{atan}\left(\frac{\text{f}^2}{(7500)^2} \right) \right.$$


Cadru semnal
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
timp [ms]


Spectru [dB]
40
30
20
10
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
frecventa [Bark]

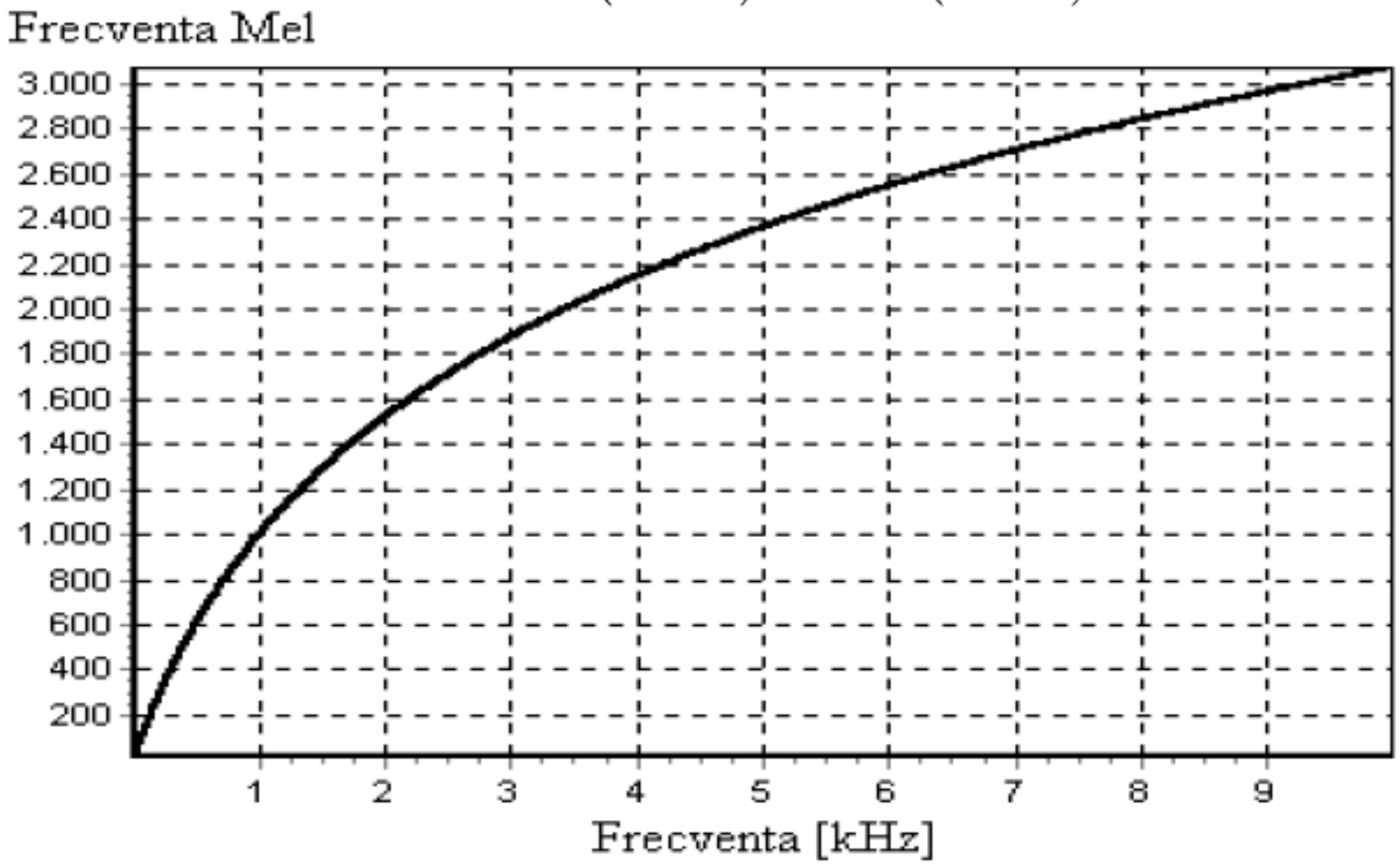
Bark Scale Bank of Filters

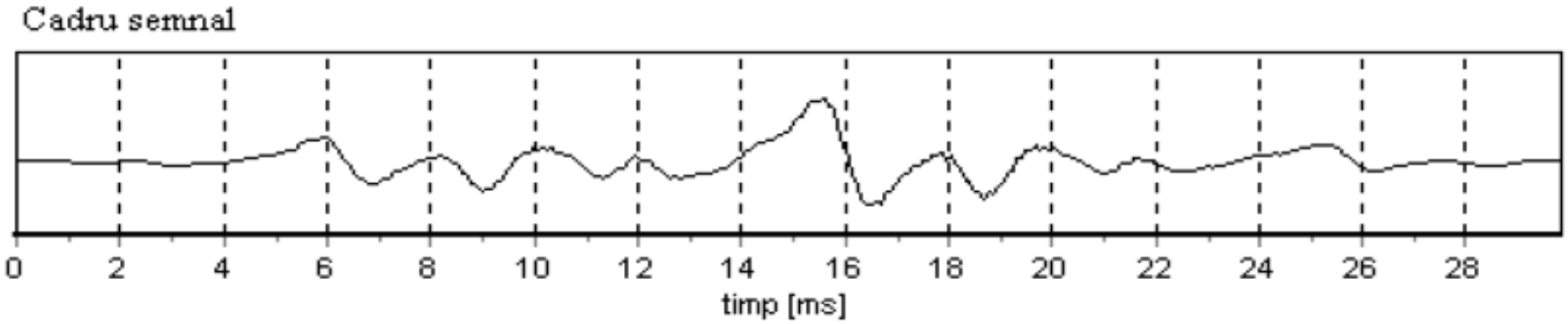


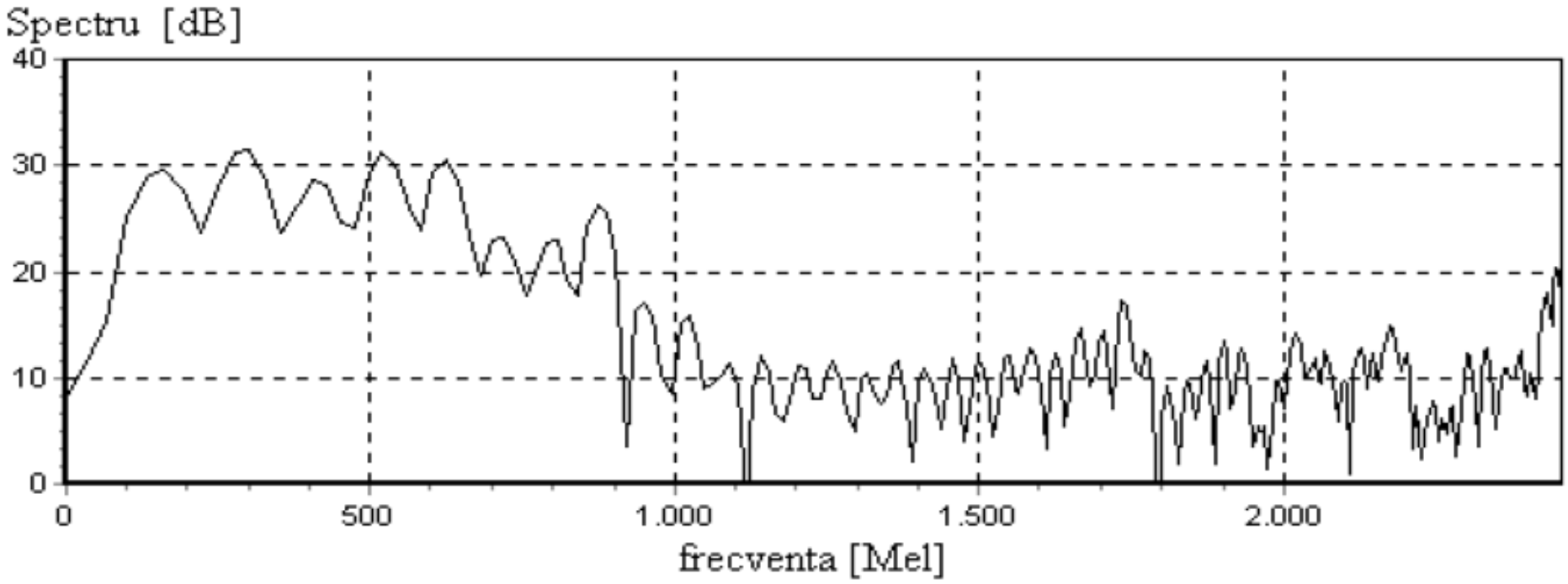
Scara Mel

$$Mel(f) = 2595 \cdot Log_{10} \cdot (1 + \frac{f}{700})$$

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) = 1127 \ln\left(1 + \frac{f}{700}\right).$$







$$BW_{crit} = 25 + 75\,[1 + 1.4 \cdot (f / 1000)^2]^{0.69}$$

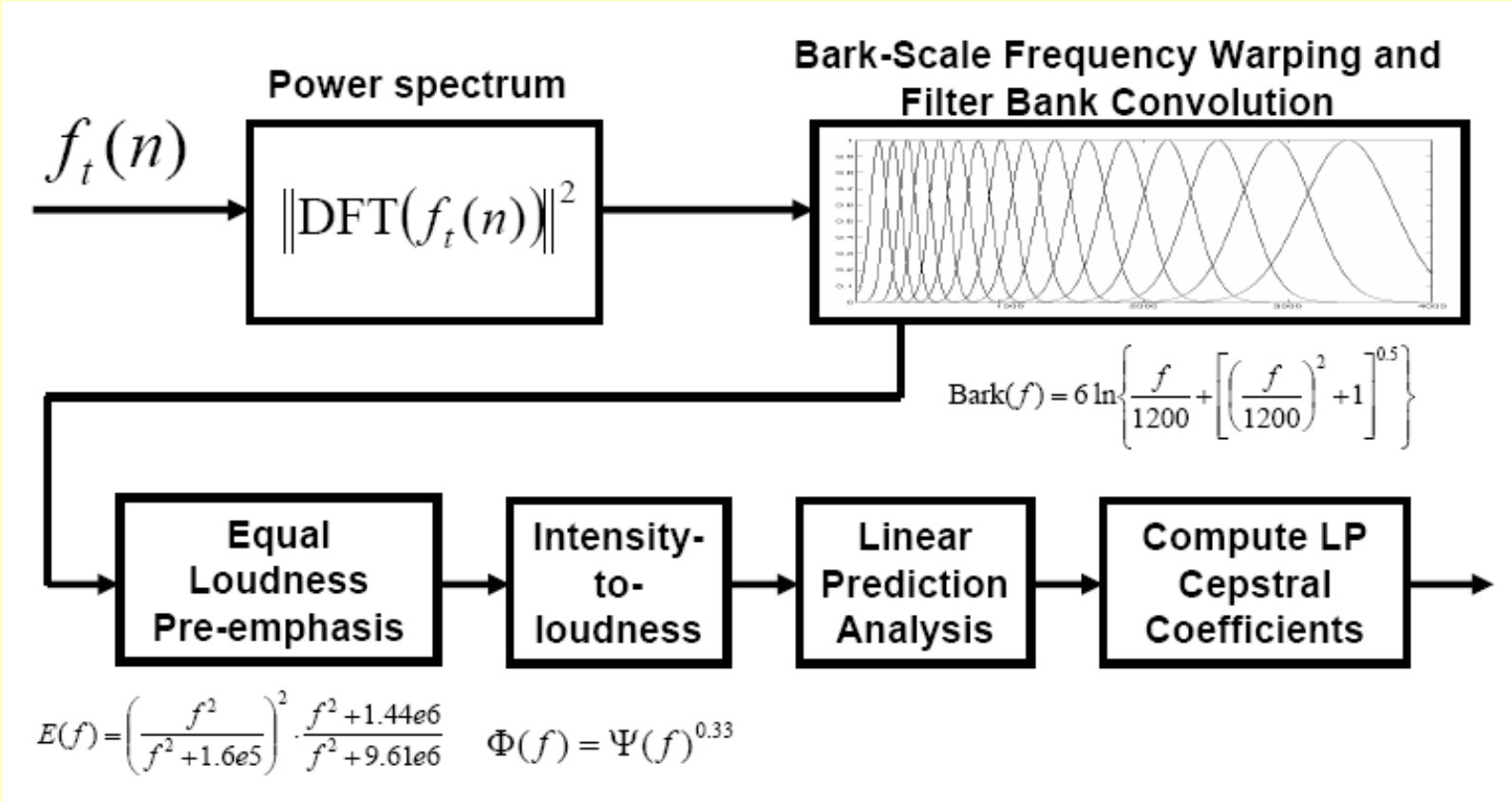
Nr.	Scara Bark		Scara Mel	
	FRECVENTA CENTRALA	BANDA BW(Hz)	FRECVENTA CENTRALA	BANDA BW(Hz)
1	50	100	100	100
2	150	100	200	100
3	250	100	300	100
4	350	100	400	100
5	450	110	500	100
6	570	120	600	100
7	700	140	700	100
8	840	150	800	100
9	1000	160	900	100
10	1170	190	1000	124
11	1370	210	1149	160
12	1600	240	1320	184
13	1850	280	1516	211
14	2150	320	1741	242
15	2500	380	2000	278
16	2900	451	2297	320
17	3400	550	2639	367
18	4000	700	3031	422
19	4800	900	3482	484
20	5800	1100	4000	556
21	7000	1300	4595	639
22	8500	1800	5278	734
23	10500	2500	6063	843
24	13500	3500	6964	969

Bancuri de filtre in scarile Bark si Mel

Analiza PLP (Perceptual Linear Prediction)

- PLP este o combinatie intre metoda FFT (Tr. Fourier) si metoda LPC
- introdusa de H. Hermansky si foloseste scara perceptuala Bark
- un mod de aliniere a spectrelor pentru a minimiza diferenţele dintre vorbitori, păstrând în acelaşi timp informaţii vocale importante.

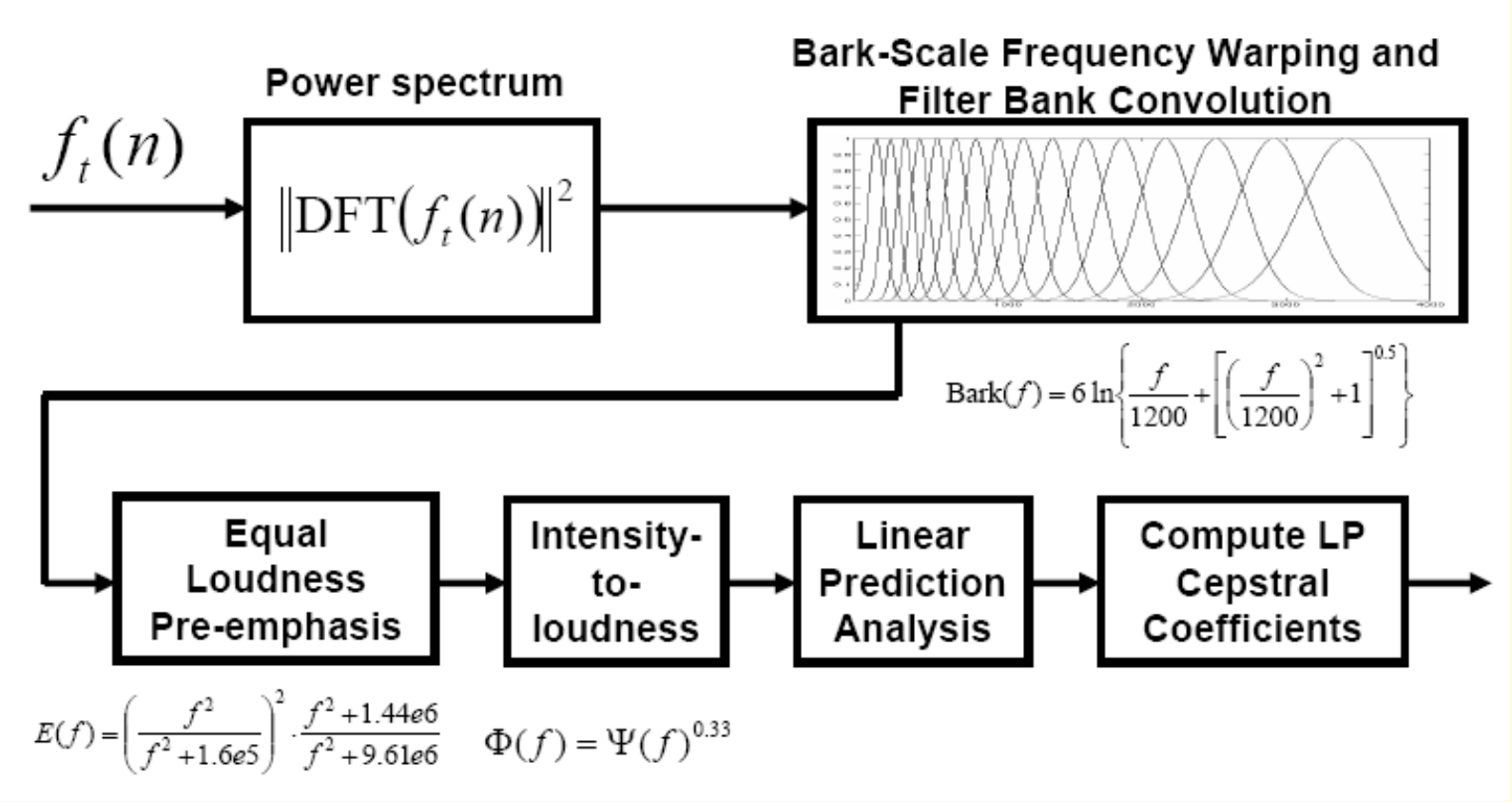
Ref: Paul R. Hill - Audio And Speech Processing With MATLAB-CRC Press (2018).



Etapele analizei PLP

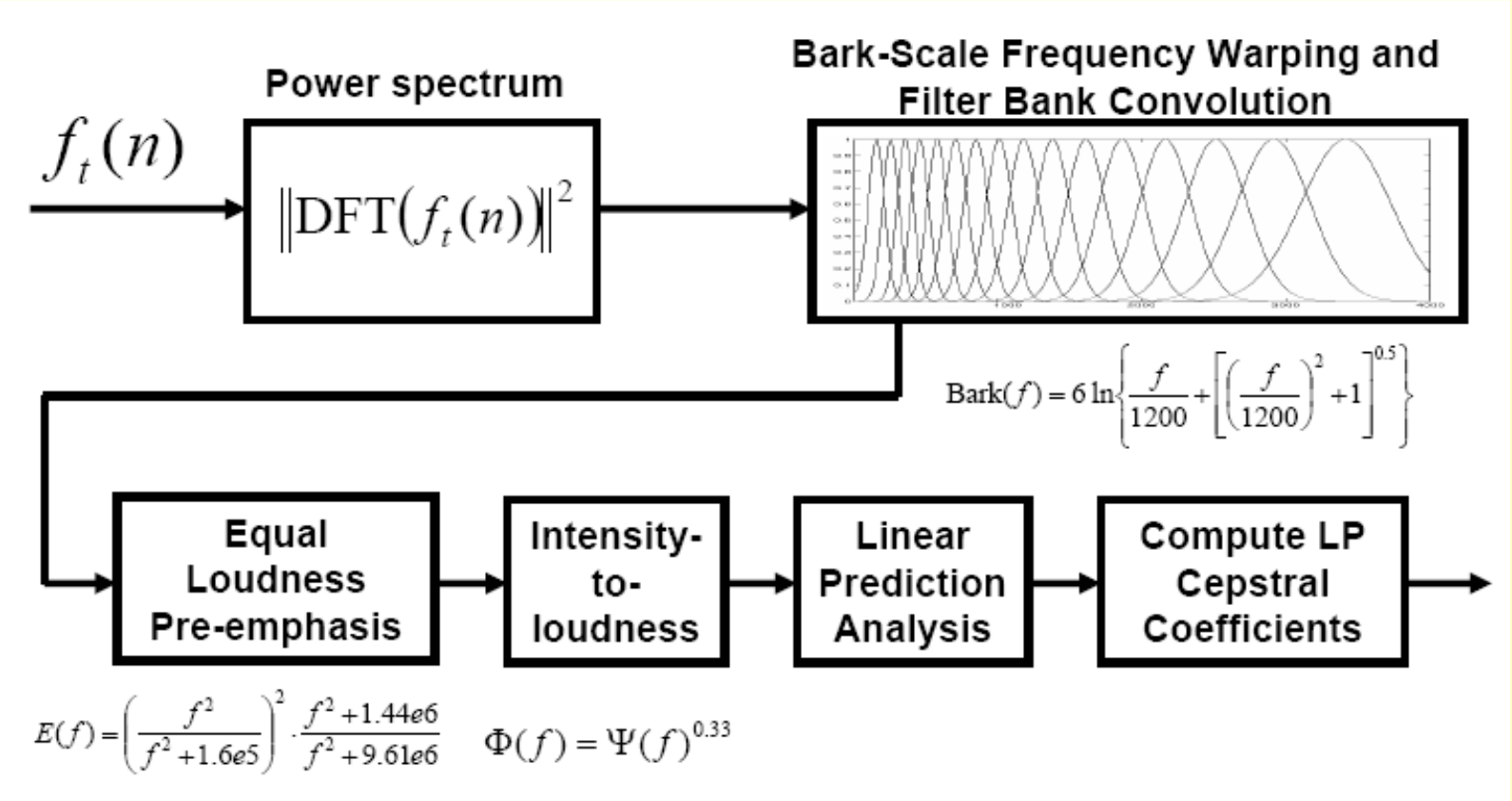
- analiza de banda critica;
- preaccentuare pentru egalizarea L;
- conversia intensitate (I) – tarie sonora (L);
- transformata Fourier inversa, IFFT;
- solutii pentru coeficientii autoregresivi;
- construire model numai poli;
- calcul coeficienti cepstrali

1. Semnalul vocal este iniţial supus unei analize spectrale, folosind segmente de 20ms lungime şi fereastra Hamming şi se obţine *spectrul de putere pe termen scurt (FFT).* Spre deosebire de analiza LPC nu se face preaccentuarea semnalului vocal deoarece în analiza PLP apare o etapă speciala de preaccentuare.



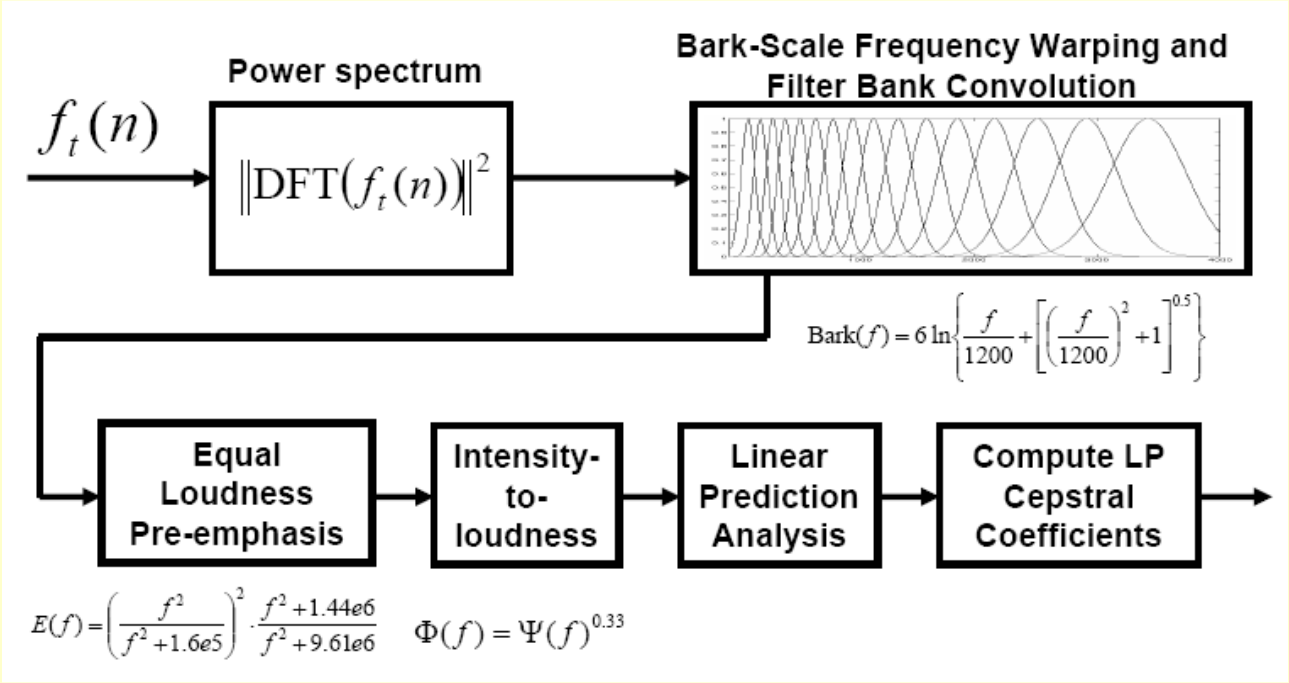
2. Spectrul de putere este aliniat dupa scara Bark :

$$\Omega(\omega) = 6\ln\left(\omega/1200\pi + \sqrt{(\omega/1200\pi)^2 + 1}\right)$$

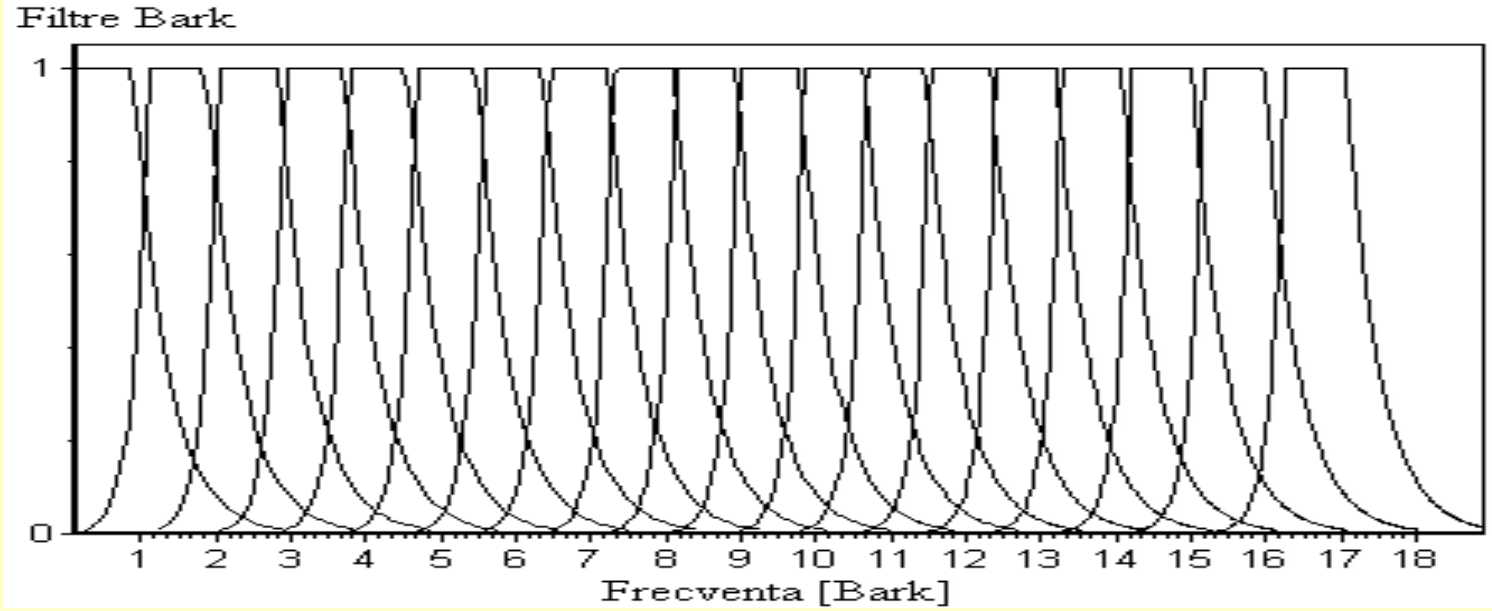


3. Spectrul aliniat Bark este supus unei convolutii cu spectrul de putere al filtrului de banda critica - se reduce rezolutia spectrala.

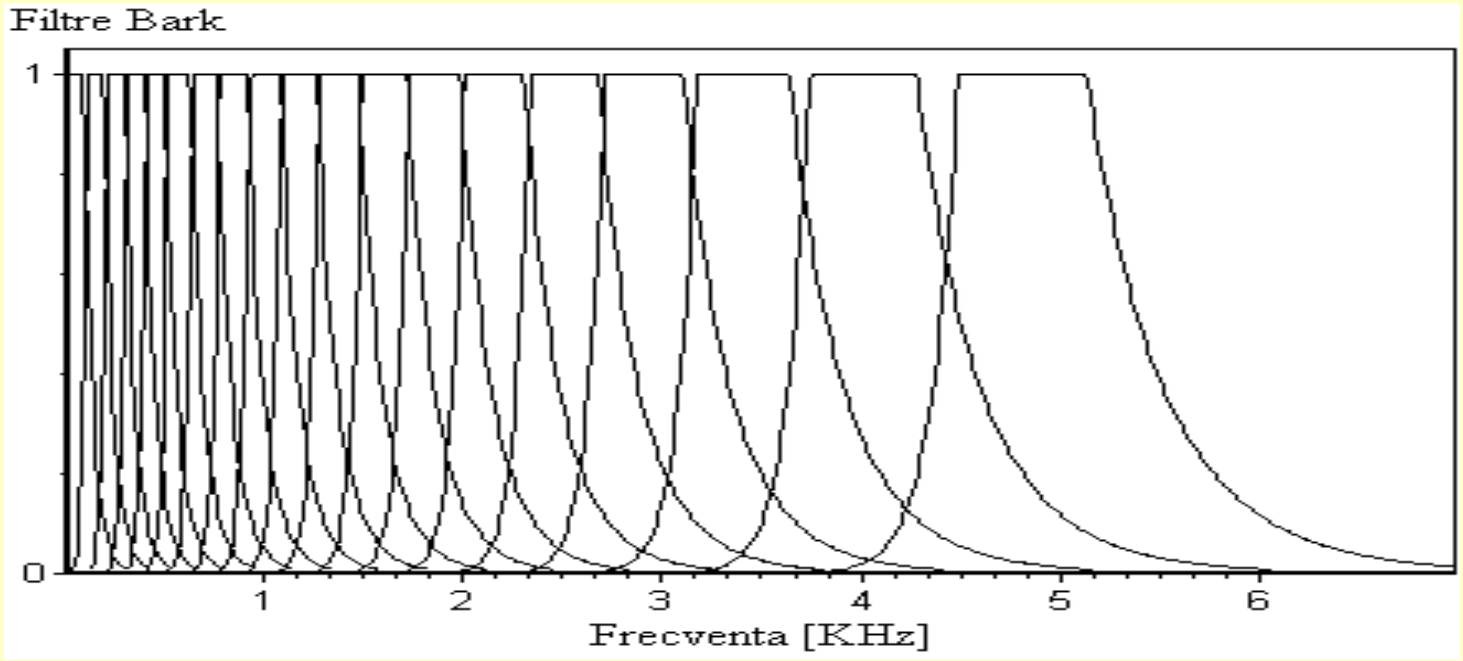
$$\Psi(\Omega) = \begin{cases} 0, & \Omega < -1.3 \\ 10^{2.5(\Omega + 0.5)}, & -1.3 < \Omega < -0.5 \\ 1, & -0.5 < \Omega < 0.5 \\ 10^{-1.0(\Omega - 0.5)}, & 0.5 \le \Omega \le 2.5 \\ 0, & \Omega > 2.5 \end{cases}$$



4. Spectrul netezit aliniat dupa scara Bark este subesantionat, prin reesantionare la 1Bark, tinand cont ca intervalului 0-5kHz de frecvente liniare ii corespunde intervalul 0-16.9 Bark.



Banc de filtre dreptunghiulare pe scara Bark



Banc de filtre dreptunghiulare pe scara liniară

5. ***Preaccentuarea pentru egalizarea tariei sonore (L)*** este necesara pentru a compensa perceptia neliniara a tariei sonore la diferite frecvente.
Preaccentuarea se face folosind relatia :

$$E(\omega) = \frac{\left(\omega^2 + 56.8 \cdot 10^6\right) \cdot \omega^4}{\left(\omega^2 + 6.3 \cdot 10^6\right) \cdot \left(\omega^2 + 0.38 \cdot 10^9\right) \cdot \left(\omega^6 + 9.58 \cdot 10^{26}\right)}$$

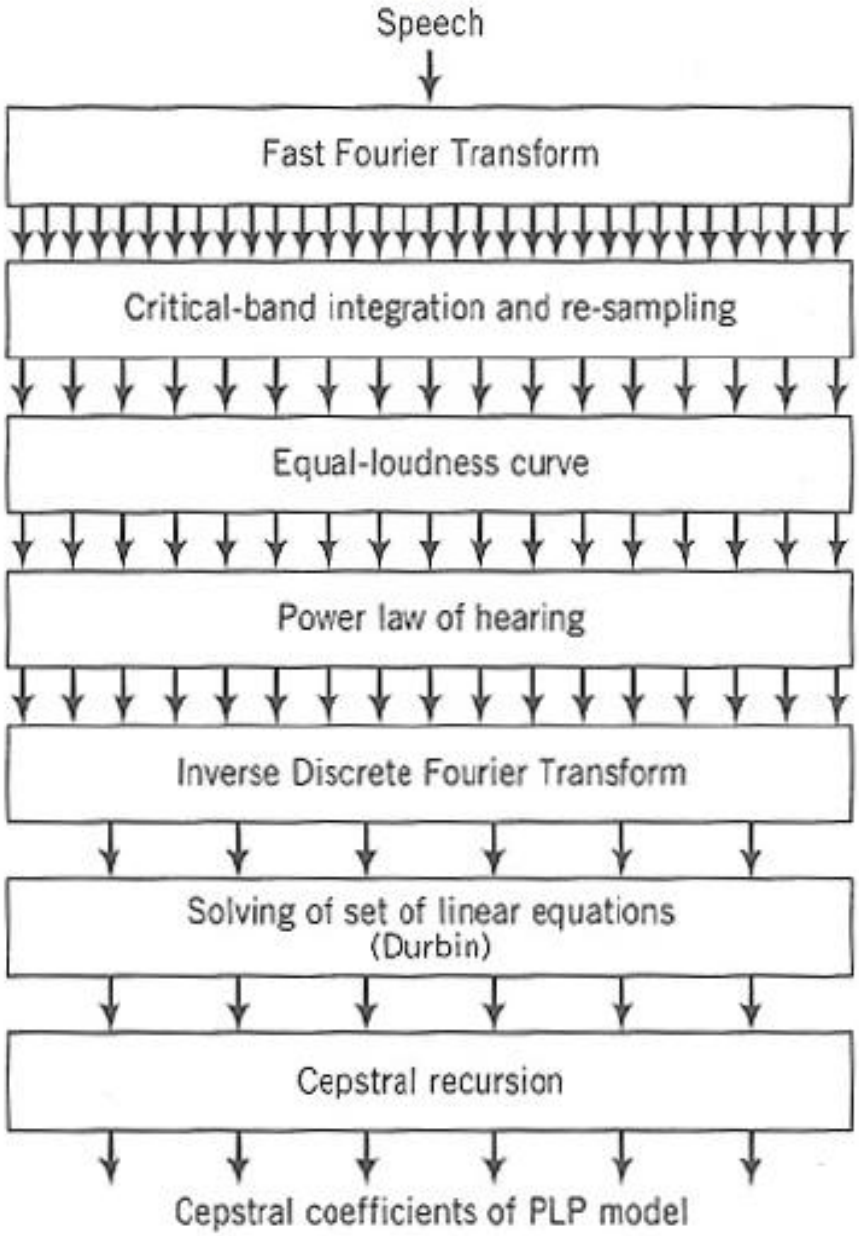
6. Conversia intensitate I(ω) – tarie sonora L(ω) se face dupa legea :

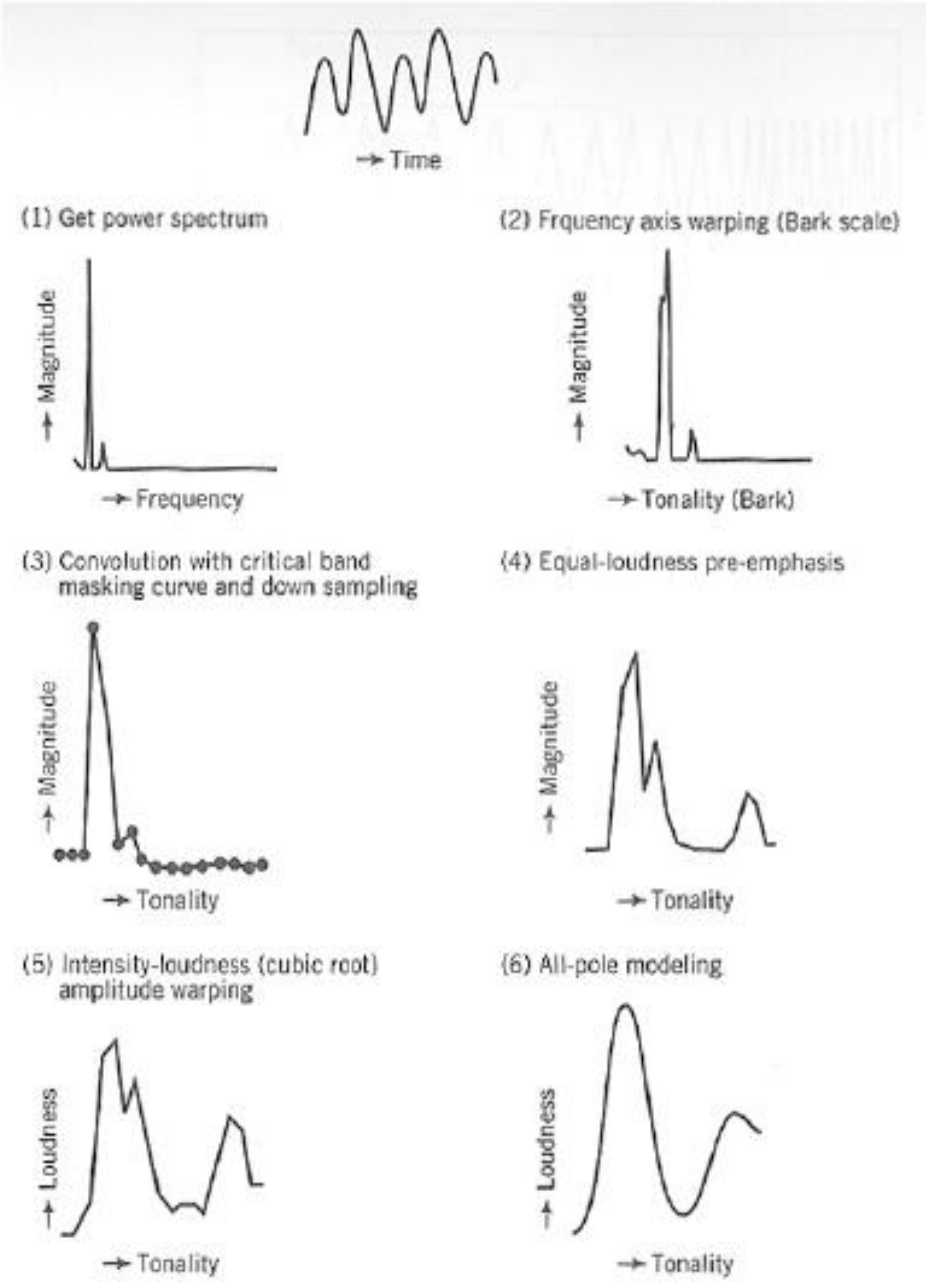
$$L(\omega) = I(\omega)^{1/3}$$

7. Aplicarea transformatei Fourier inverse (IFFT) >>> functia de autocorelatie

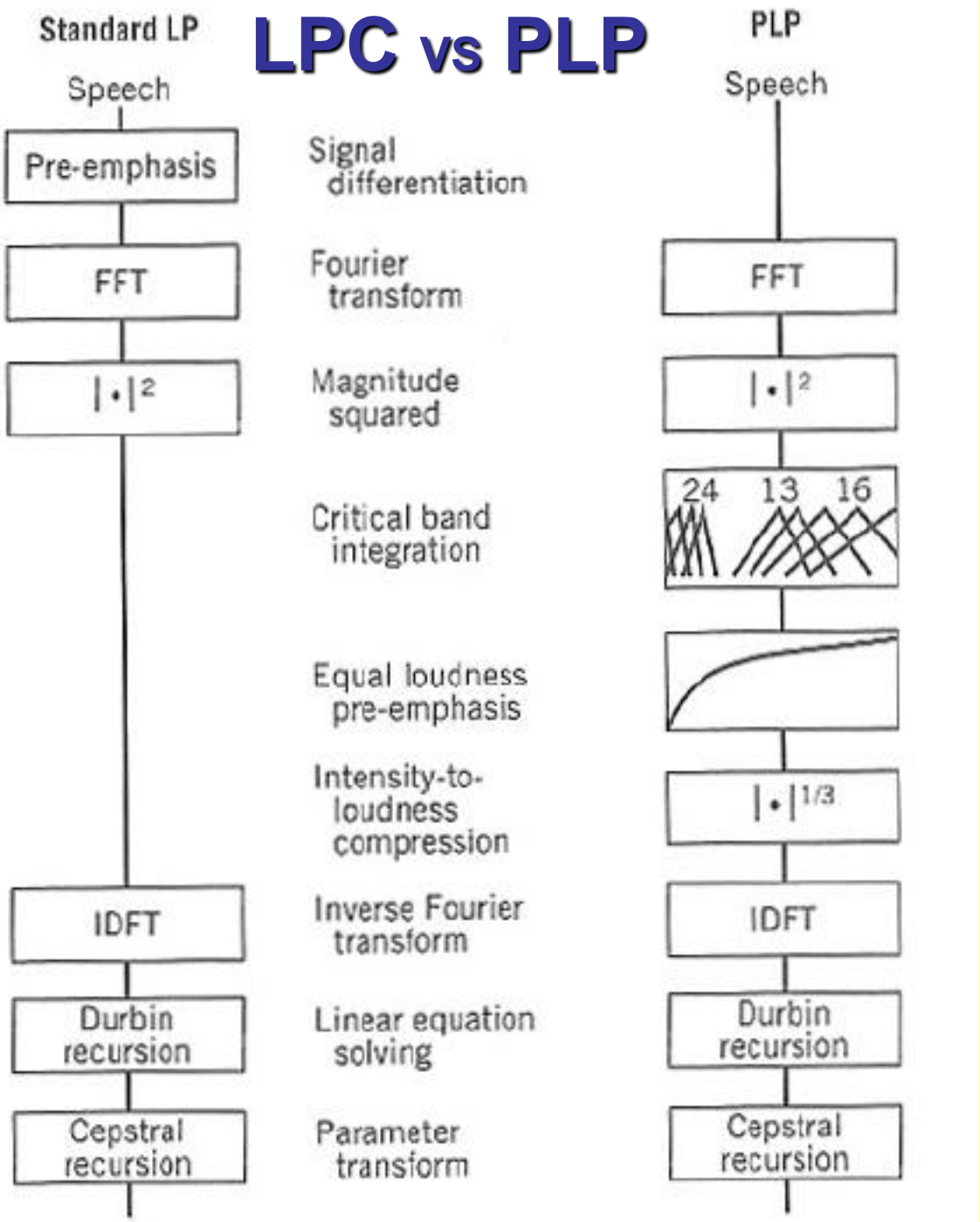
8. Rezulta coeficientii modelului de predictie liniara LPC (eventual coef. cepstrali)

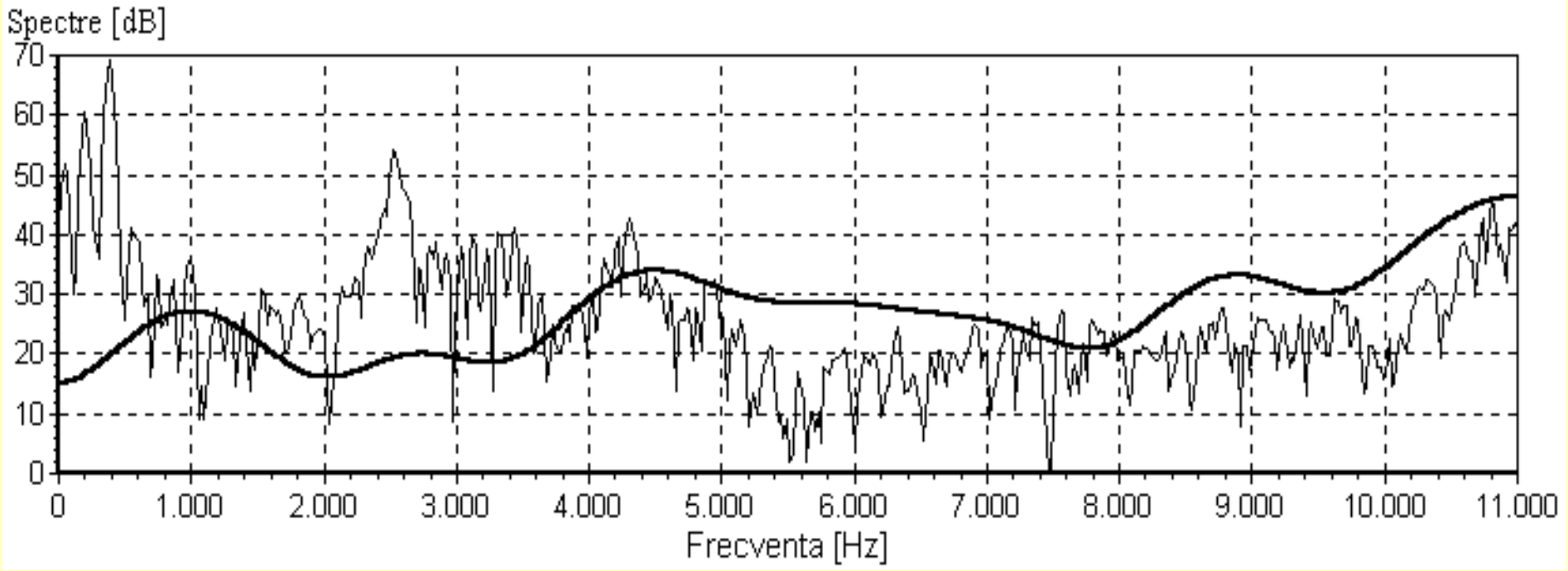
Perceptual Linear Prediction



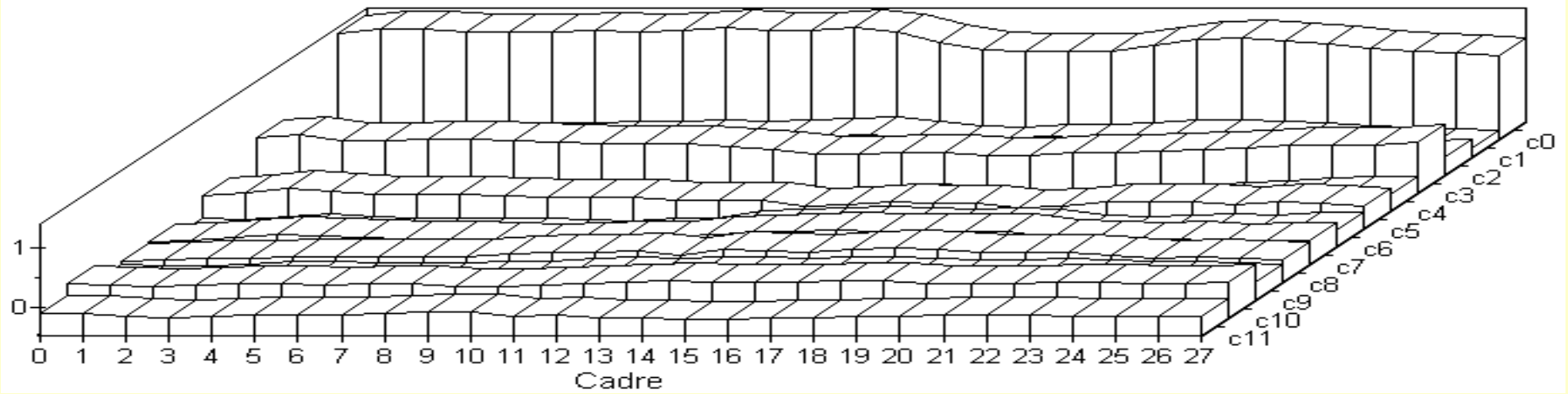


LPC vs PLP

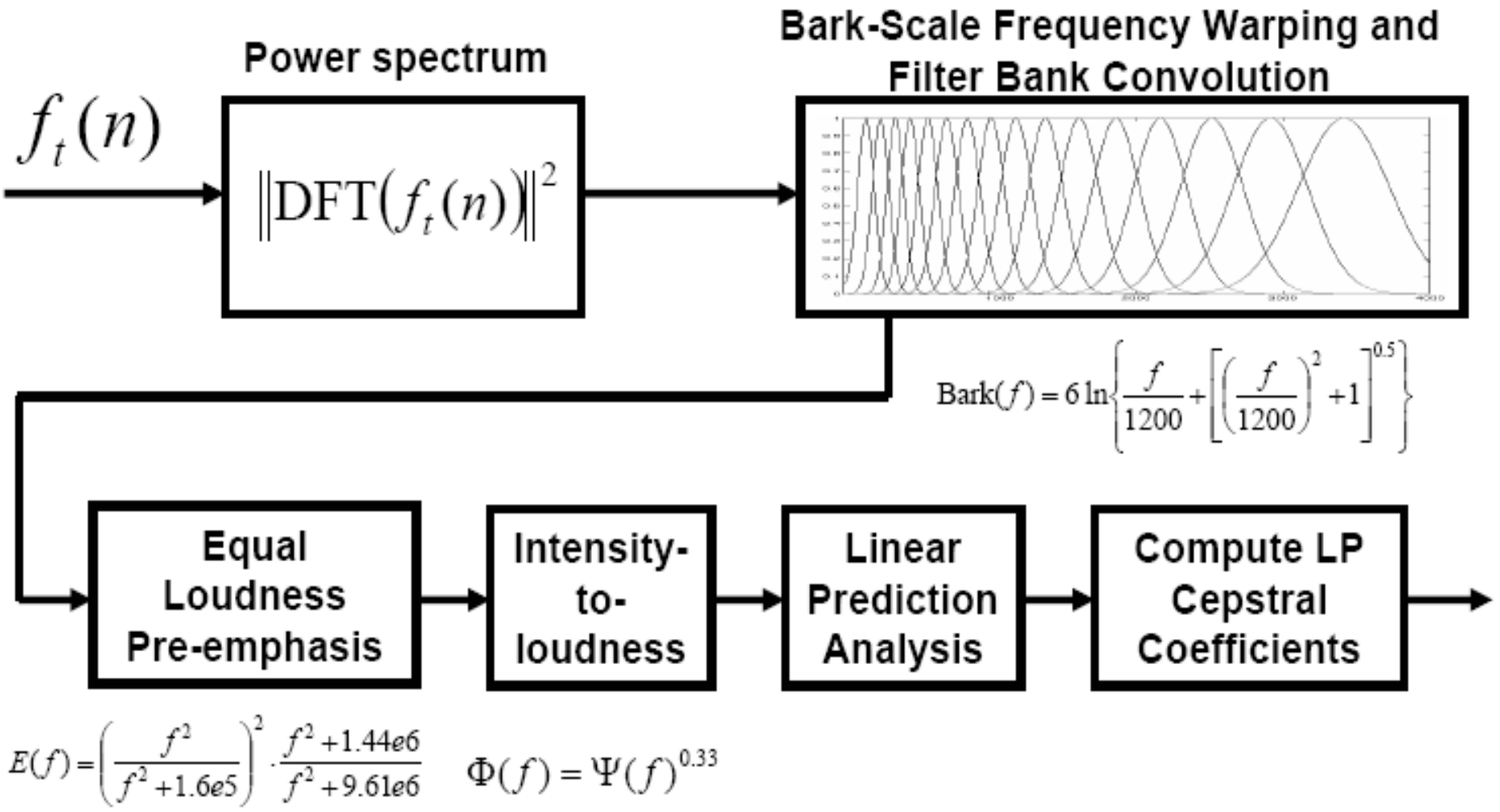

Standard LP
Speech
Pre-emphasis
FFT
|•|²
IDFT
Durbin recursion
Cepstral recursion
Signal differentiation
Fourier transform
Magnitude squared
Critical band integration
Equal loudness pre-emphasis
Intensity-to-loudness compression
Inverse Fourier transform
Linear equation solving
Parameter transform
PLP
Speech
FFT
|•|²
24
13
16
|•|^1/3
IDFT
Durbin recursion
Cepstral recursion



Spectrul FFT şi spectrul PLP, P=11.



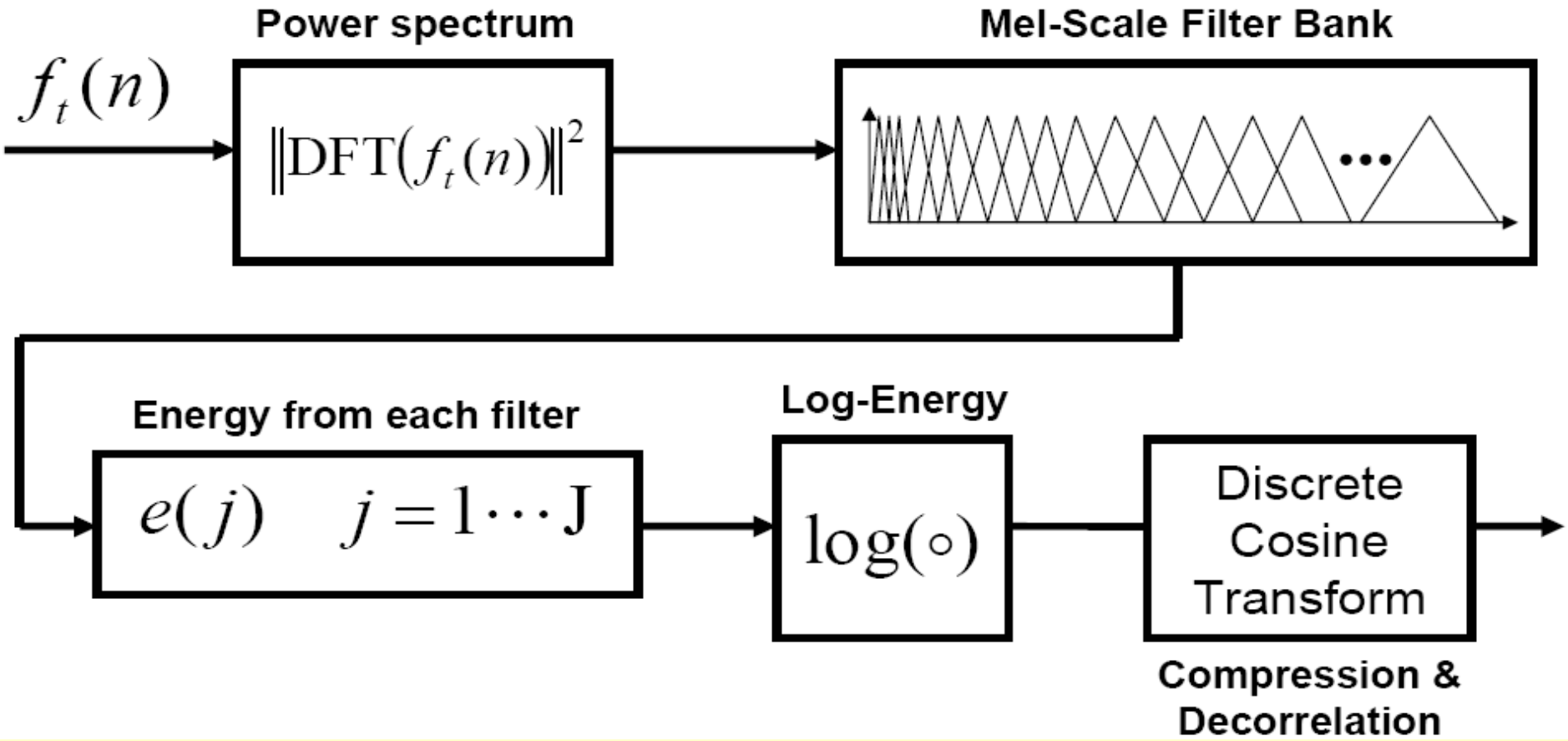
Evoluţia coeficienţilor PLP (p=12) pentru cuvântul “unu”.



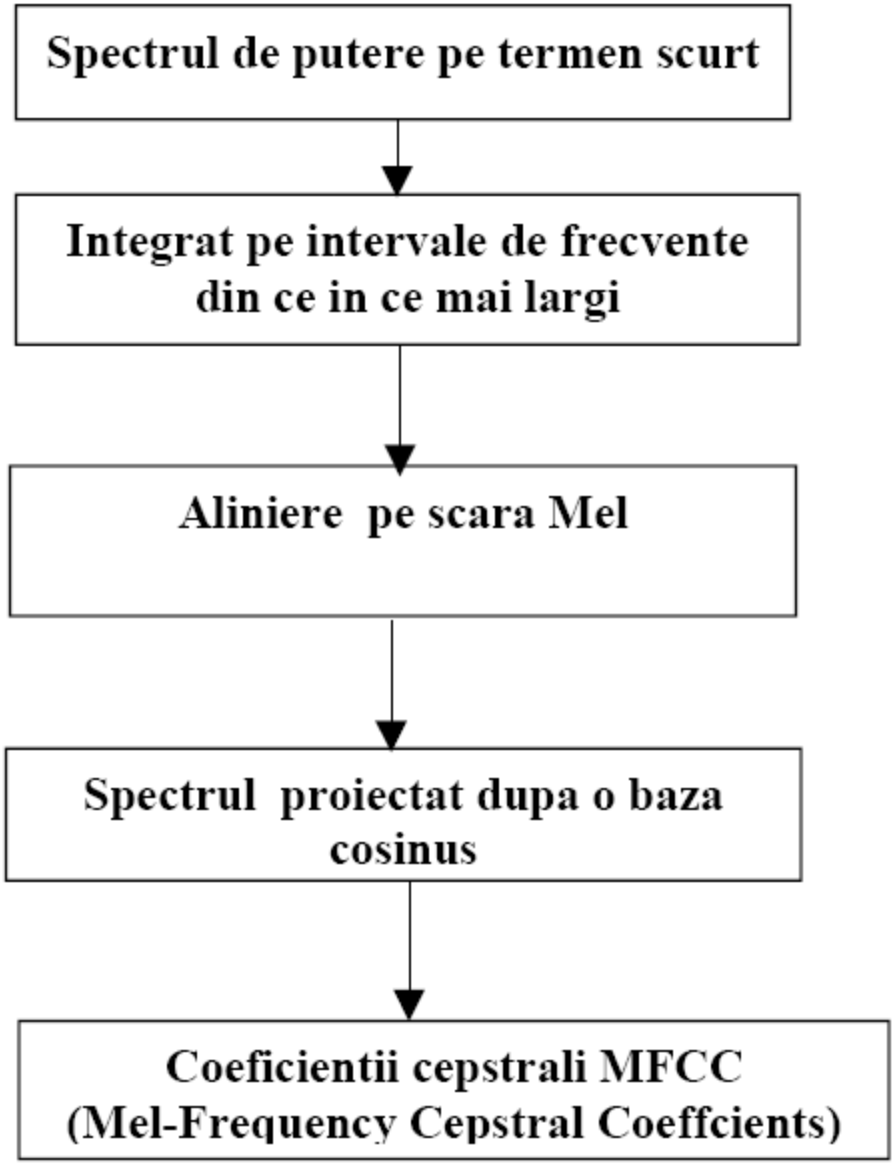
PLP - Include aspecte perceptuale în recunoastere - mai robusta decât coef. cepstrali de predicţie liniară (LPCCs).

- STUDIU: Analiza RASTA-PLP

Analiza Cepstrala MEL



http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/
http://mirlab.org/jang/books/audiosignalprocessing/speechFeatureMfcc.asp?title=12-2%20MFCC
https://www.youtube.com/watch?v=hF72sY70_IQ



- $X_a(k) = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N}, 0 \le k \le N$

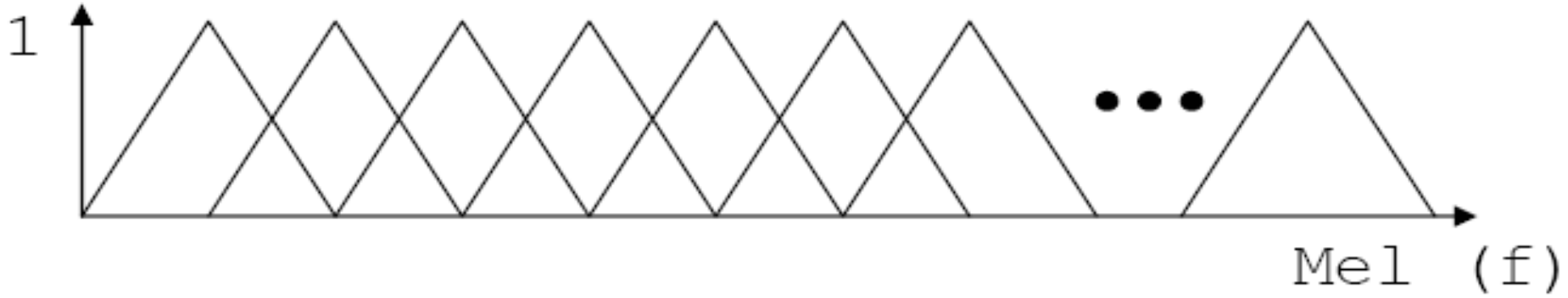
- Bancul de filtre folosit :

$$w_j(f) = \begin{cases} \dfrac{f - f_{j-1}}{f_j - f_{j-1}}, & f_{j-1} \le f \le f_j \\ \dfrac{f_{j+1} - f}{f_{j+1} - f_j}, & f_j \le f \le f_{j+1} \\ 0, & altfel \end{cases}$$

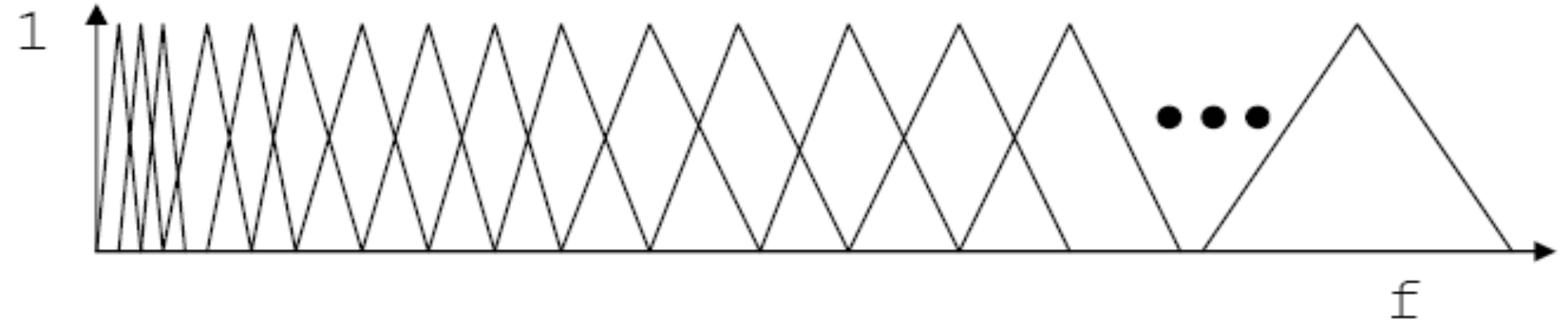
unde frecventele fj sunt date de relatia :

$$f_j = \begin{cases} 100 * j, & 0 \le j \le 10 \\ 1000 * (1.15)^{j-10}, & j > 10 \end{cases}$$

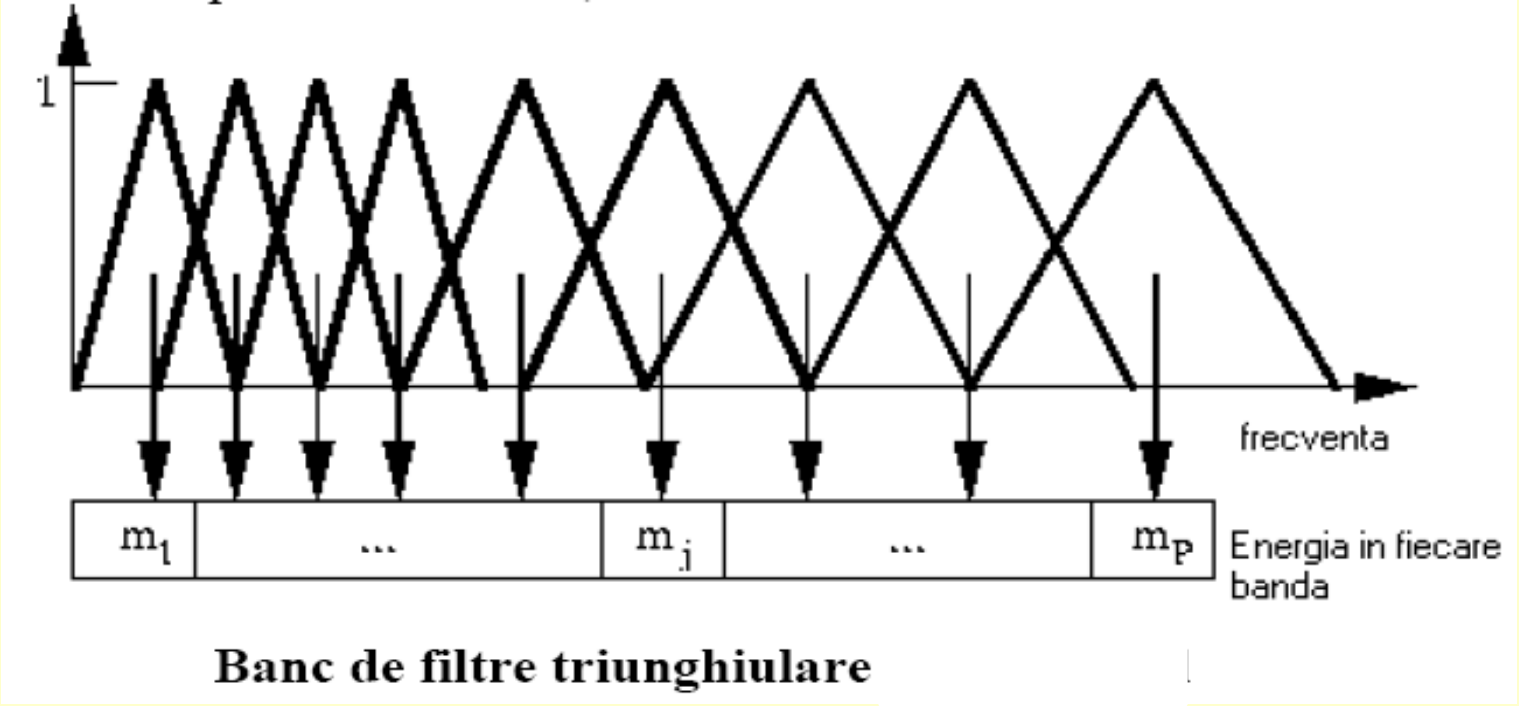
ptr. F_e = 16KHz , f_{max} =8KHz rezulta P=25.



Banc de filtre triunghiulare pe scara MEL



Banc de filtre triunghiulare pe scara liniara



Banc de filtre triunghiulare

- Rezulta vector cu energii logaritmate pentru cele P filtre :

$$E_j = \log\left[\frac{\sum_{i=0}^{N-1} |X_i|^2 \cdot w_j\left(\frac{i}{N}\cdot f_{\max}\right)}{\sum_{i=0}^{N-1} w_j\left(\frac{i}{N}\cdot f_{\max}\right)}\right], j = 1,\ldots, P$$

- Valorile energiilor Ej sunt medii ponderate ale valorilor spectrului de putere, pe zona de frecvente acoperita de filtrul cu indicele j.

• forma spectrului dat de tractul vocal este netezita, nivelele de energie din benzile adiacente tind sa fie corelate. De aceea pentru a obtine coeficientii MFCC, se aplica transformarea cosinus pentru a obtine decorelarea specifica coeficientilor cepstrali:

$$Y_i = \sqrt{\frac{2}{P}} \cdot \sum_{j=1}^{P} E_j \cdot \cos\left[i \cdot \left(j - \frac{1}{2} \right) \cdot \frac{\pi}{P} \right], \quad i = 0,\ldots,\ M$$

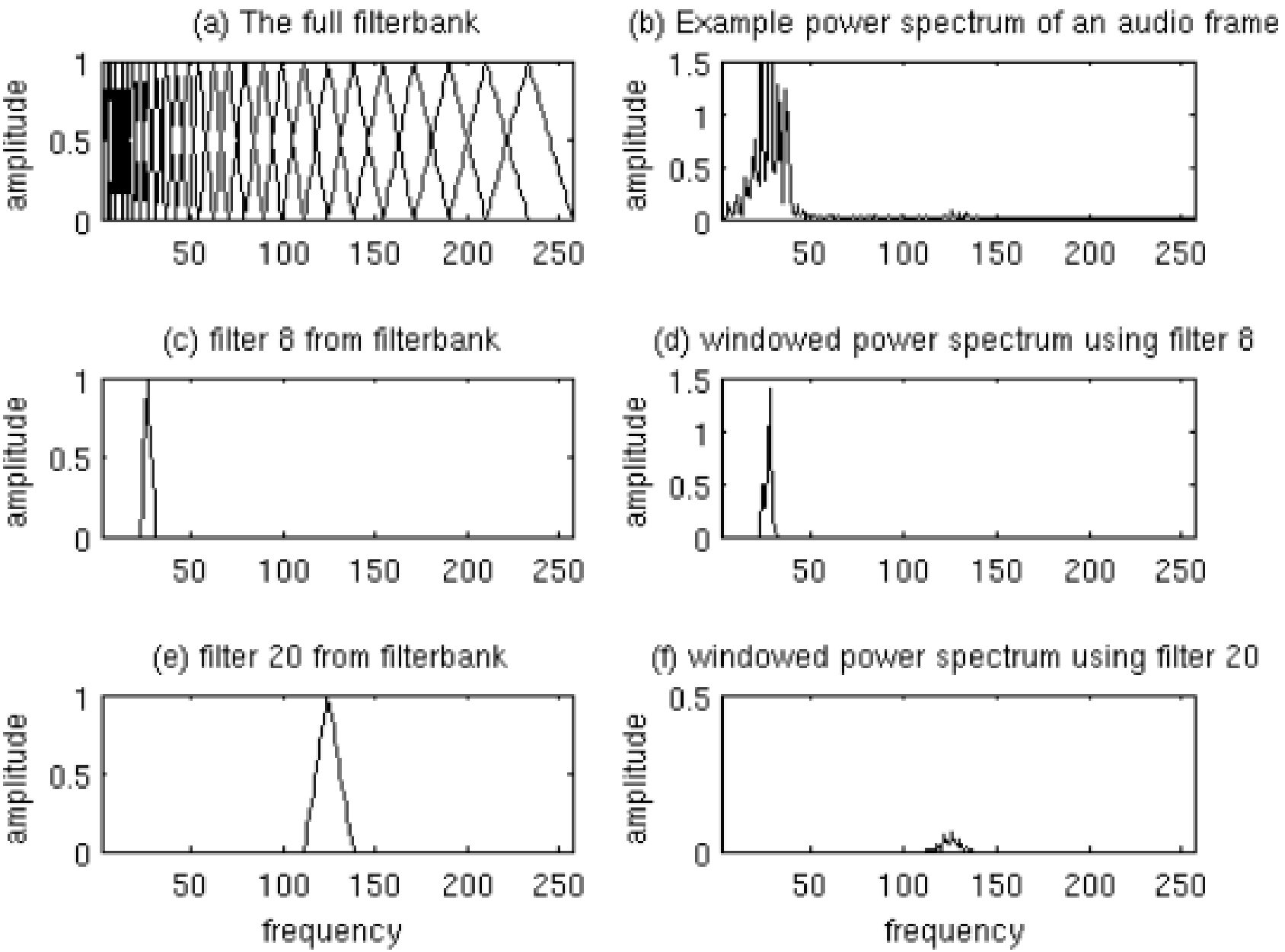
unde M este ordinul MFCC (24…..40).

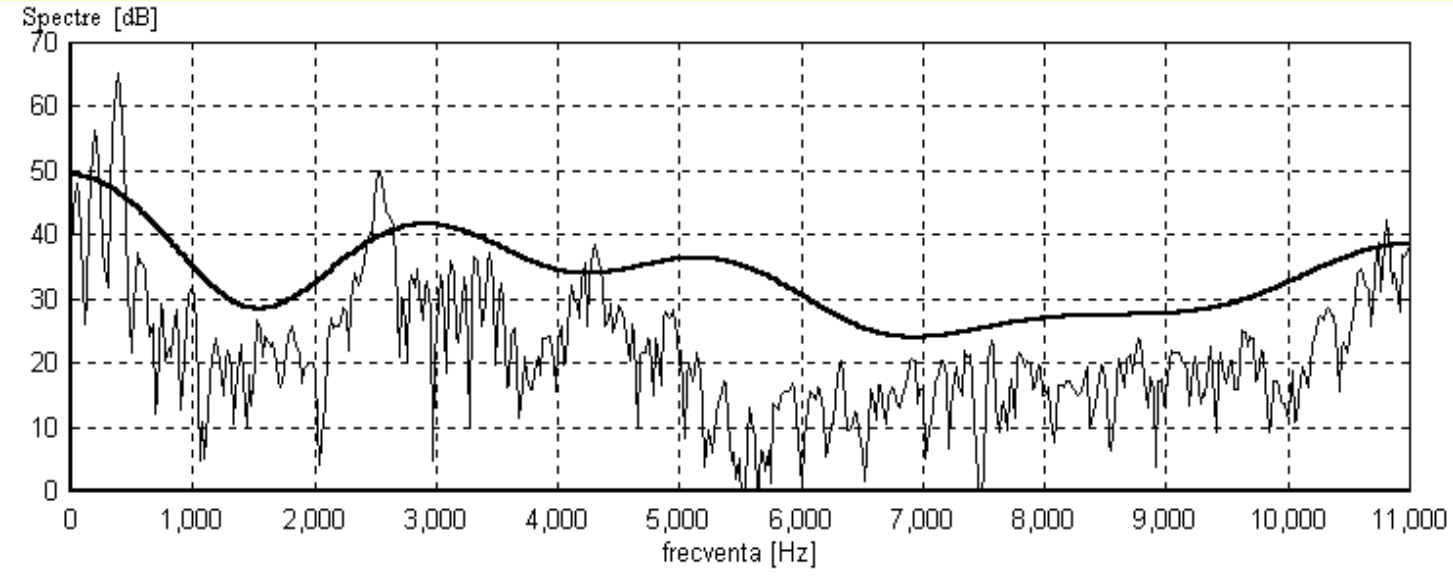
• rescalare a coeficientilor MFCC, prin "liftare" pana la valoarea L, utlizand formula :

$$c'_n = \left(1 + \frac{L}{2} \cdot \sin \frac{\pi \cdot n}{L} \right) \cdot c_n$$

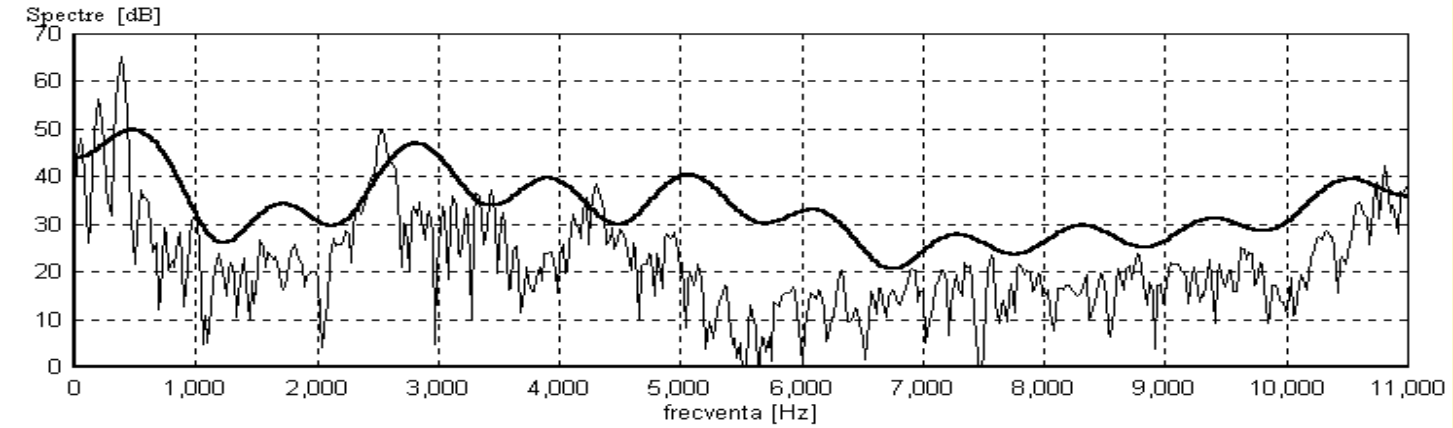
- Primul coeficient DCT (c0) reprezintă energia logaritmică medie a cadrului (offset DC).
- Coeficienții următori (c1, c2, c3...) reprezintă forma spectrală în detalii din ce în ce mai fine.
- (c1) corespunde înclinării spectrale (există mai multă energie în frecvențele joase sau în cele înalte?).
- (c2) corespunde „denivelării" spectrului.

Acest lucru ne permite să efectuăm o selecție extrem de eficientă a caracteristicilor: pur și simplu *păstrăm primii 12-13 coeficienți și renunțăm la restul*. Renunțăm la coeficienții de ordin superior, care reprezintă detalii spectrale rapide și fine, adesea mai mult legate de zgomot sau de caracteristicile individuale ale vorbitorului decât de conținutul fonetic.

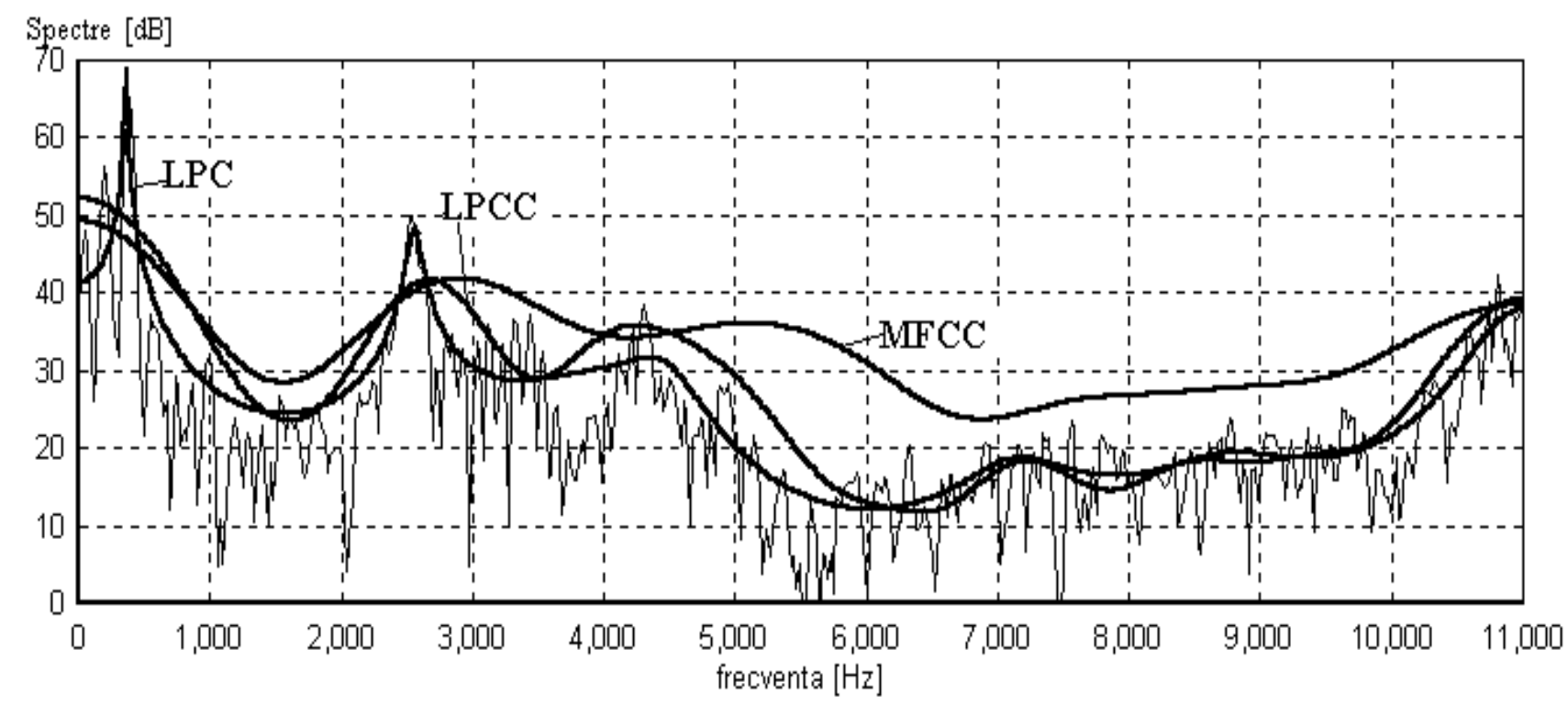

(a) The full filterbank
amplitude
1
0.5
0
50
100
150
200
250
(b) Example power spectrum of an audio frame
amplitude
1.5
1
0.5
0
50
100
150
200
250
(c) filter 8 from filterbank
amplitude
1
0.5
0
50
100
150
200
250
(d) windowed power spectrum using filter 8
amplitude
1.5
1
0.5
0
50
100
150
200
250
(e) filter 20 from filterbank
amplitude
1
0.5
0
50
100
150
200
250
frequency
(f) windowed power spectrum using filter 20
amplitude
0.5
0
50
100
150
200
250
frequency



Spectrul FFT şi spectrul MFCC, m=14.



Spectrul FFT şi spectrul MFCC, m=20.



Spectrele unui segment vocal (FFT, LPC–16, LPCC-20, MFCC-16).

De ce sunt coeficientii MFCC atât de populari?

-Forma tractului vocal se manifestă în anvelopa spectrului de putere pe timp scurt, iar sarcina MFCC este de a reprezenta cu precizie acest anvelopă.
- sunt eficienti şi relativ simplu de calculat
- sunt „dotati” cu o scară de frecvenţă perceptuala
- bancul de filtre reduce impactului excitaţiei în setul final de caracteristici
- DCT decoreleaza caracteristicile (sursa, tract)

-Valorile MFCC nu sunt foarte robuste în prezența zgomotului aditiv, astfel încât este obișnuit să se normalizeze valorile acestora în sistemele de recunoaștere a vorbirii pentru a reduce influența zgomotului. Unii cercetători propun modificări ale algoritmului MFCC de bază pentru a îmbunătăți robustețea, cum ar fi ***ridicarea amplitudinilor log-mel*** la o putere adecvată (în jur de 2 sau 3) înainte de a efectua transformarea cosinus discretă (DCT), ceea ce reduce influența componentelor cu energie redusă.

Tutorial MFCC

-http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

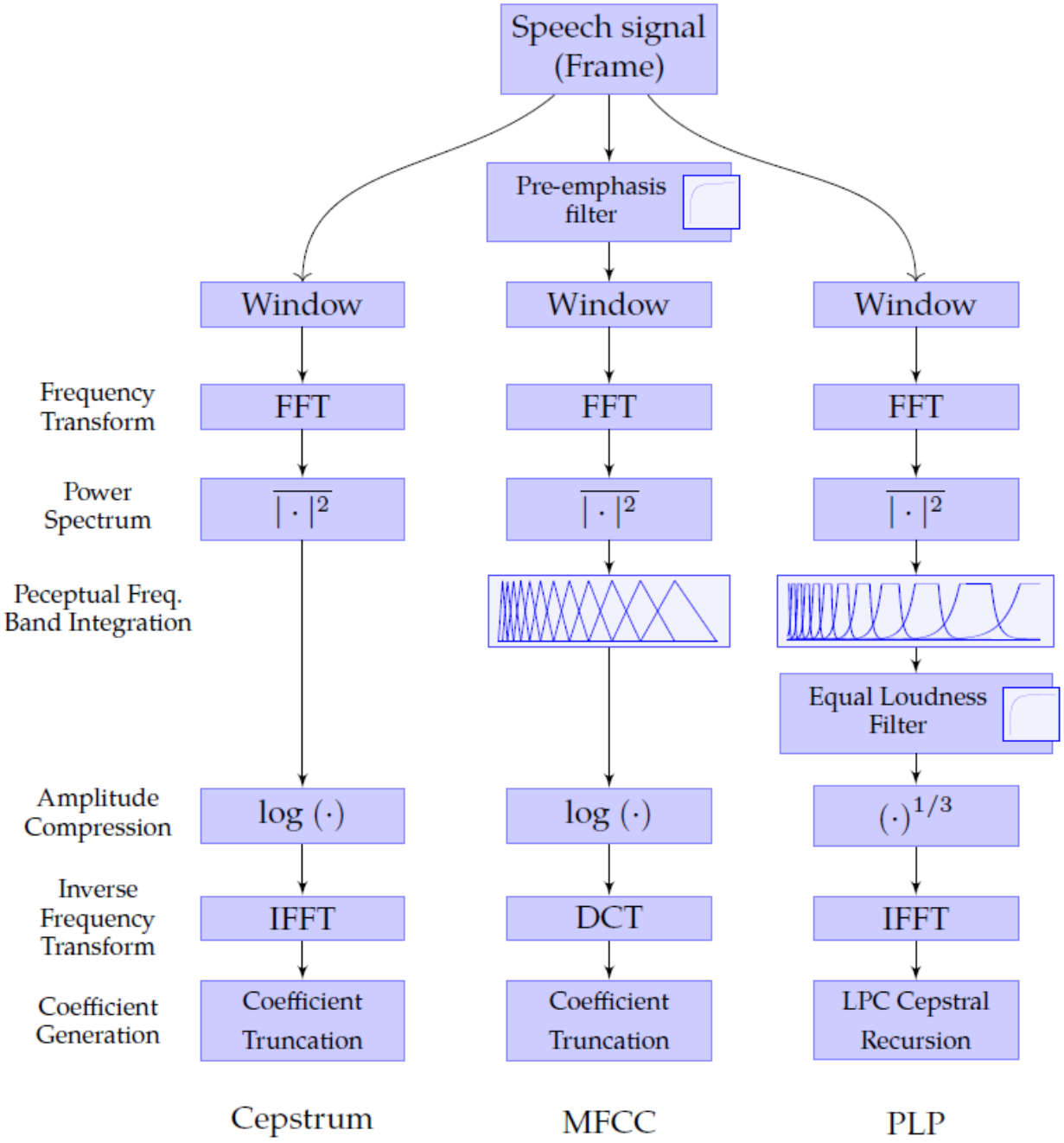


FIGURE 9.12: Cepstrum, PLP and MFCC feature extraction methods.

Modelele auditive

- Analiza de frecvenţă neuniforma este esenţială pentru a potrivi percepţia pitch-ului

- La unele tipuri de compresie log este esenţială pentru a potrivi percepţia tariei

- Unele tipuri de AGC (Automatic Gain Control) este esenţială a se potrivi mascarea percepţiei

- Ambele caracteristici *temporale şi spectrale* sunt esentiale şi poartă informaţii importante

- Atât pe termen lung (~ 200 msec) cat şi pe termen scurt (~ 20 msec) analizele sunt esenţiale pentru a caracteriza unităţile de lungimea silabei şi unităţile de lungimea fonemelor

Parametrii dinamici

• Unele sisteme de recunoaşterea vorbirii/vorbitorului folosesc viteza de schimbare în timp a parametrilor statici obţinuţi în urma analizei în frecvenţă pe termen scurt *(parametrii statici nu contin informatii temporale)*

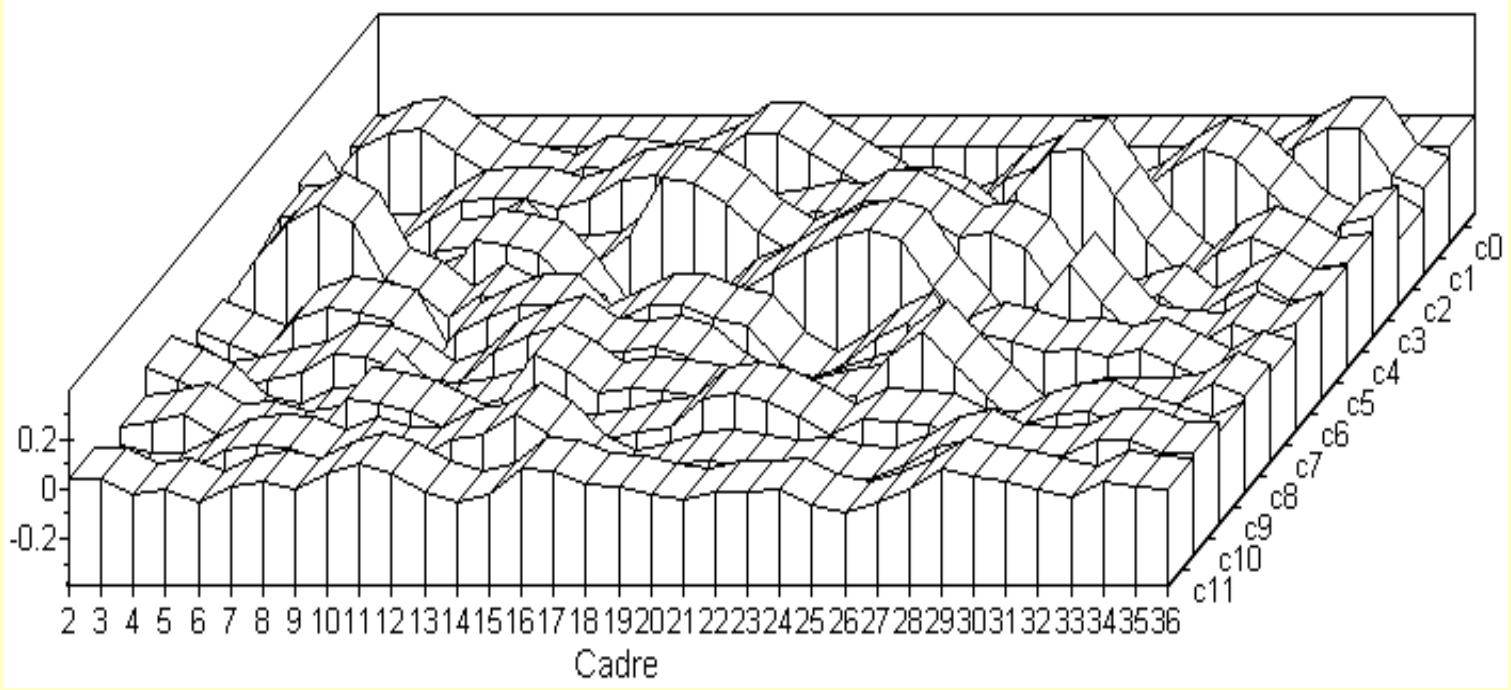
• modificările în timp ale acestor parametri joacă un rol important în percepţia umană.

• se obţin rezultate mai bune în recunoaştere cu până la 20% [Hua01].

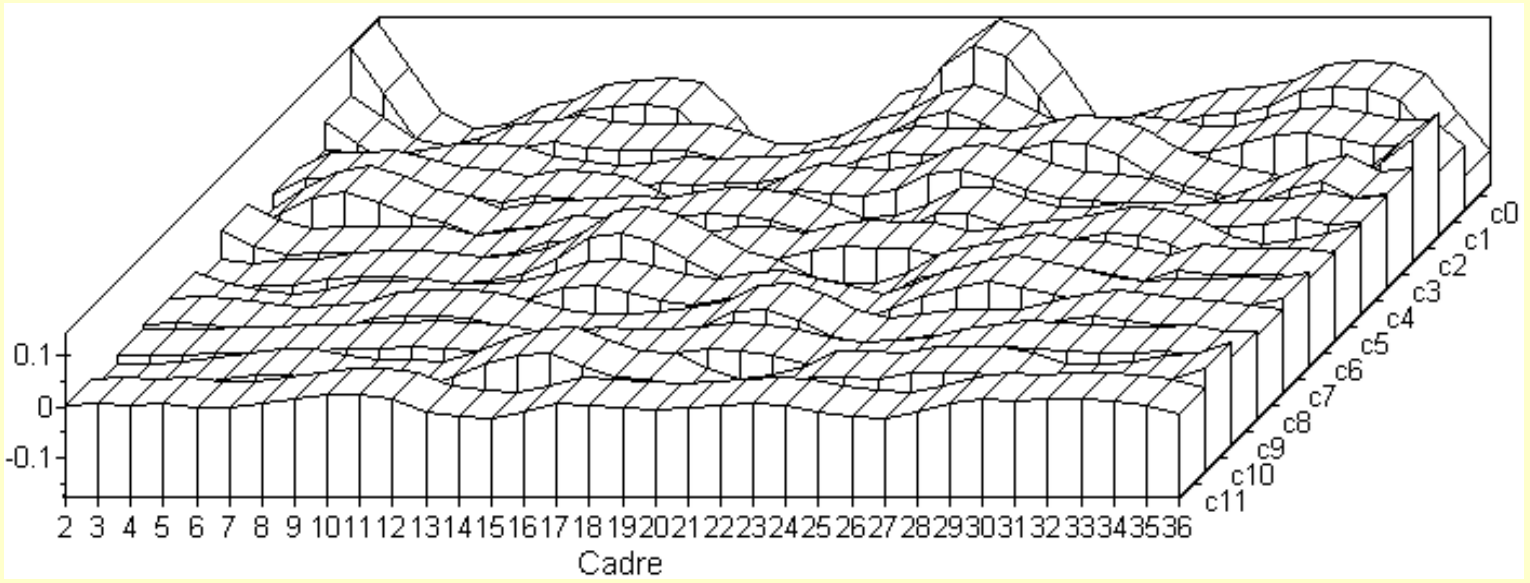
• Cel mai simplu mod de a obţine această informaţie este de a face diferenţa între coeficienţii din cadre consecutive

$$\Delta_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2}$$

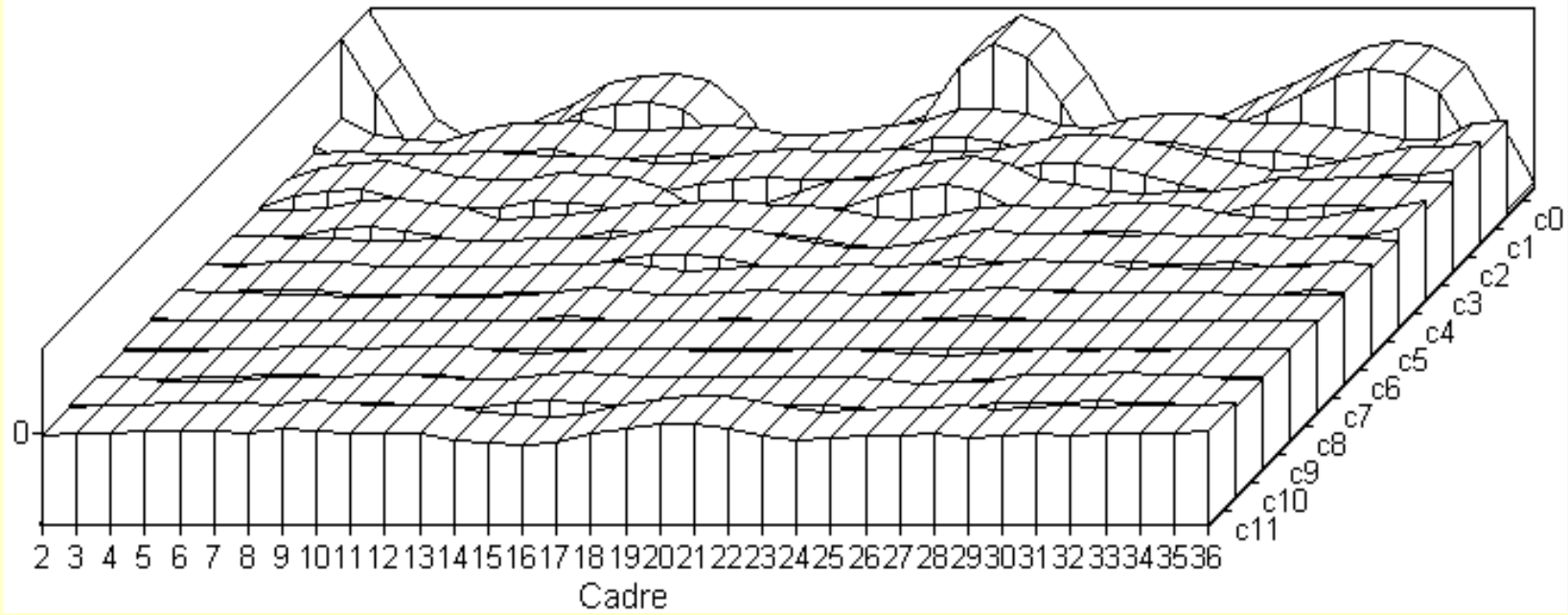
- Δ_t sunt parametri delta, iar Θ dă nr. de cadre pe care se face calculul.
- Parametri delta se calculează prin interpolarea parametrilor statici din cateva cadre consecutive (Θ = 2, tipic), ce acoperă un interval de ~ 50-100ms.



Evoluţia coeficienţilor delta LPCC (p=12) pentru cuvântul “unu”.



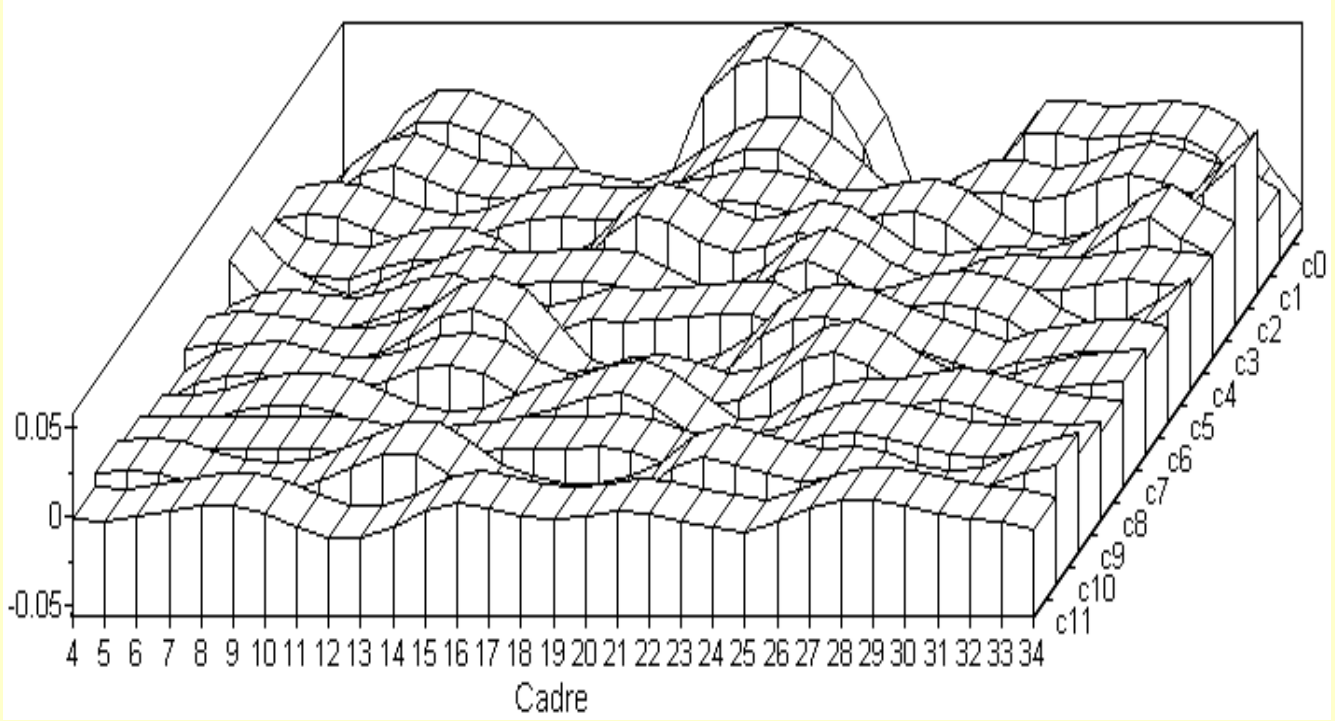
Evoluţia coeficienţilor delta PLP (p=12) pentru cuvântul “unu”.



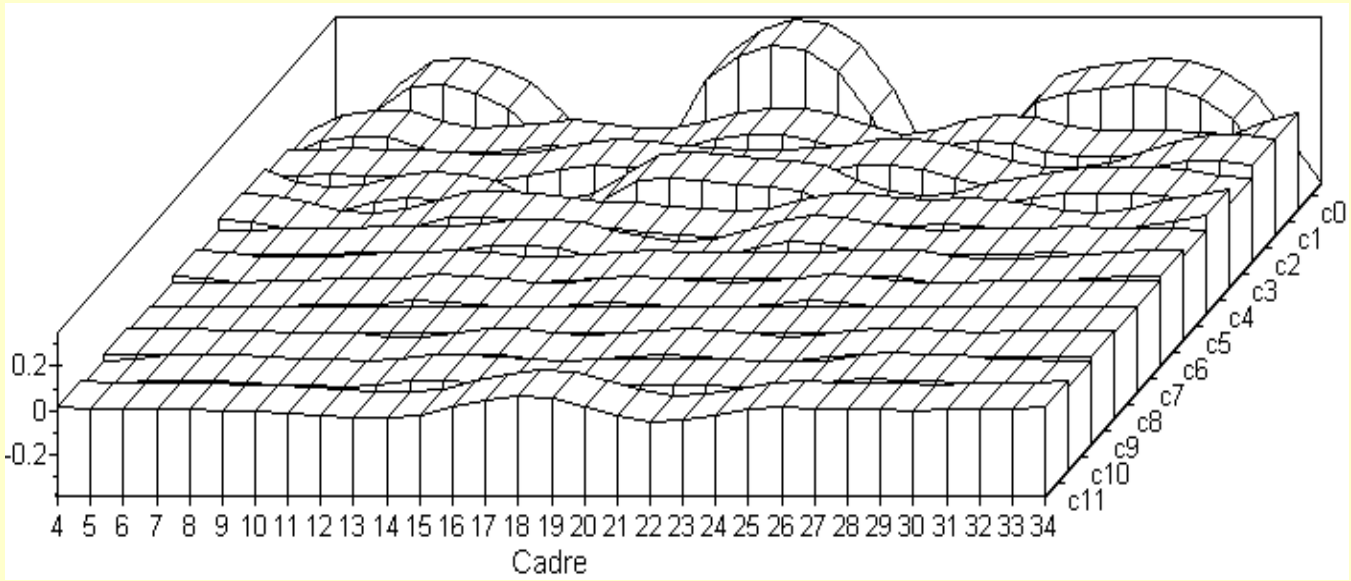
Evoluţia coeficienţilor delta MFCC (p=12) pentru cuvântul “unu”.

Dacă se aplică aceeaşi formulă pentru parametri delta se obţine acceleraţia sau parametri delta-delta. [Her99] :

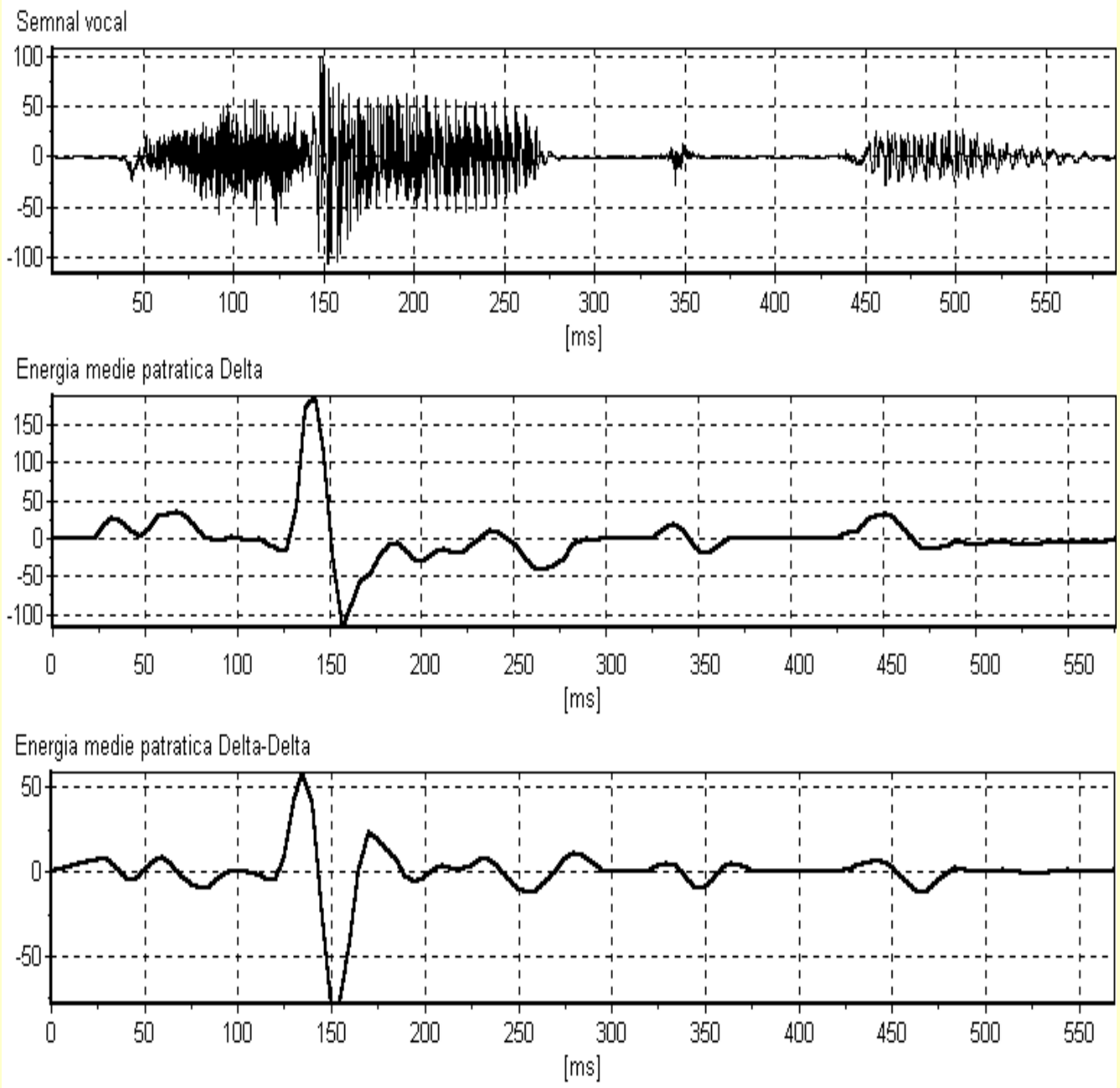
$$\Delta\Delta_t = \frac{\sum_{\theta=1}^{\Theta} \theta(\Delta_{t+\theta} - \Delta_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2}$$



Evoluţia coeficienţilor delta-delta PLP (p=12) pentru cuvântul “unu”.



Evoluţia coeficienţilor delta-delta MFCC (p=12) pentru cuvântul “unu”.



Evoluţia coeficienţilor delta şi delta-delta pentru energia medie pătratică.

Vector acustic/caracteristic pentru recunoasterea automata a vorbirii

A single feature vector,

- 12 cepstral coefficients (PLP, MFCC, …) + 1 norm energy
- + 13 delta features
- + 13 delta-delta

Frame energy is a typical feature used in speech recognition. Frame energy is computed from the windowed frame,

$$e[t] = \sum_{m} s^2(n)$$

Typically a normalized log energy is used. E.g.,

$$e_{\max} = \arg\max_{t} \{0.1 \cdot \log(e[t])\}$$

$$E[t] = \arg\max \{-5.0, 0.1 \cdot \log(e[t]) - e_{\max} + 1.0\}$$