

ASRSV – curs 4

Analiza Semnalului Vocal (2)

http://liceu.uab.es/~joaquim/phonetics/fon_anal_acus/herram_anal_acus.html

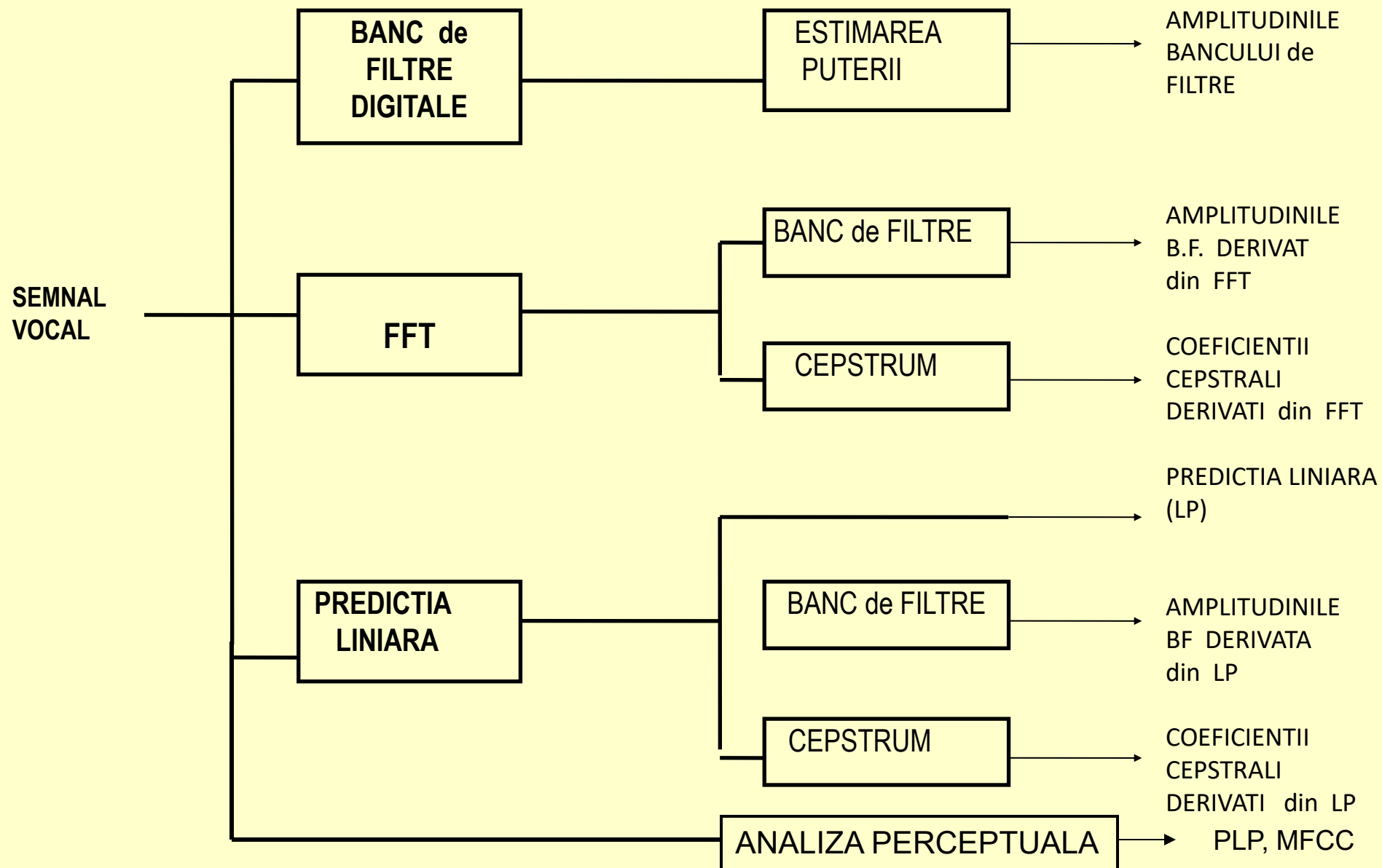
ANALIZA SV

In domeniul timp:

- Amplitudinea medie/maxima
- Densitatea de amplitudine
- Energia medie
- Energia TEAGER
- Numarul trecerilor prin zero
- Frecventa fundamentala (F0)
- Codarea TESPAP

In domeniul frecventa:

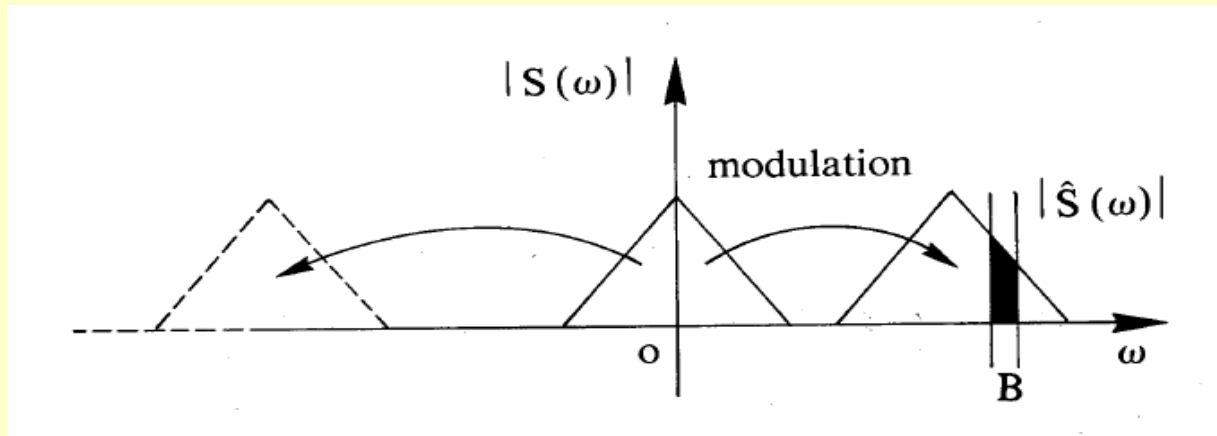
- TFD (TFR)
- Banc de filtre
- Analiza homomorfica
- Analiza LPC
- Analiza perceptuala



Algoritmi folosiți în analiza spectrală [Picone]

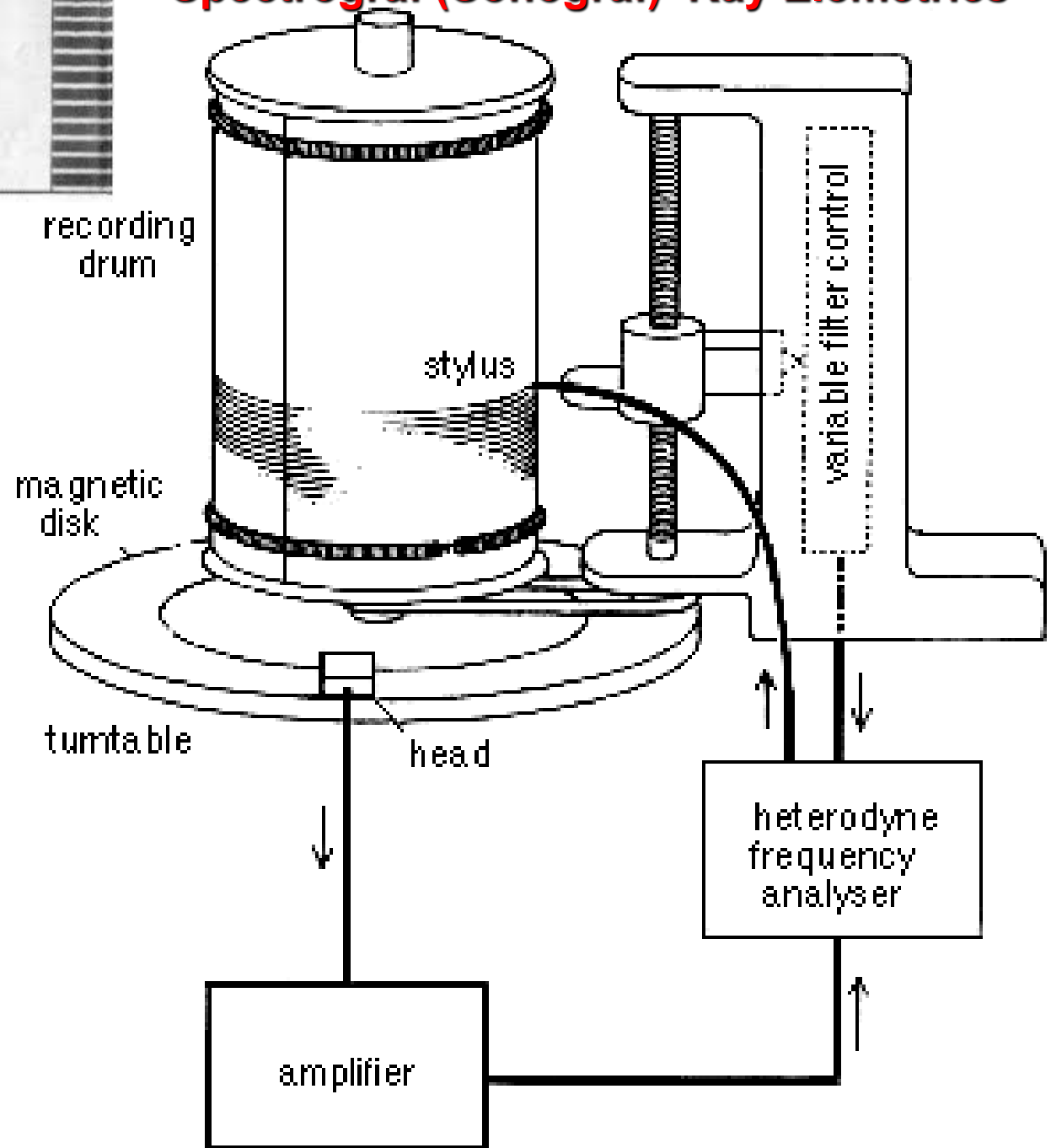
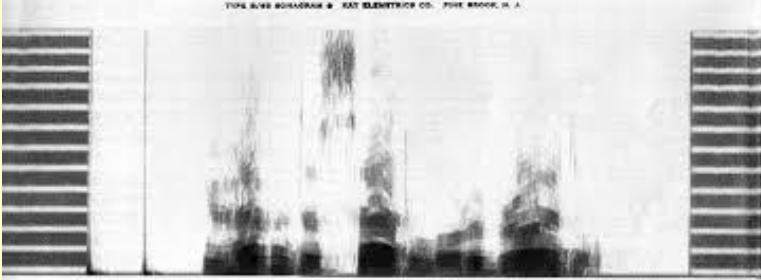
1. SPECTROGRAMA (reprezentare timp-frecventa-amplitudine)

- Evolutia functiei spectrale a SV in timp – spectrograful
- Plotarea spectrogramei poate fi : de contur sau luminos/intunecat;
- Primul instrument folosit de foneticieni – Key Elemetrics;
- Compus din : modulator, filtru, tambur de plotare sonograma (f/t)
- Principiul este filtrajul superheterodina, daca SV este $s(t)$, semnalul modulat este $\hat{s}(t) = s(t) * \cos 2\pi f t$, iar spectrul este deplasat spre frecvente mai inalte si baleiaza intrarea unui FTB;

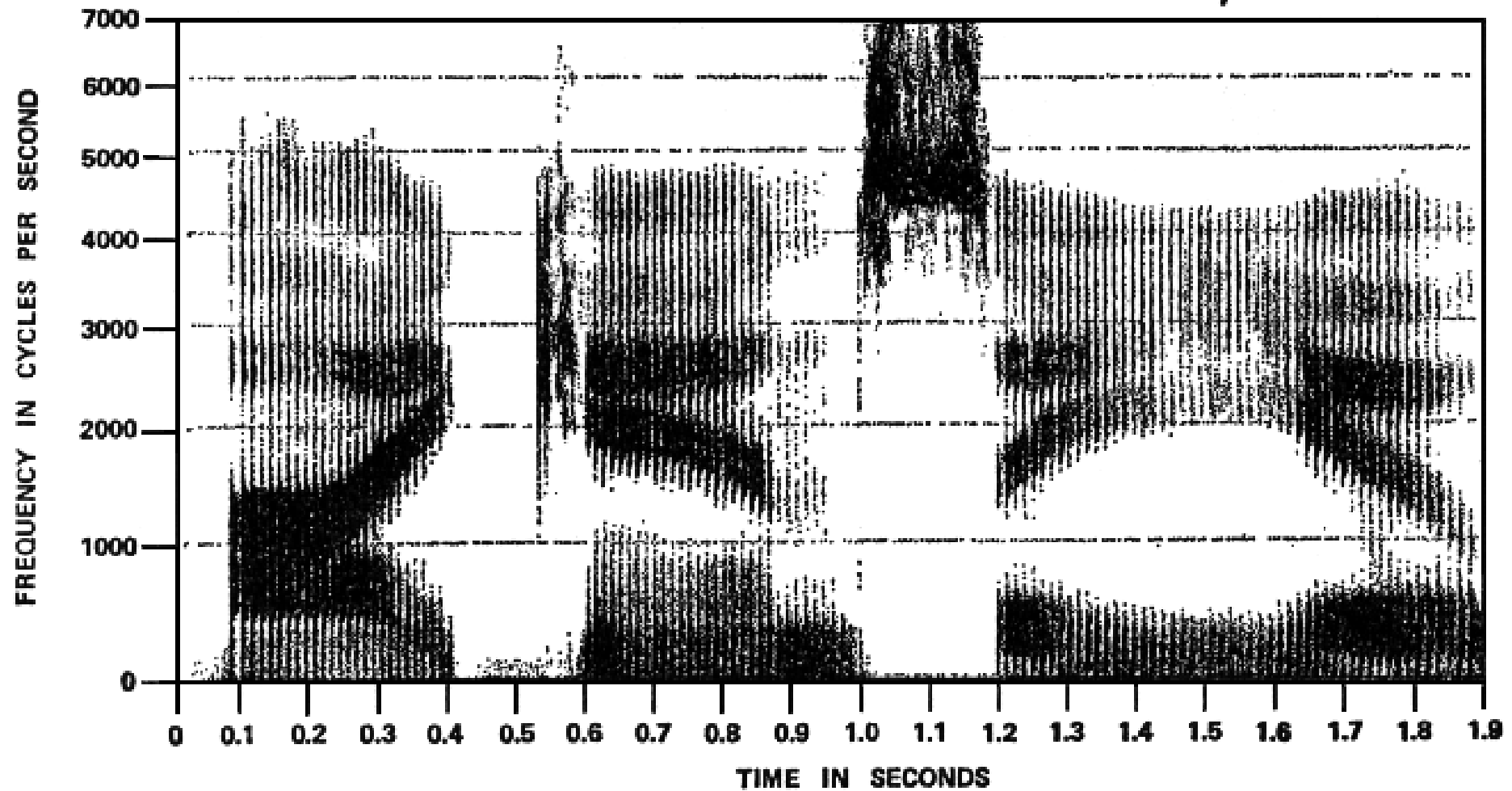


- Energia in aceasta banda este obtinuta printr-o redresare-integrare;
- Se modifica frecventa purtatoarei pentru a baleia tot spectrul semnalului in banda filtrului;
- $B=300\text{Hz}$ – *spectrograma de banda larga* - care subliniază *modificările temporale* ale semnalului
- $B=45\text{Hz}$ – *spectrograma de banda ingusta* - subliniază în semnal *modificările de frecvență*

Spectrograf (Sonograf) Kay Elemetrics

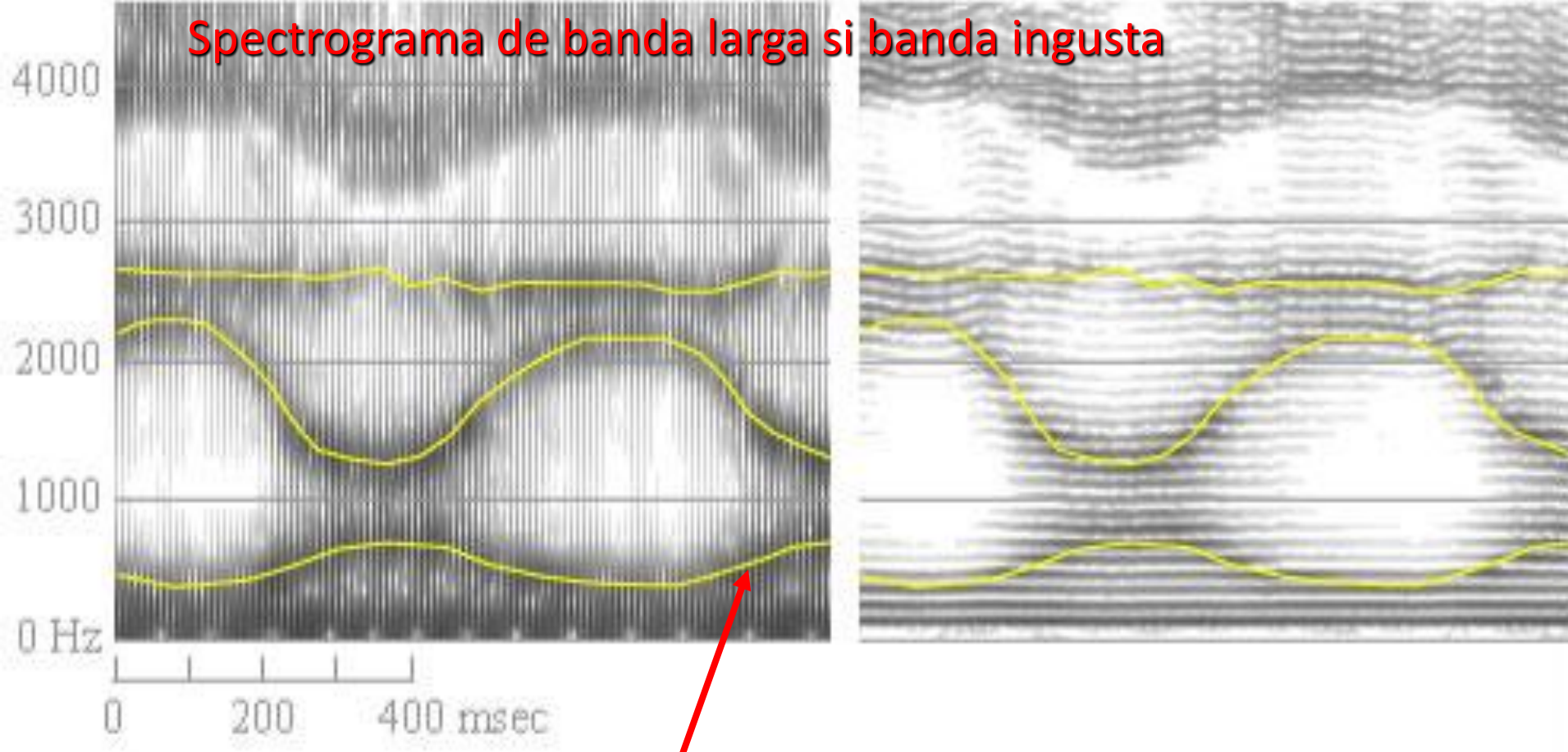


I c a n s e e y ou



Spectrogram (sonograma)

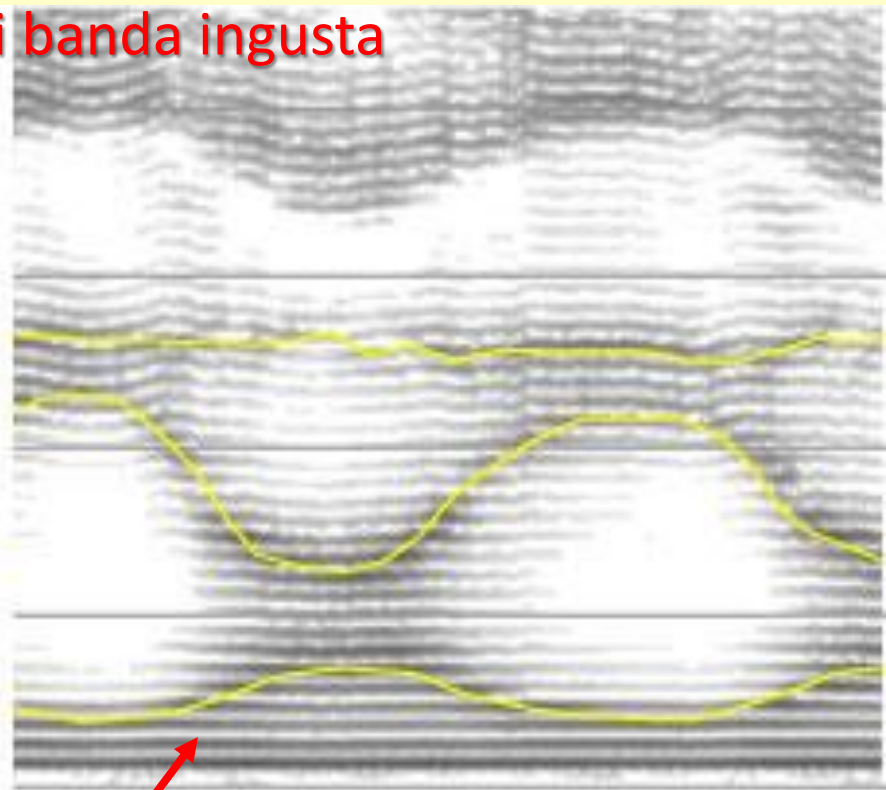
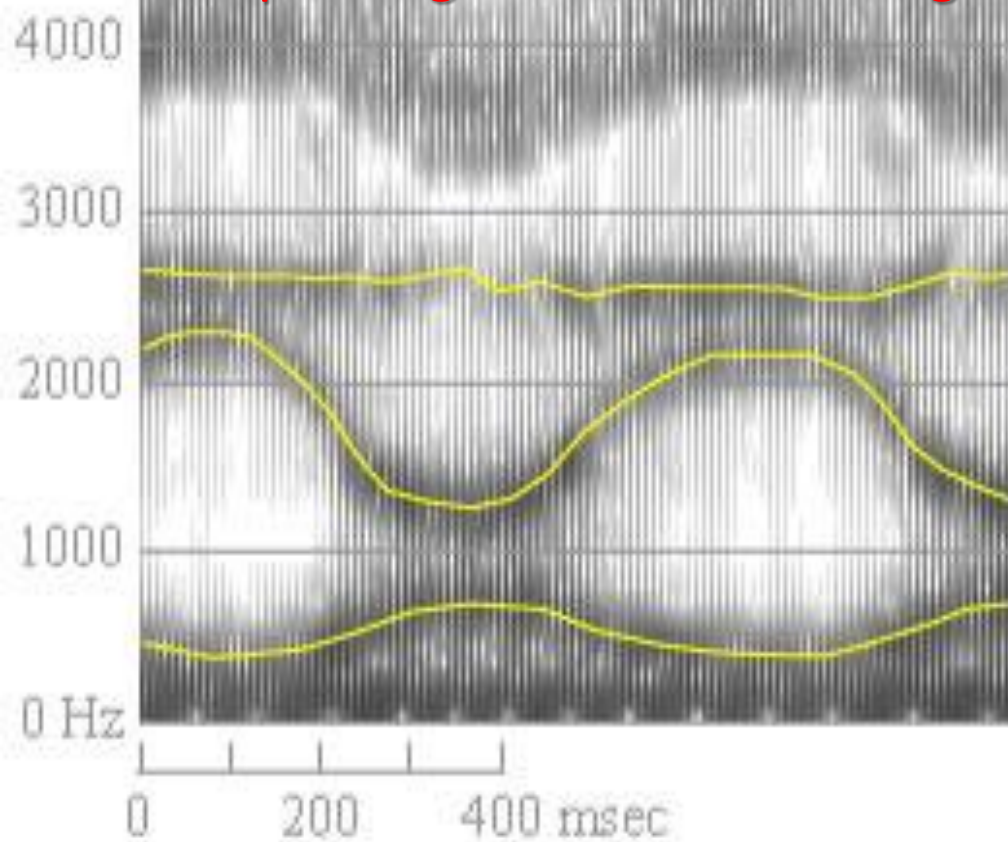
Spectrograma de banda larga si banda ingusta



Spectrograma de bandă largă (~3ms)

- urmează vârfuri spectrale largi (formanti) in timp
- evidentiaza cele mai multe perioade individuale ale F0 ca striuri verticale, deoarece banda filtrului de analiza este comparabilă în timp cu o perioada a FF (3ms/300Hz)
- pentru vorbirea nesonora nu există striuri de pitch verticale

Spectrograma de banda larga si banda ingusta



Spectrograma de bandă îngustă (~20ms)

- armonicile sunt evidentiuate in regiunile sonore
- frecvențele formantilor sunt încă evidentiuate
- de obicei, se poate vedea si frecvența fundamentală
- regiunile nesonore nu arată o structura bine conturata

2. Analiza prin TFR

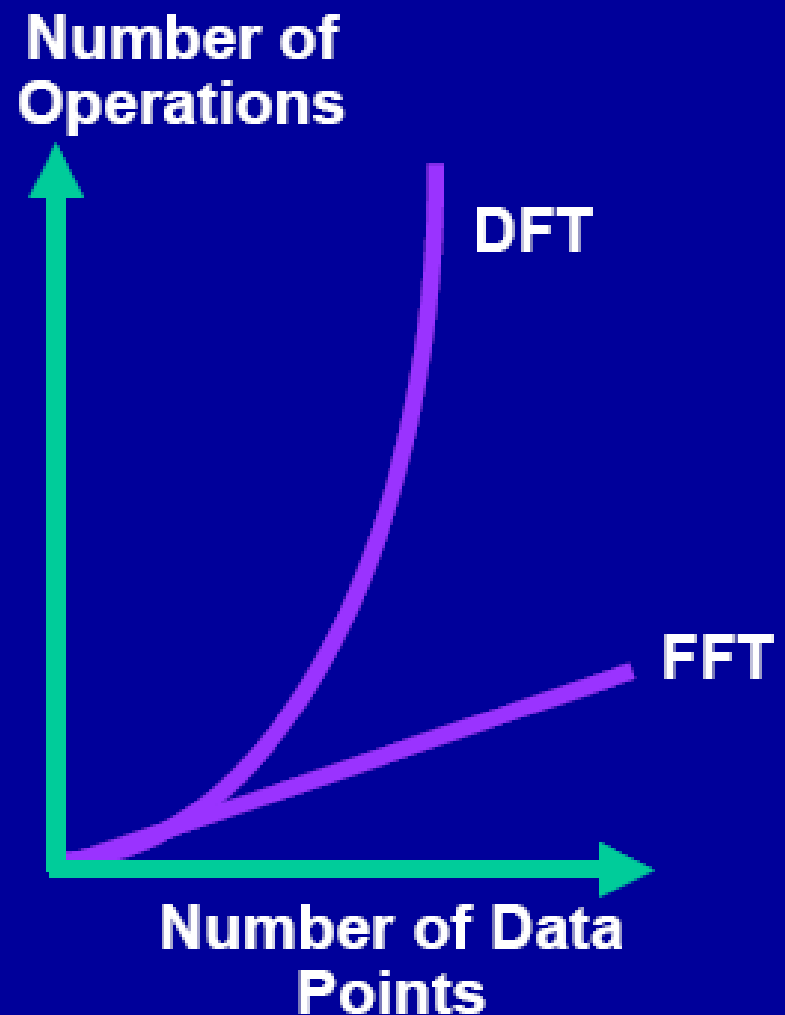
- reprezintă semnalul ca suma de exponentiale complexe si aceasta conduce la soluții convenabile pentru diferite probleme (estimare formanti, estimare F0, și de analiza semnalului în sine)
- reprezentările Fourier furnizeaza :
- Mijloace convenabile pentru a determina răspunsul sistemelor liniare la o sumă de sinusoid
- Evidențierea clara a proprietăților semnalului care sunt ascunse în semnalul original
- Algoritmul FFT (Cooley-Tukey 1965, Gauss) – prelucrari de semnal in timp real (reduce operatiile de la $N^2 > N \cdot \log_2 N$; ex. $10^6 \gg 10^4$)

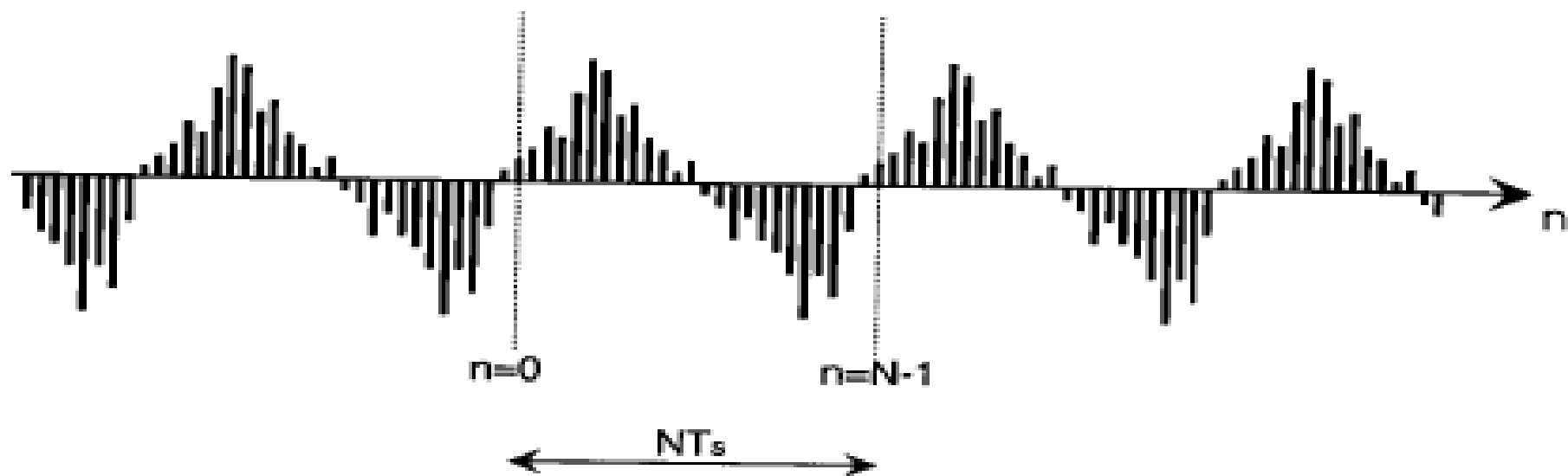
$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n) e^{-j\omega n} = \mathcal{DTFT} \{x(n)\}$$

$$x(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) e^{j\omega n} d\omega = \mathcal{DTFT}^{-1} \{X(e^{j\omega})\}$$

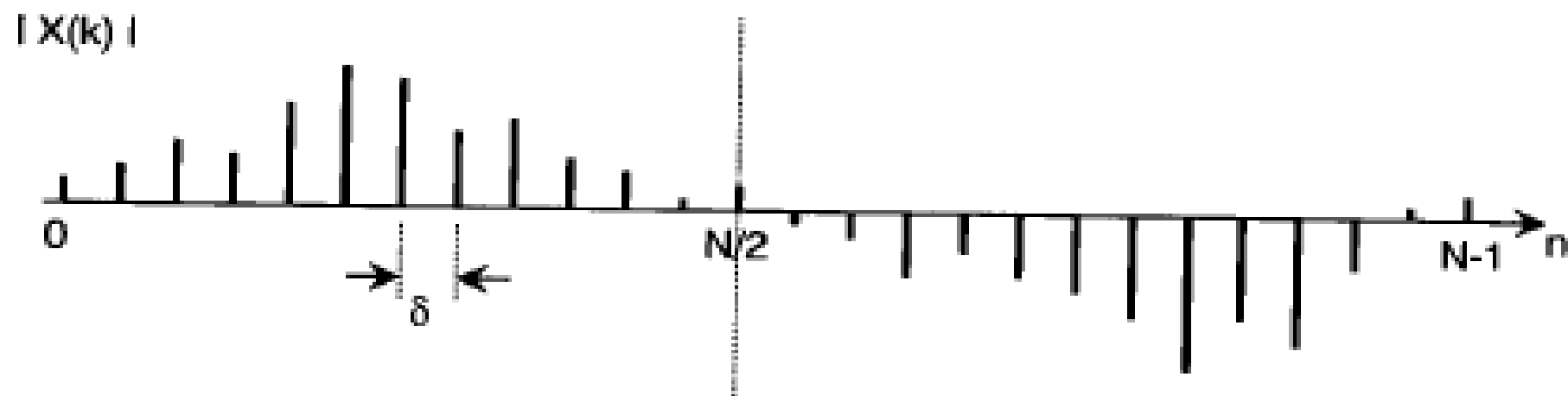
Fast Fourier Transform

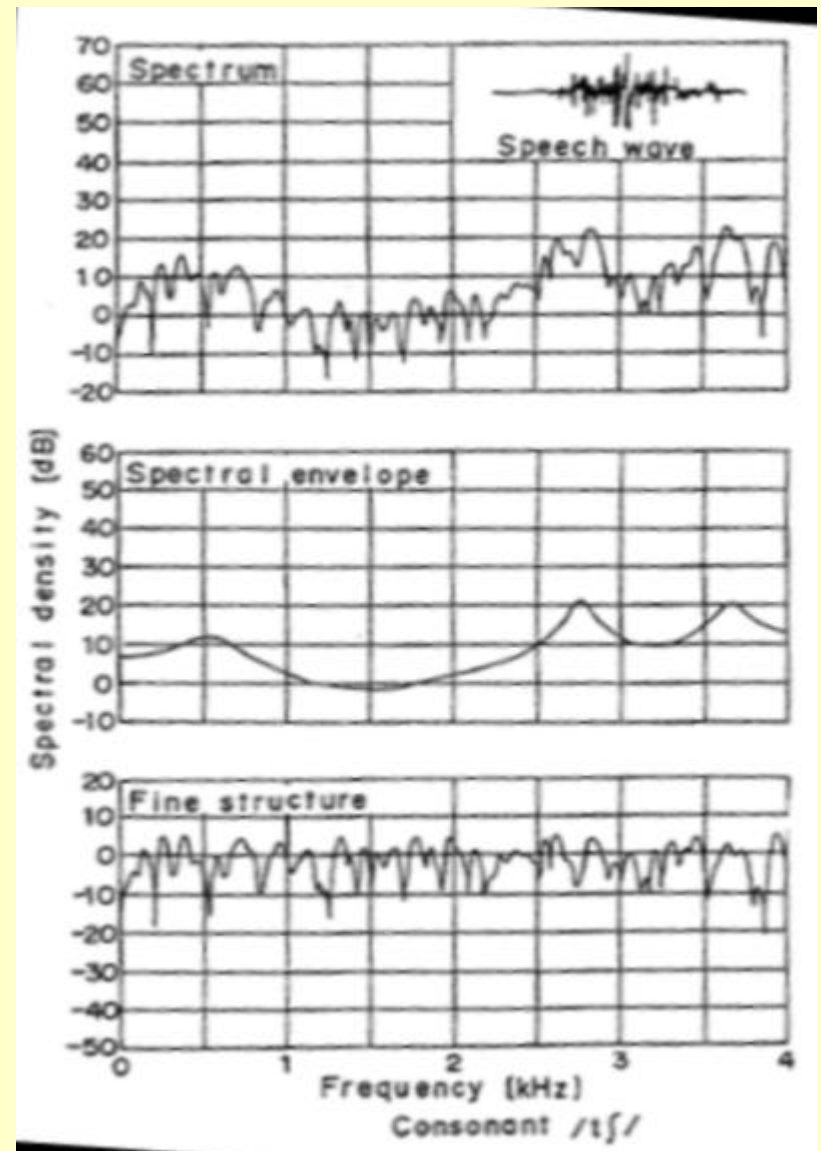
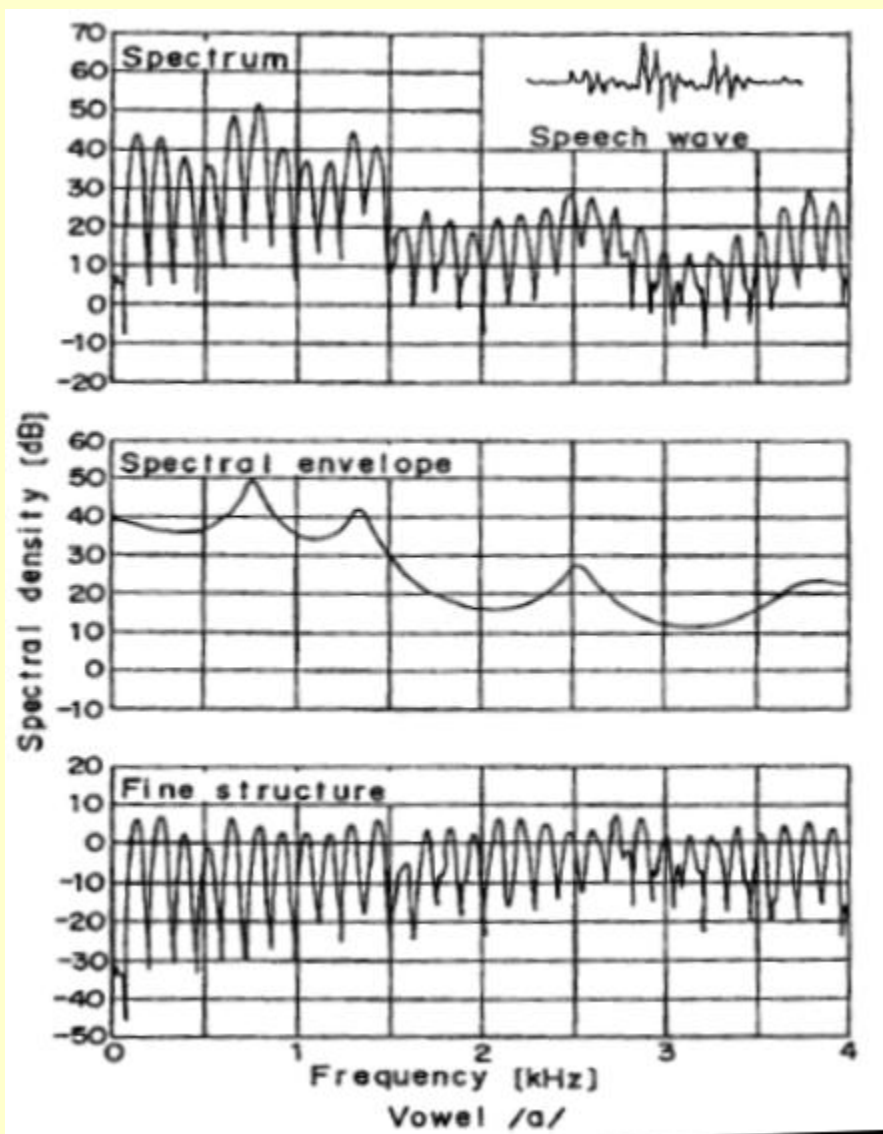
N	DFT	FFT
16	256	64
32	1,024	160
64	4,096	384
128	16,384	896
256	65,536	2,048
512	262,144	4,608
1,024	1,048,576	10,240
2,048	4,194,304	22,528
4,096	16,777,216	49,152





$$N\delta = \omega_s$$





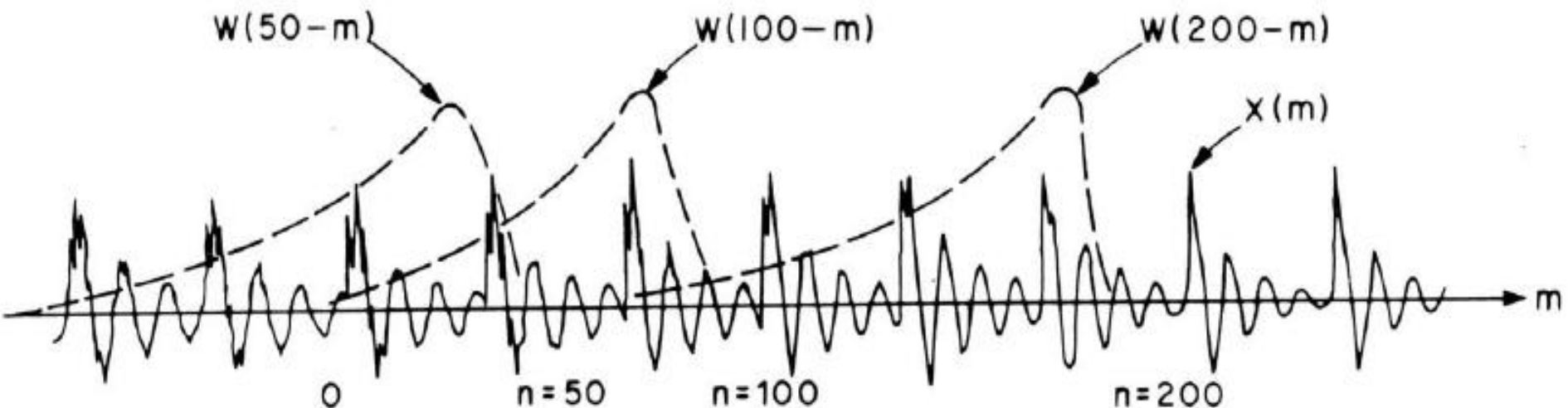
- **Spectrul de putere pe timp scurt** este compus din **anvelopa spectrala globala** (se modifica lent cu tractul-rezonante/antirezonante, forma spectrului sursei si caract. de radiatie la nivel buze-nazal) si **structura fina** (rapid- periodicitatea sursei glotale)

TFTS (STFT) – TF pe timp scurt

- TF nu poate face localizare simultana si in timp si in frecventa, nu este ft. utila ptr. analiza semnalelor variabile in timp, nestationare

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m) w(n-m) e^{-j\omega m}$$

- unde $w(n-m)$ este o fereastră care determina portiunea din $x(n)$ care se foloseste la calculul $X_n(e^{j\omega})$



Fereastra larga → rezolutie buna in frecventa si slaba in timp

$W(t)=1$, infinit de lunga: → STFT se transformă în FT, oferind o localizare excelentă a frecvenței, dar fără localizare în timp.

Fereastra ingusta → rezolutie buna in timp si slaba in frecventa

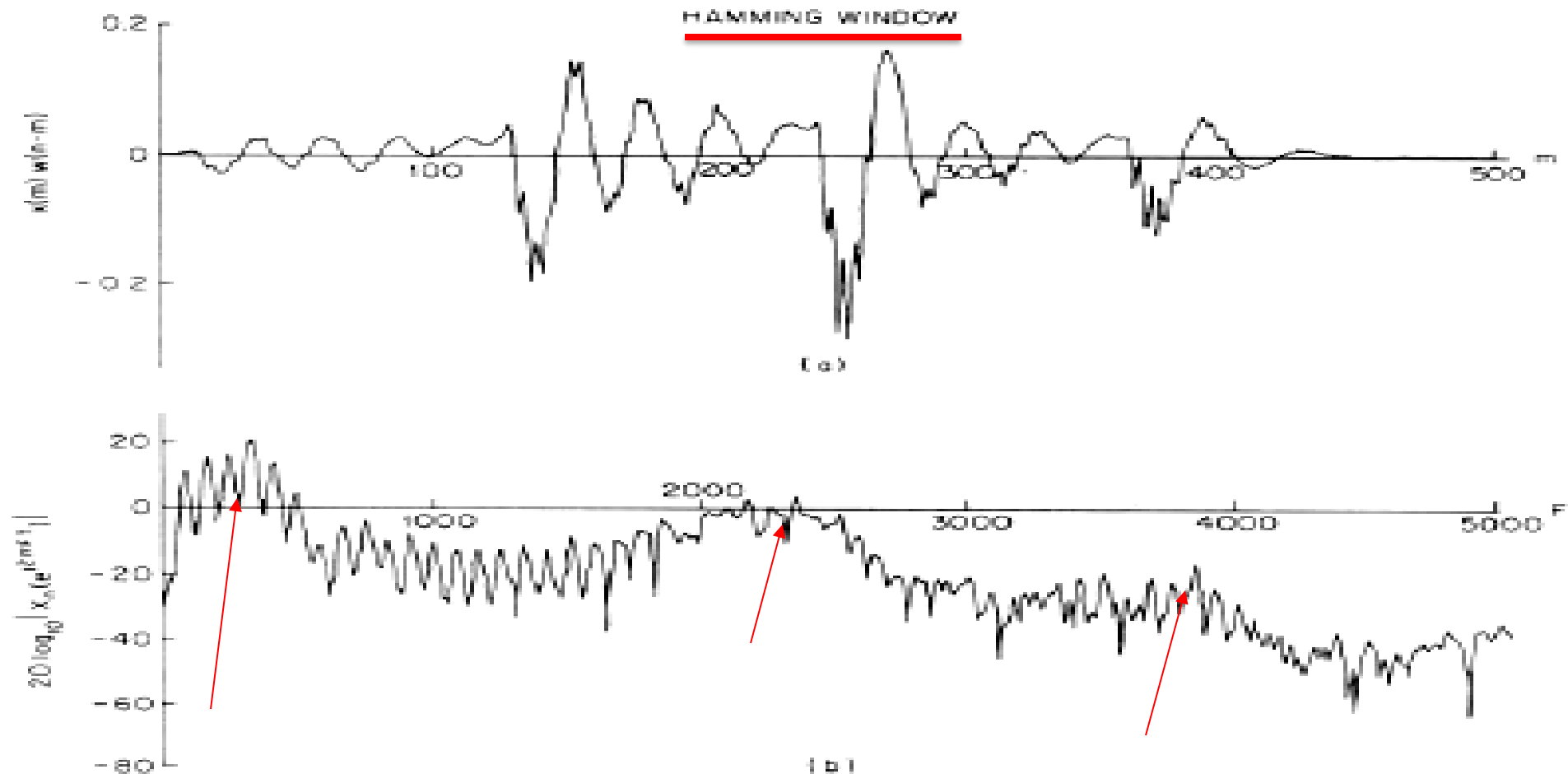
$W(t)=\delta(t)$, infinit de scurta: → are ca rezultat semnalul in timp (cu un factor de fază), oferind o localizare excelentă a timpului, dar fără localizare de frecvență.

$$STFT_f^u(t', u) = \int_t [f(t) \cdot \delta(t - t')] \cdot e^{-j2\pi ut} dt = f(t') \cdot e^{-jut'}$$

Δt si Δf nu pot fi facute arbitrare de mici!!

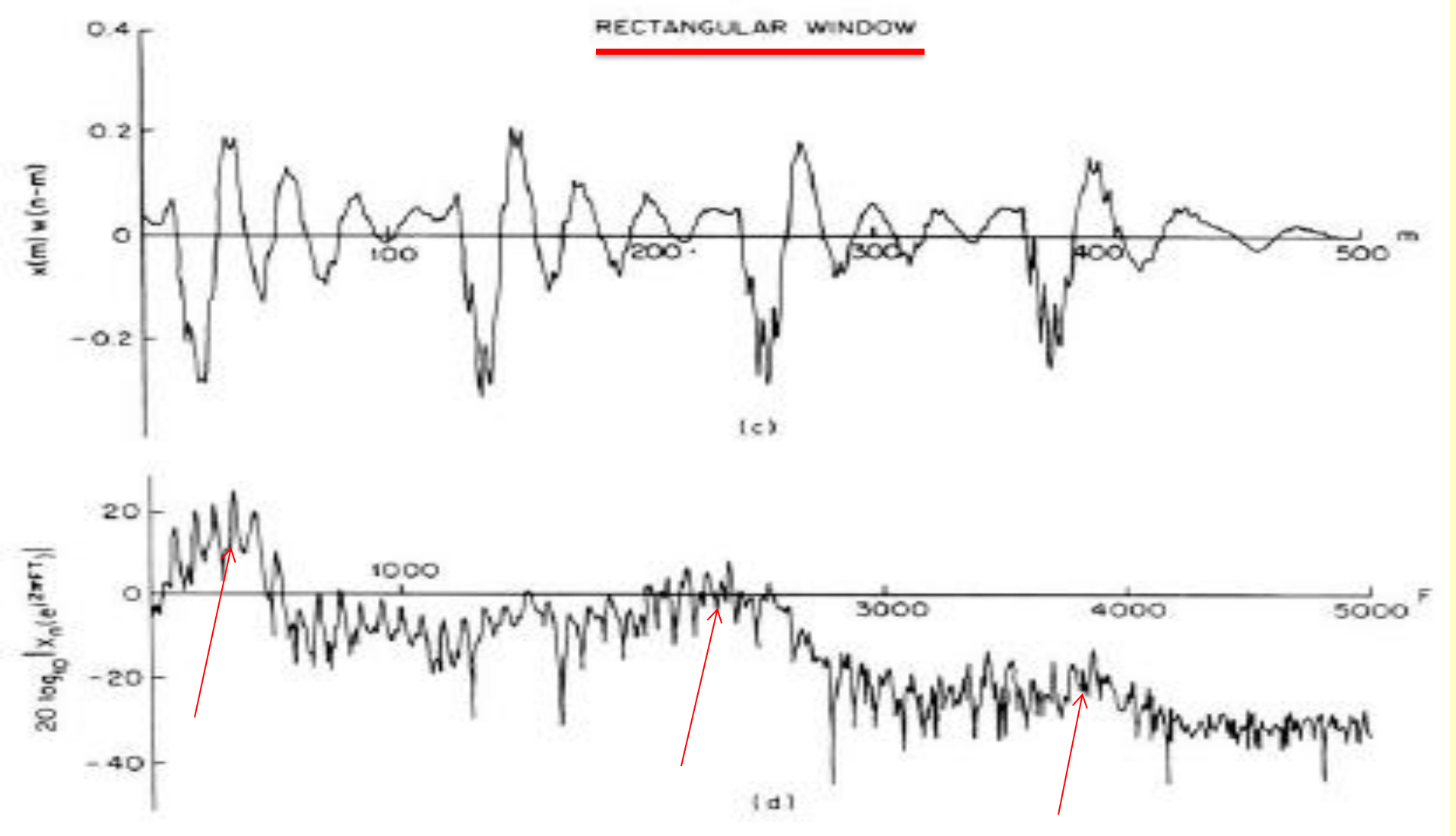
- Principiul de incertitudine a lui Heisenberg

$$\Delta t \cdot \Delta f \geq \frac{1}{4\pi}$$

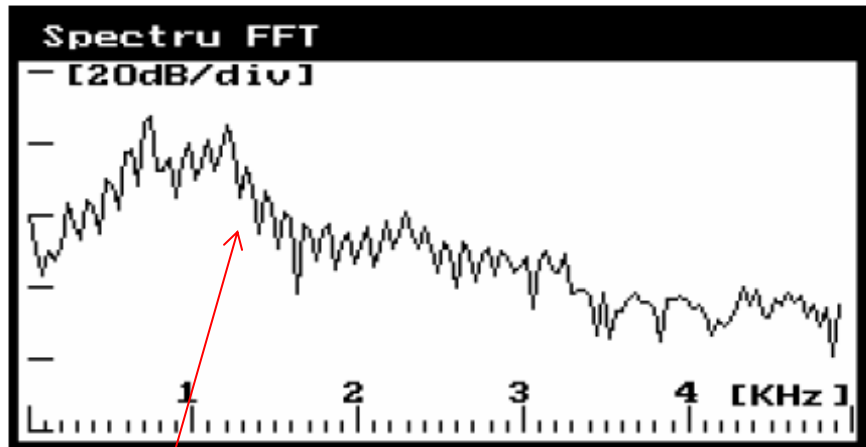
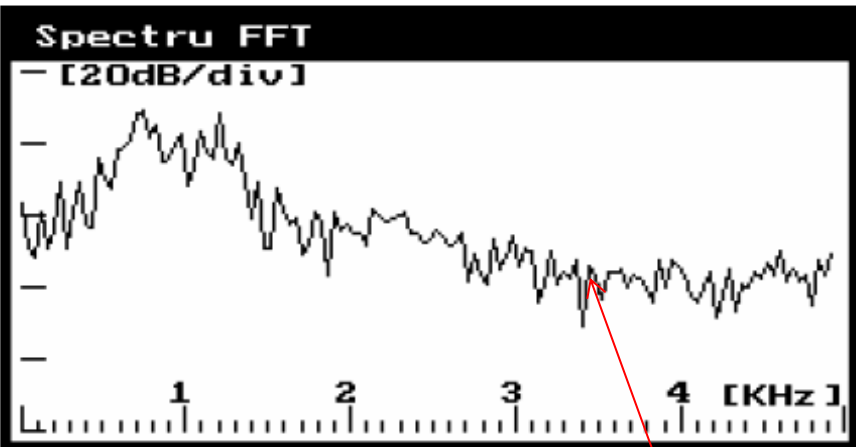
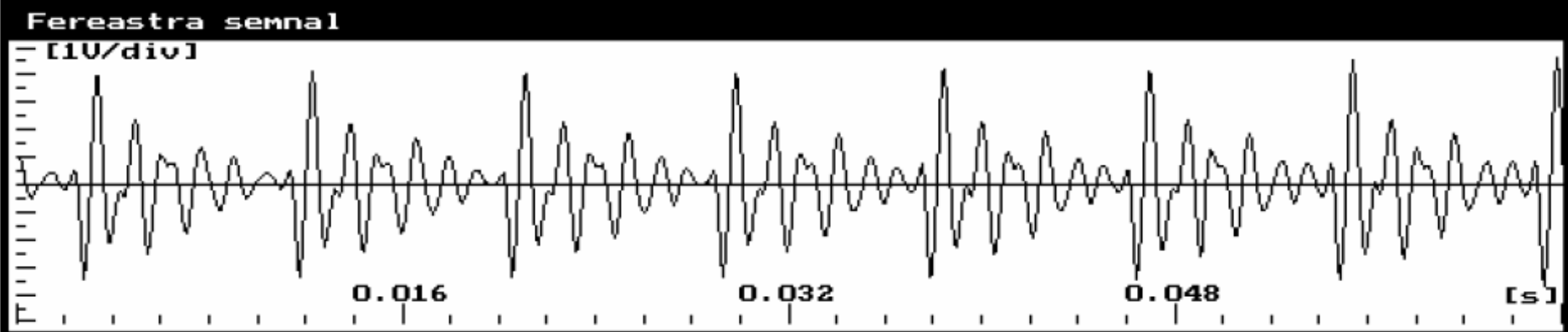


Spectrum analysis for voiced speech using a 50 msec (a,b) Hamming window; (c,d) rectangular window. Parts (a) and (c) show time waveforms; parts (b) and (d) show corresponding spectra.

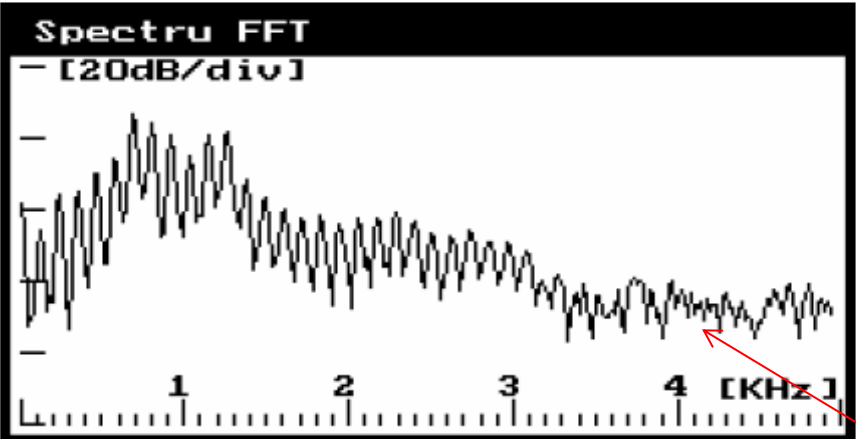
- 500 de esantioane/ferestra (50 msec)
- periodicitatea se poate vedea în timp și în frecvență
- se poate vedea primul formant (350 Hz), II^a rezonanță la 2200 Hz, III^a rezonanță la 3800 Hz



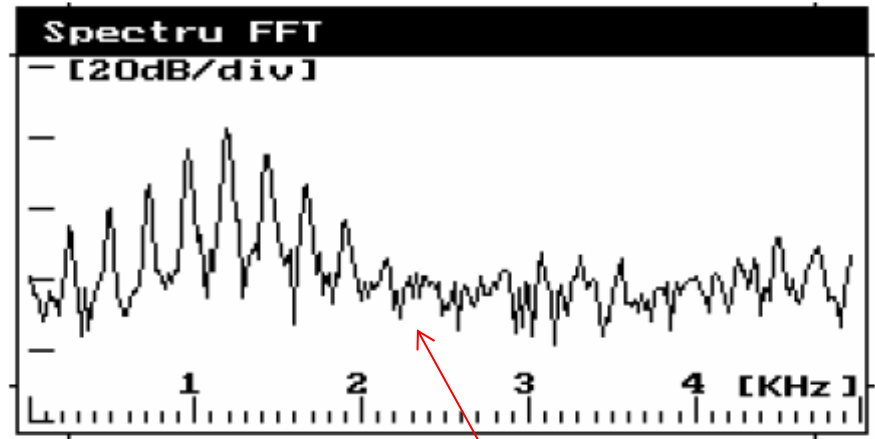
- caderea spectrului de frecvențe, datorită formei impulsurilor glotale
- o claritate mai bună a armonicilor FF din cauza lobului principal mai îngust al ferestrei rectangulare (FR)
- spectrul de frecvențe mai zgomotos (ptr. FR), datorită interferenței inter-armonice, deoarece lobii laterali ai RW au doar -14 dB atenuare



Segment din vocala “a” și spectrul său Fourier (N=256) cu fereastră dreptunghiulară (a) și Hamming (b)



(a)



(b)

Fig. 3.5.2. Spectrul Fourier al vocalei "a" folosind fereastră Hamming pe 512 puncte pentru locutor masculin(a) și feminin (b)

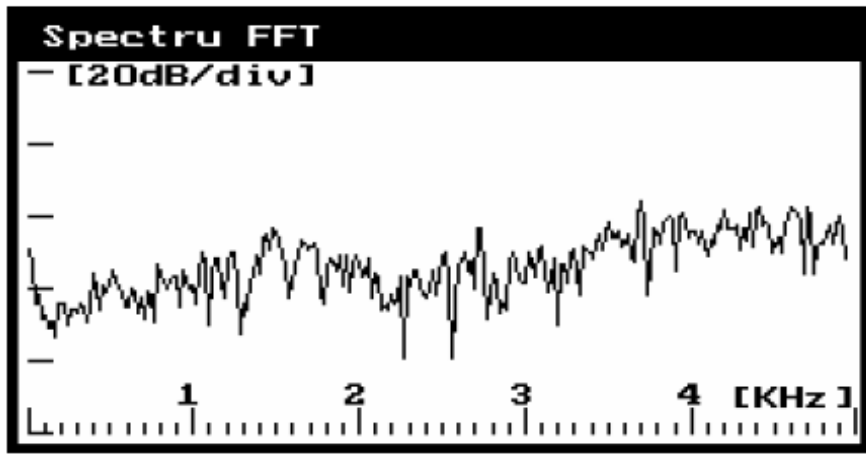
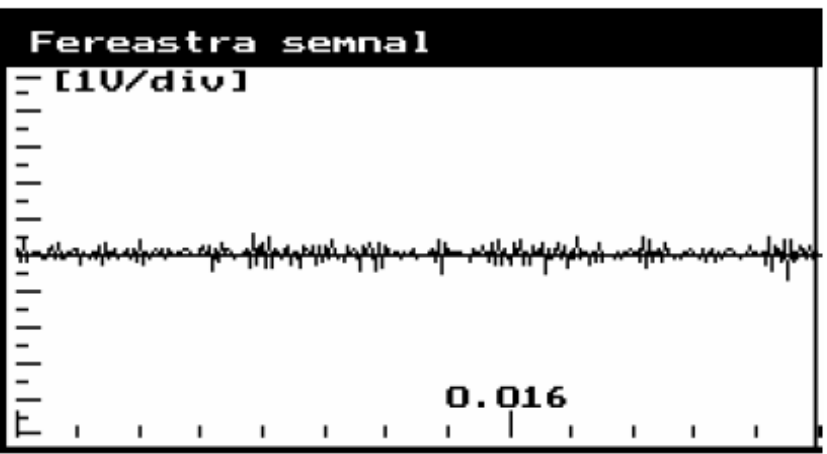


Fig.3.5.3. Segment nesonor din consoana "s" și spectrul său Fourier (N=512)

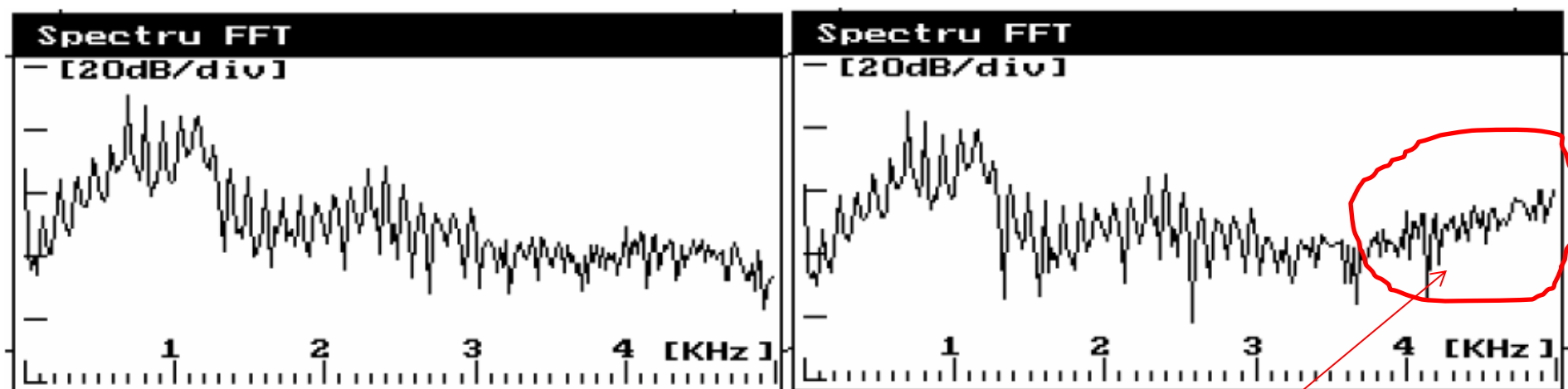
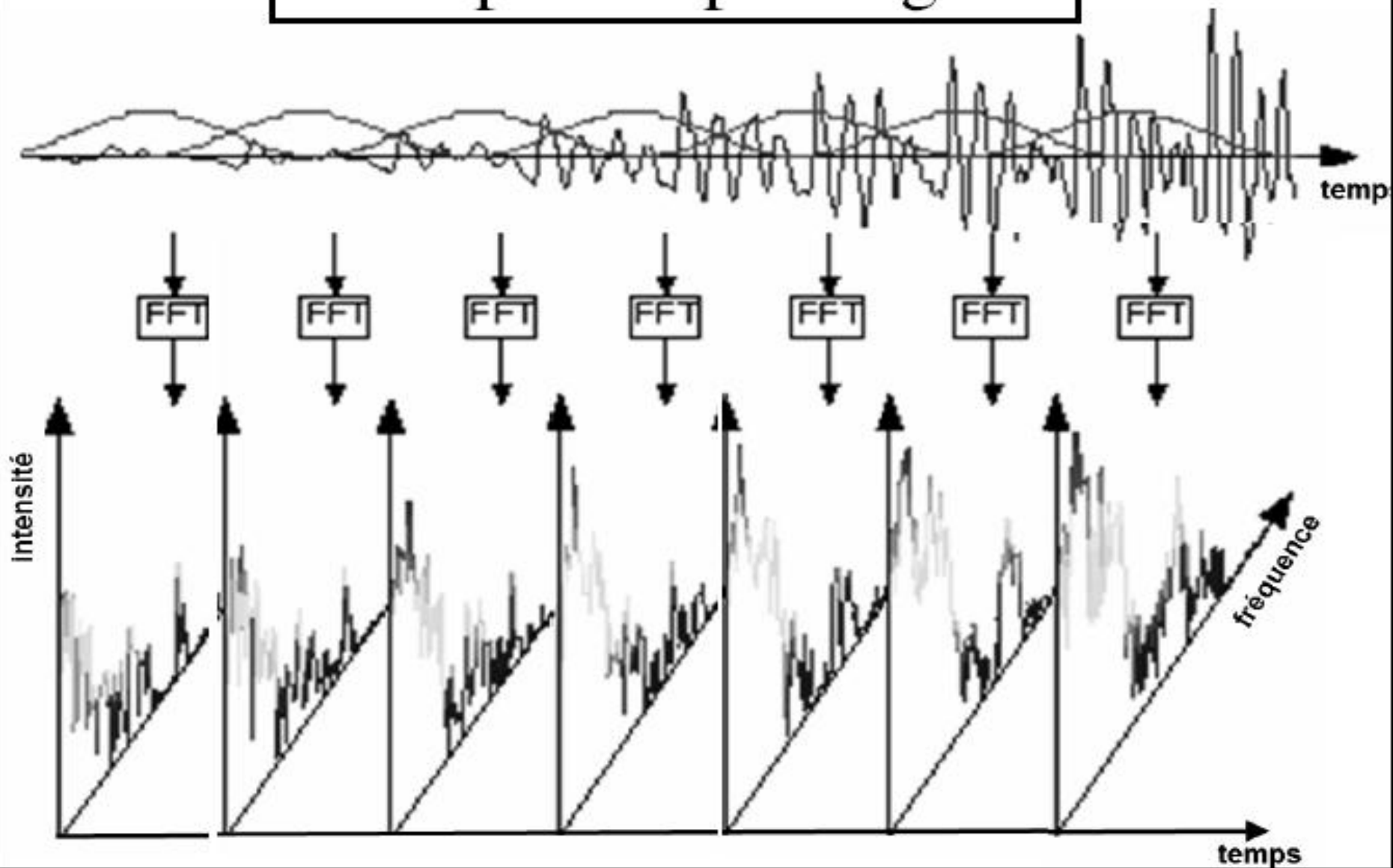


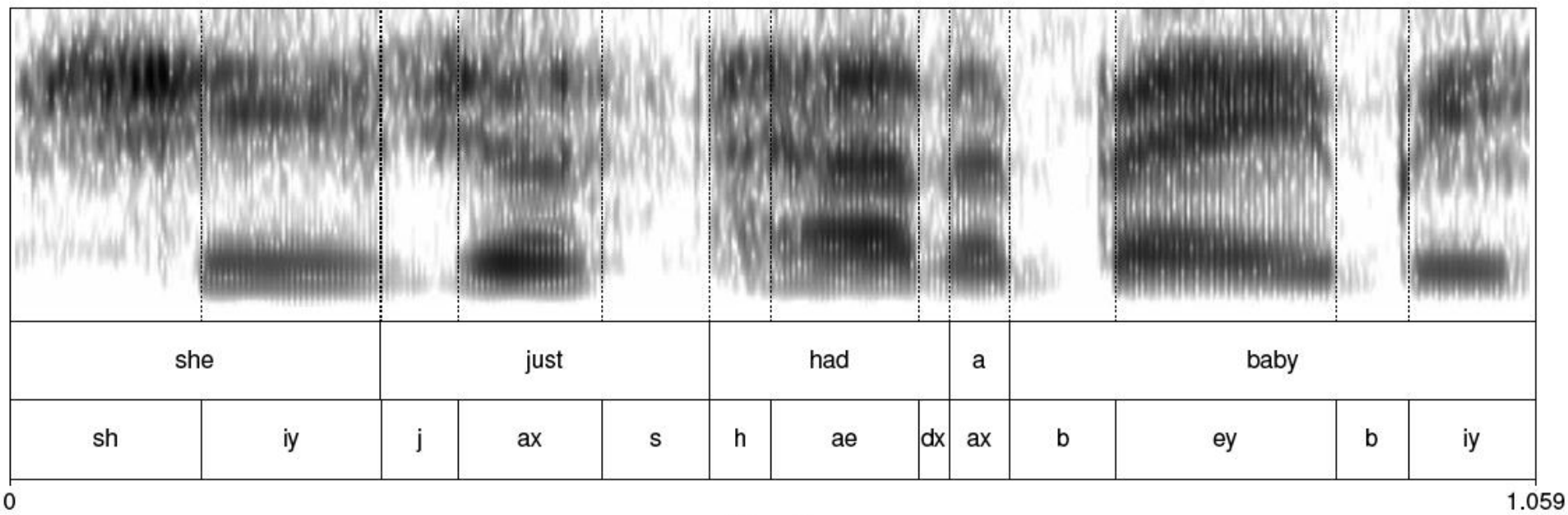
Fig.3.5.11. Spectrul Fourier al unui segment de 512 puncte din vocala "a" fără și cu preaccentuare

Principle of spectrogram



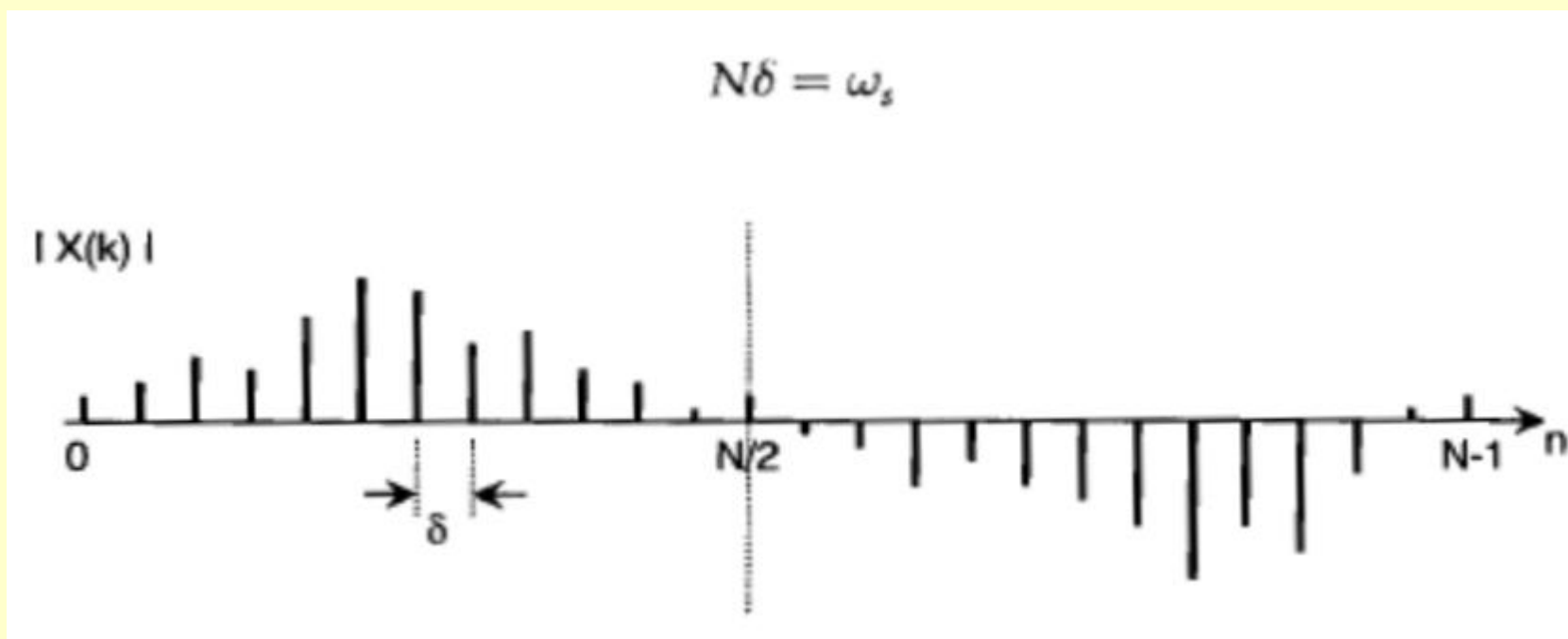


F[Hz]



t[s]

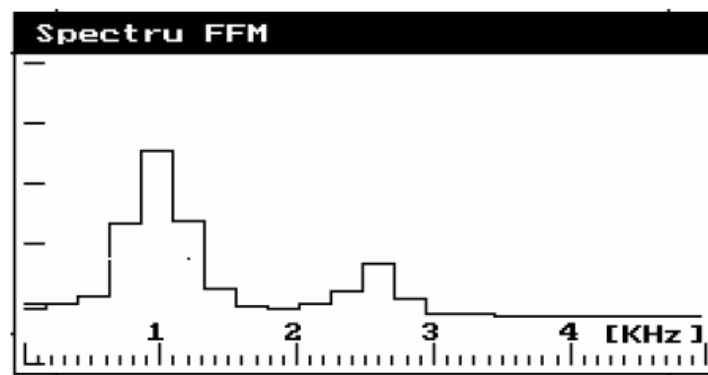
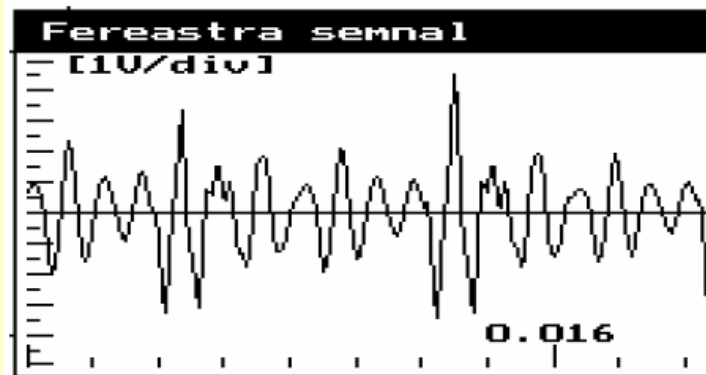
3. Banc de filtre derivat din TFR



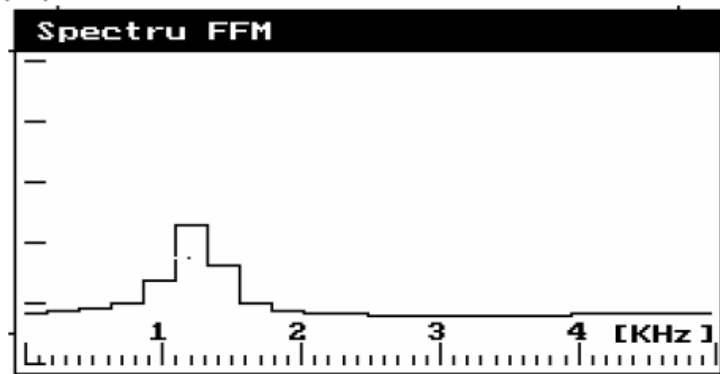
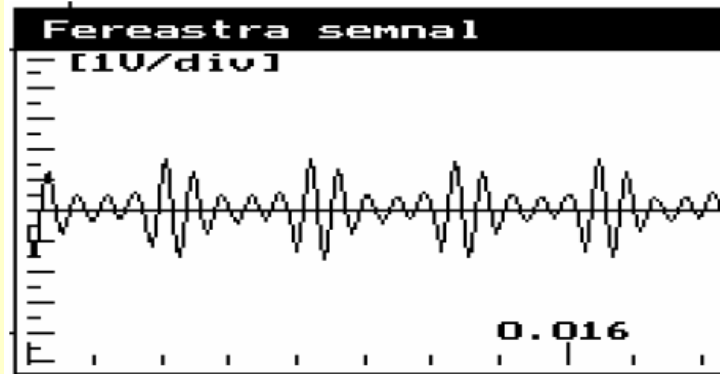
banda 1 (0-200Hz)
benzile 2-13 (200-3kHz)
benzile 14-17 (3-5kHz)

4 coeficienți Fourier
câte 6 coeficienți Fourier/bandă
câte 13 coeficienți Fourier/bandă.

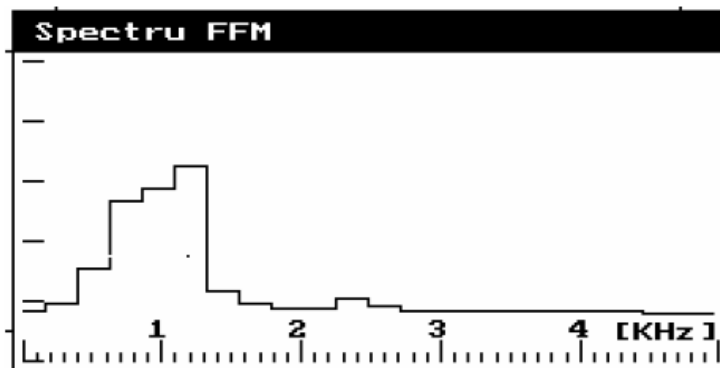
$N=256, f_s=10\text{kHz} \Rightarrow \delta \sim 40\text{Hz}$



V2(m)



V3(f)



V4(m)

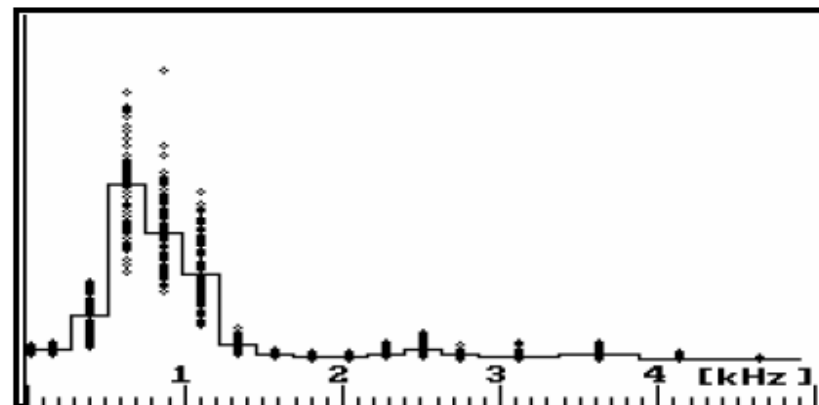
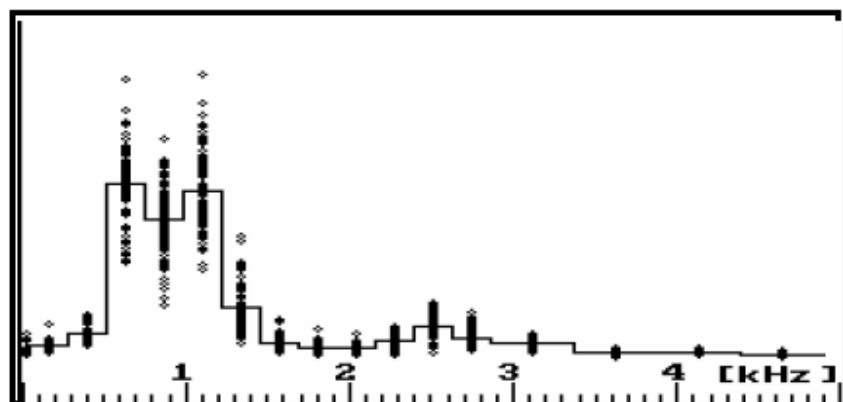
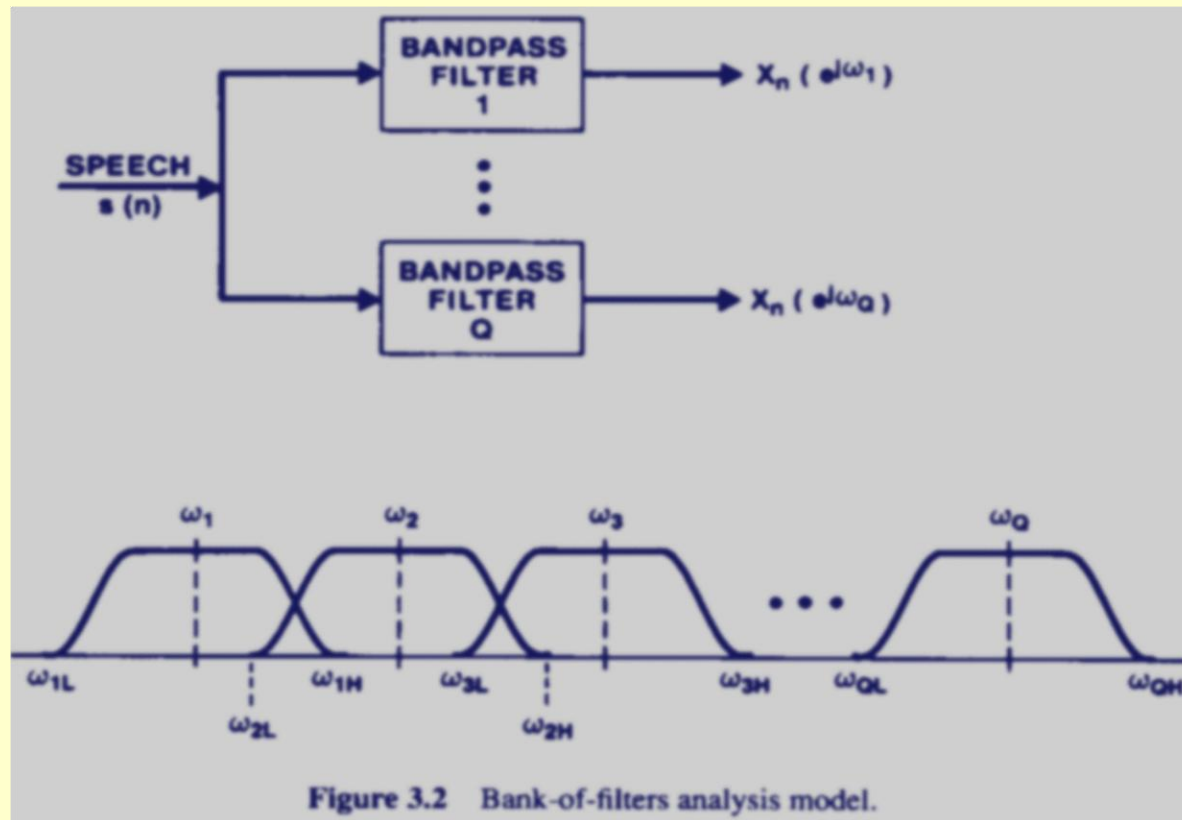


Fig.3.5.18. Variabilitatea intravorbitor reflectată în spectrul mediat pentru 100 de vectori rezultați din rostirea vocalei "a" (doi vorbitori masculini).

4. ANALIZA PRIN BANC DE FILTRE DIGITALE

$$\begin{aligned}x_i(n) &= s(n) \otimes h_i(n) \\ &= \sum_{m=0}^{L-1} h_i(m)s(n-m)\end{aligned}$$

$$X_i(z) = S(z)H_i(z) = Z\{s(n) \otimes h_i(n)\}$$



- Masurarea energiei SV in anumite benzi
- Filtrele analogice (Dudley 1939, Bell Labs) >>> filtre digitale

Analiza prin Banc de Filtre Digitale

$$s_i(n) = s(n) \otimes h_i(n)$$

$$= \sum_{m=0}^{L-1} h_i(m)s(n-m)$$

Deplaseaza spectrul benzii semnalului spre banda de JF si creaza imagini de IF

Pastreaza componenta DC
Si elimina imaginile de IF
create de neliniaritati

Fiecare filtru da masura energiei SV in banda sa

Reduce datele

Compresie
Log, u-law
(7bit/esantion)

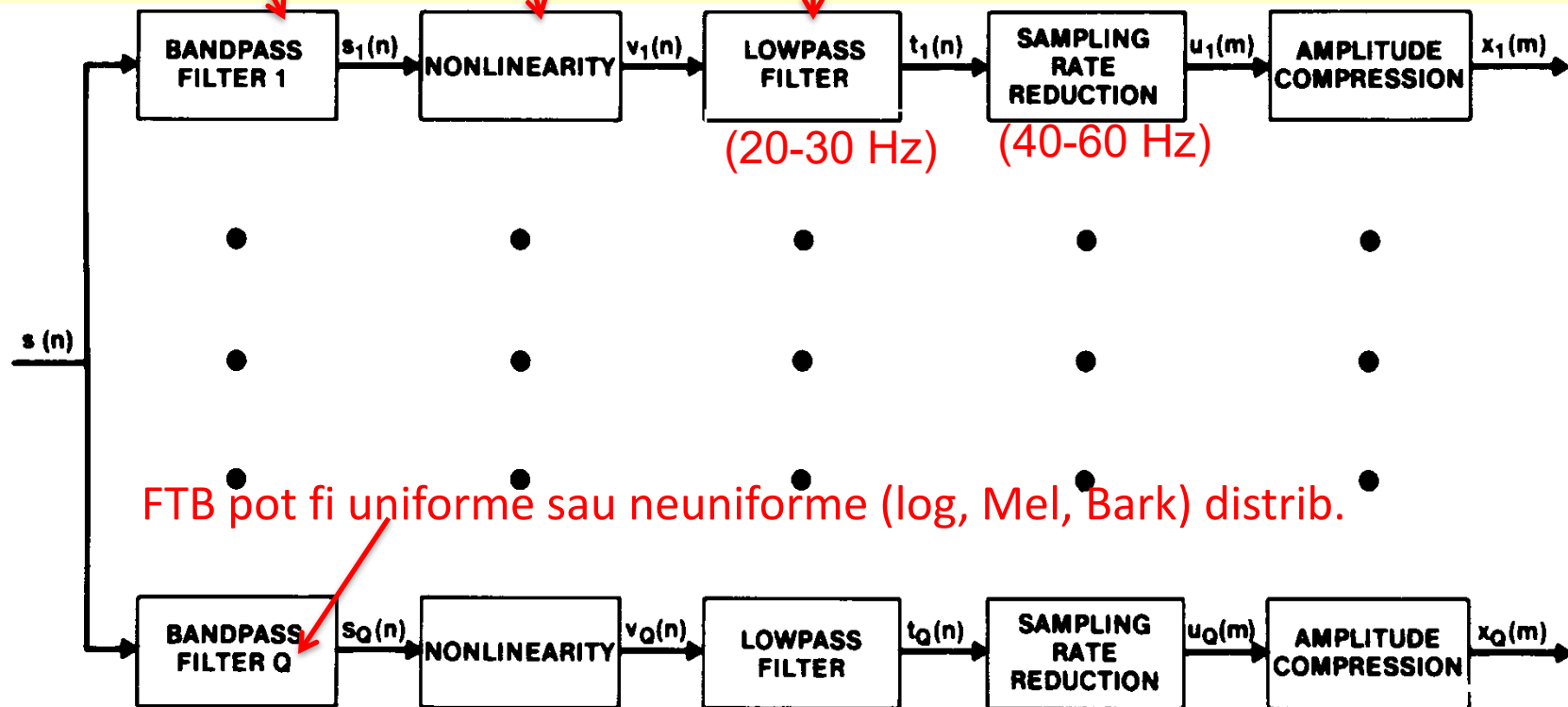


Figure 3.4 Complete bank-of-filters analysis model.

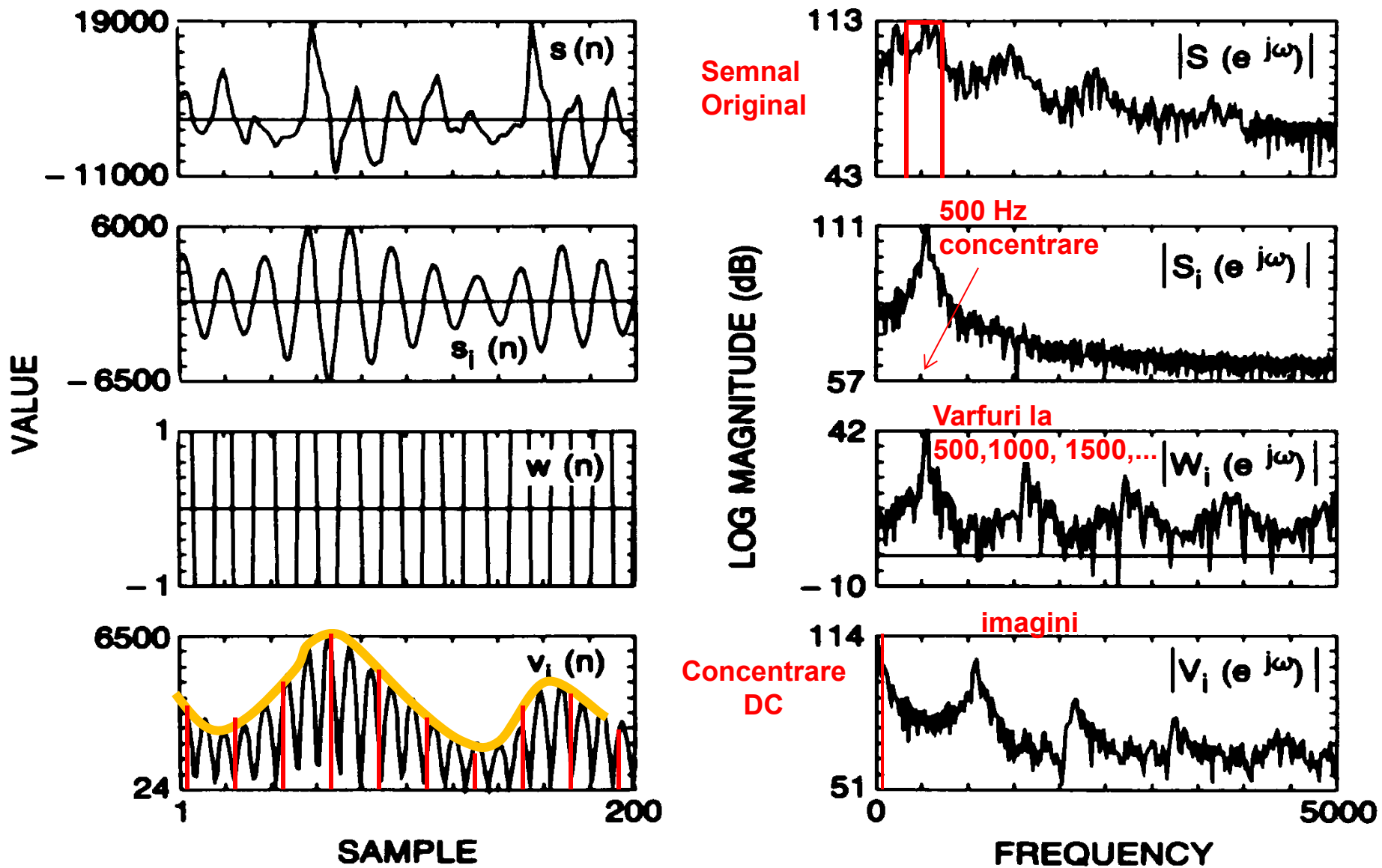


Figure 3.6 Typical waveforms and spectra of a voice speech signal in the bank-of-filters analysis model.

Banc de filtre Uniform

Banda Filtrului
(no overlapping)

Frecventele
Centrale

$$f_i = \left(\frac{F_s}{N} \right) i, \quad 1 \leq i \leq Q$$

Numarul
filtrelor

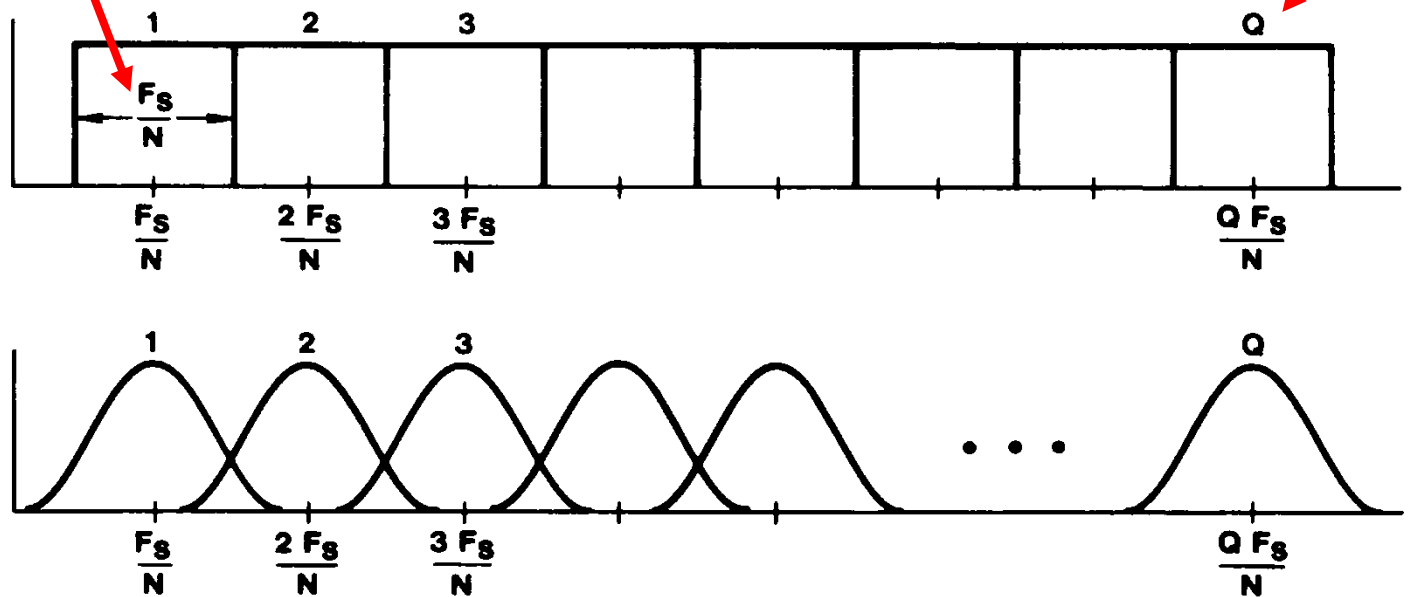


Figure 3.7 Ideal (a) and realistic (b) set of filter responses of a Q -channel filter bank covering the frequency range F_s/N to $(Q + 1/2)F_s/N$.

Q – numarul filtrelor uniform distribuite pe gama de frecvente a SV

BF Neuniform

- Scara de frecventa logaritmica
- Scara de banda critica
 - Mel Scale
 - Bark Scale

Scala Logaritmica de Frecventa

- Ptr. Q FTB, frecventa centrala, f_i si banda b_i :

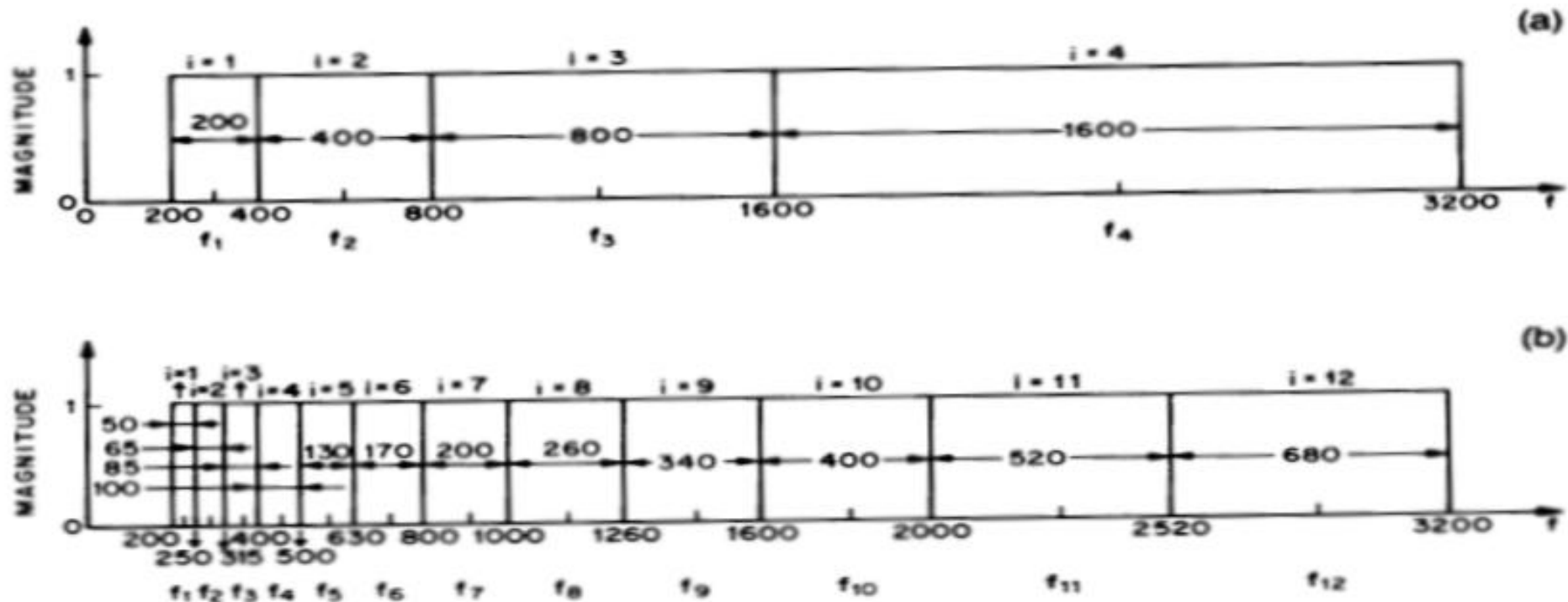
$$\begin{aligned} b_1 &= C \\ b_i &= \alpha b_{i-1}, \quad 2 \leq i \leq Q \\ f_i &= f_1 + \sum_{j=1}^{i-1} b_j + \frac{(b_i - b_1)}{2} \end{aligned}$$

Banda arbitrara ptr.
primul filtru

α =Factor logaritmic
de crestere, uzual 2.

Frecventa centrala filtru i
F1 frecventa arbitrara la primul filtru

$$C = 200\text{Hz}; \quad f_1 = 300\text{Hz}; \quad \alpha = 2; \quad Q = 4;$$



$$C = 50\text{Hz}; \quad f_1 = 225\text{Hz}; \quad \alpha = 1.33 \quad Q = 12;$$

Figure 3.8 Ideal specifications of a 4-channel octave band-filter bank (a), a 12-channel third-octave band filter bank (b), and a 7-channel critical band scale filter bank (c) covering the telephone bandwidth range (200–3200 Hz).

Scara de banda critica (perceptuala)

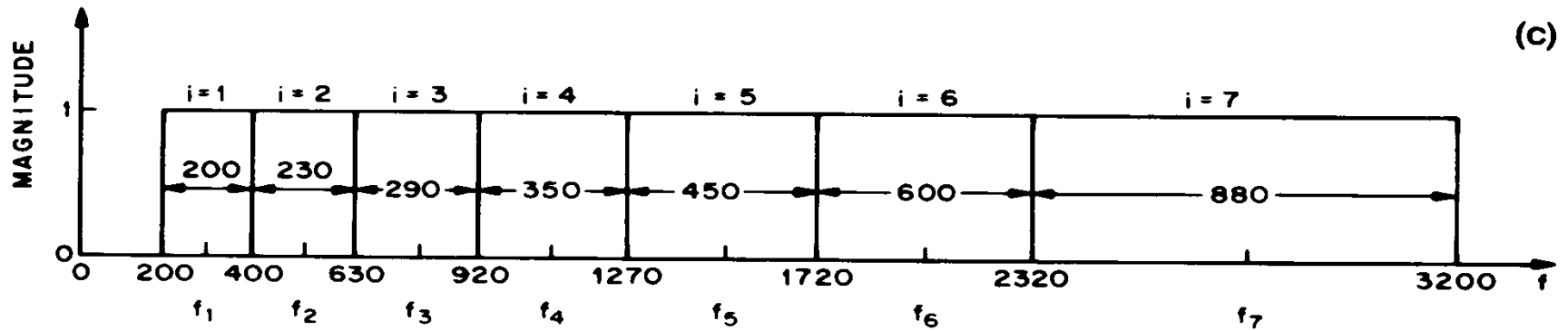


Figure 3.8 Ideal specifications of a 4-channel octave band-filter bank (a), a 12-channel third-octave band filter bank (b), and a 7-channel critical band scale filter bank (c) covering the telephone bandwidth range (200–3200 Hz).

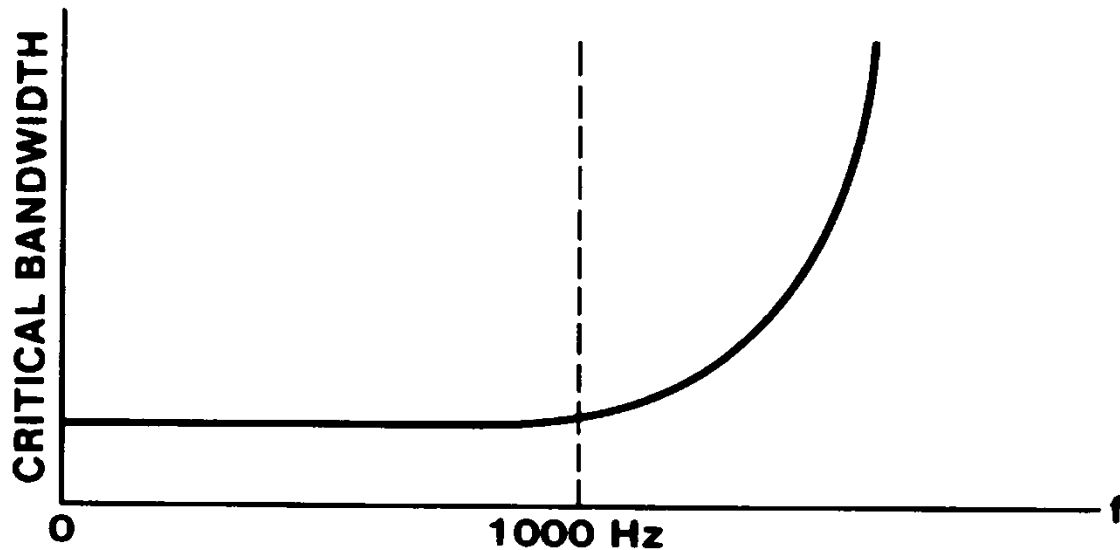
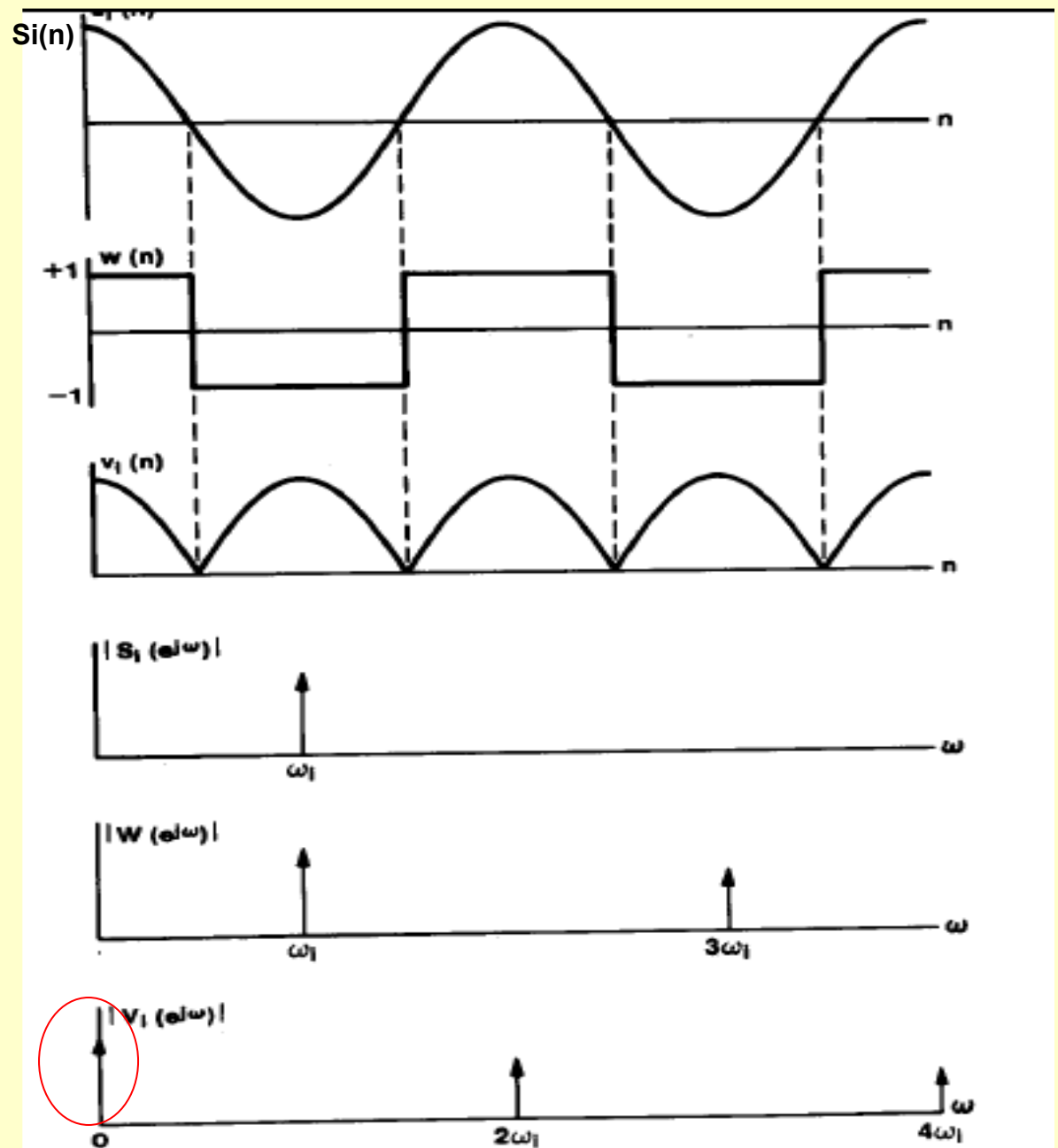
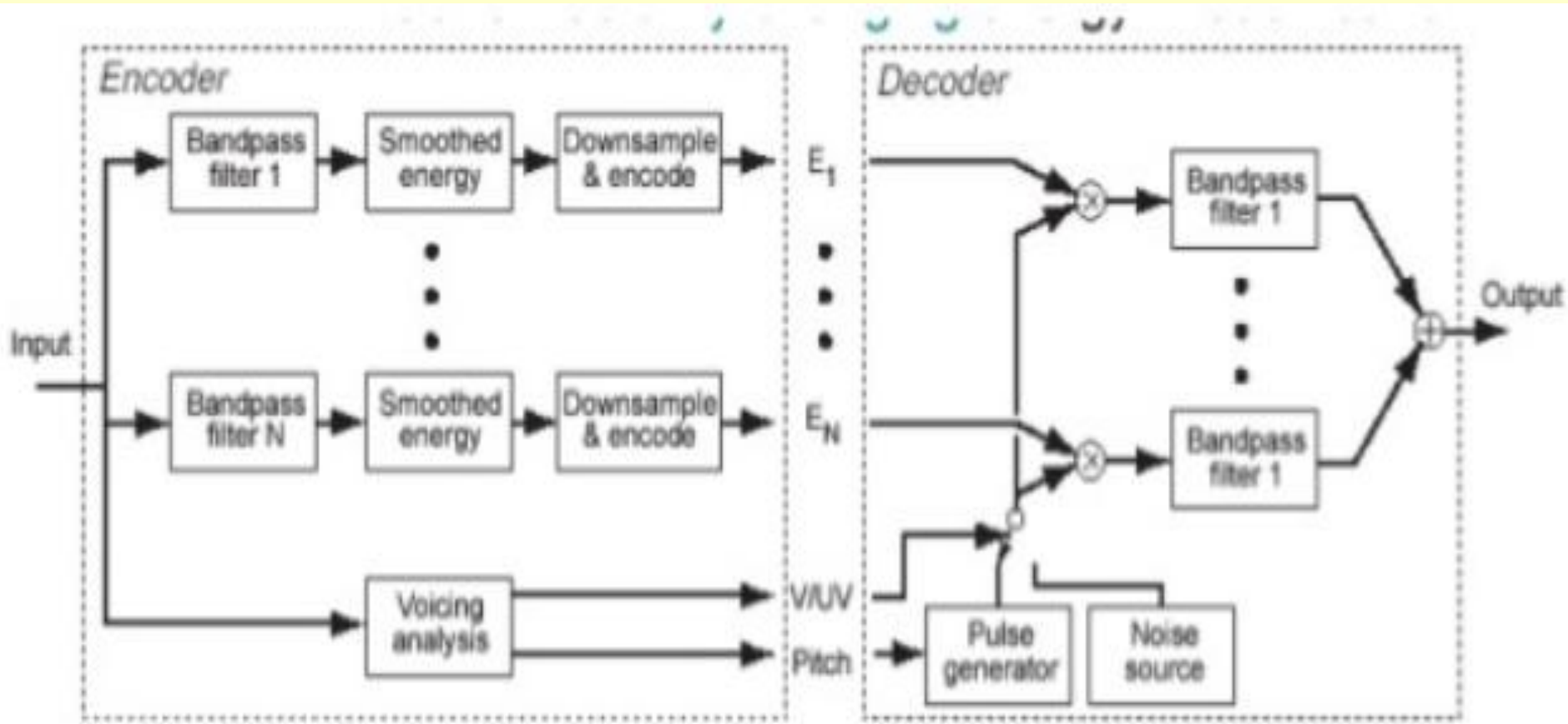


Figure 3.9 The variation of bandwidth with frequency for the perceptually based critical band scale.



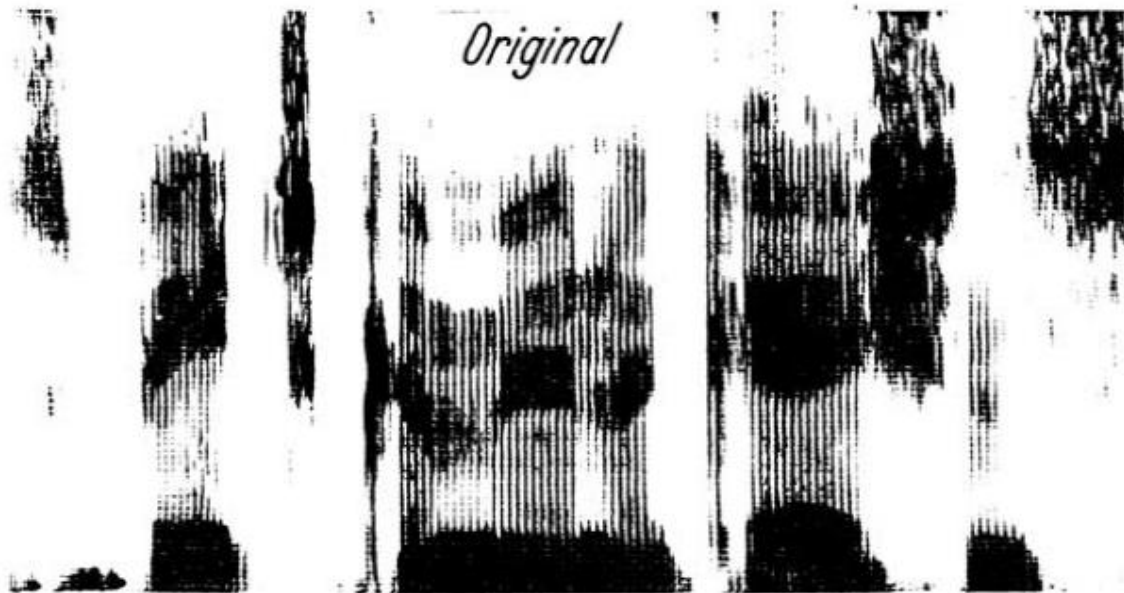
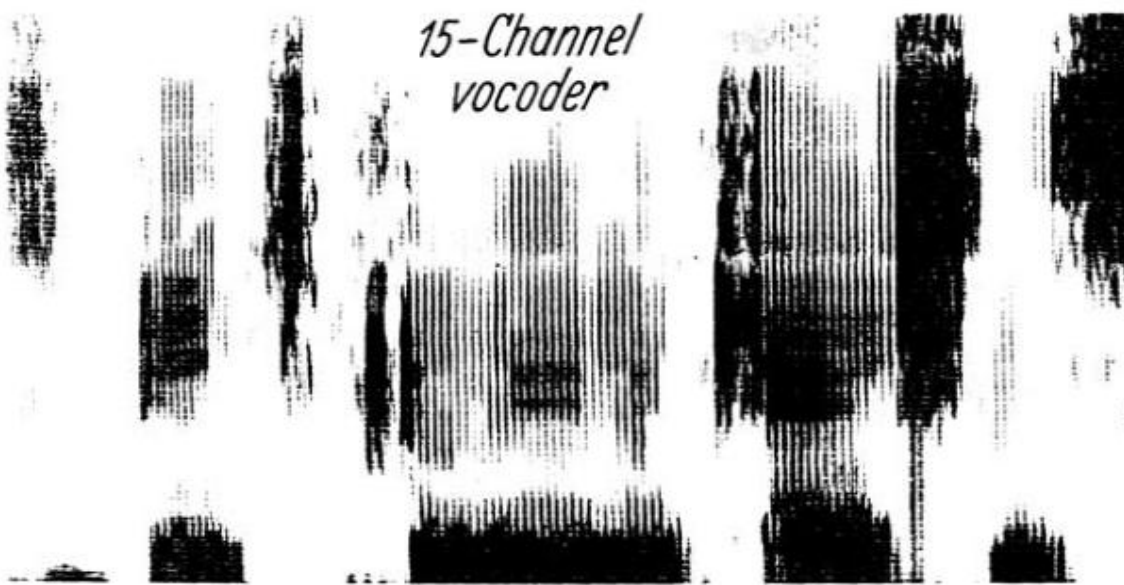
Typical waveforms and spectra for analysis of a pure sinusoid in the filter-bank model.

Exemplu simplificat



Vocoder de canal - separarea sursa-tract

- bancul de filtre imparte spectrul in benzi spectrale
- transmite schimbarea lenta a energiei in benzi
- 10-20 benzi aliniata perceptual
- subesantionare
- excitarea cu impulsuri/zgomot



S P E E C H C O M M U N I C A T I O N S

Ex. Care este rata de compresie a unui vocoder de canal cu $Q=16$ FTB, la prelucrarea unui SV cu $B < 8\text{kHz}$ fata de codarea PCM? Presupunem $f_{\text{es}}=16\text{kHz}$, si rez. CAN 8 biti.

PCM

$$\text{BitRate1} = 16\text{kHz} \cdot 8\text{b} = 128\text{kbps}$$

Vocoder

$$\text{BR2} = 16 \cdot 50\text{Hz} \cdot 7\text{b} = 5.6\text{kbps}$$

$$C = \text{BR1} / \text{BR2} \sim 23$$

Fig. 6.63 An example of a 15 channel vocoder. (After Flanagan [31]).

5. ANALIZA HOMOMORFICA

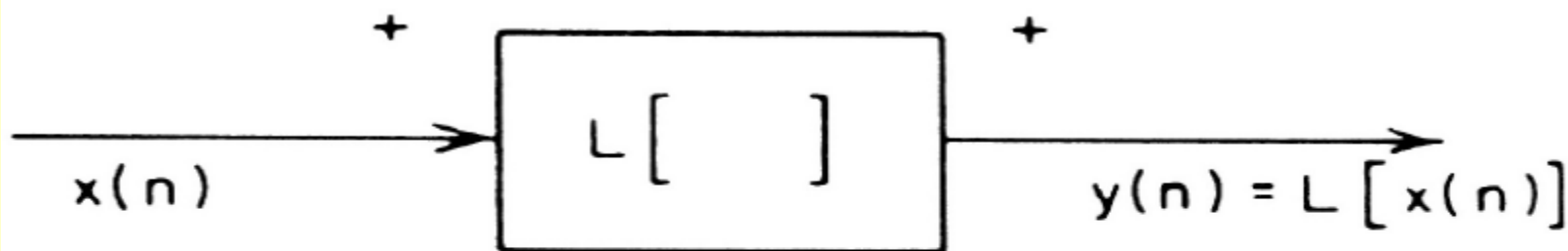
- Filtrarea/transformarea homomorfica este o transformare neliniară, $x[n] = D(x[n])$, aplicată de obicei, la prelucrări de imagine și voce, utilizată pentru a transforma un semnal obținut din convoluția a două semnale în suma celor două semnale.*

Analiza vorbirii=>estimarea parametrilor unui model de producere a SV si măsurarea variațiilor lor

$$\text{SV} = \text{excitație} * \text{răspuns tract}$$

⇒ Dacă doresc să separ din SV excitația de răspunsul tractului se pot folosi si metode de filtrare homomorifice

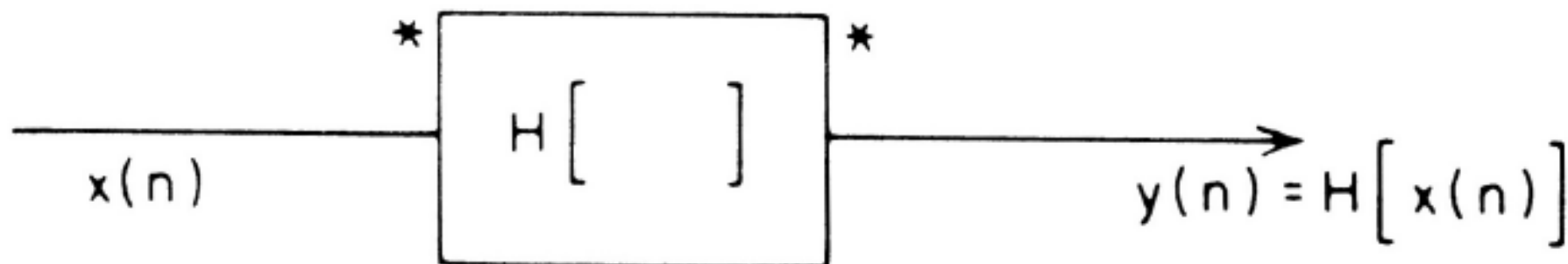
⇒ **Sistemele liniare respecta principiul superpozitiei:**



$$x(n) = ax_1(n) + bx_2(n)$$

$$y(n) = L[x(n)] = aL[x_1(n)] + bL[x_2(n)]$$

⇒ Sistemele homomorfe respecta principiul superpozitiei generalizate (ptr. convolutie):



- Pentru un sistem LTI (linear invariant in timp) avem:

$$y(n) = x(n) * h(n) = \sum_{k=-\infty}^{\infty} x(k) h(n-k)$$

- principiul superpozitiei “generalizate” inlocuieste + prin convolutie (*):

$$x(n) = x_1(n) * x_2(n)$$

$$y(n) = H[x(n)] = H[x_1(n)] * H[x_2(n)]$$

Filtru homomorfic => sistem homomorfic $[H]$ care permite semnalului dorit sa treaca nealterat, in timp ce opreste pe cel nedorit

$$x(n) = x_1(n) * x_2(n) \quad ; X1 - \text{nedorit}$$

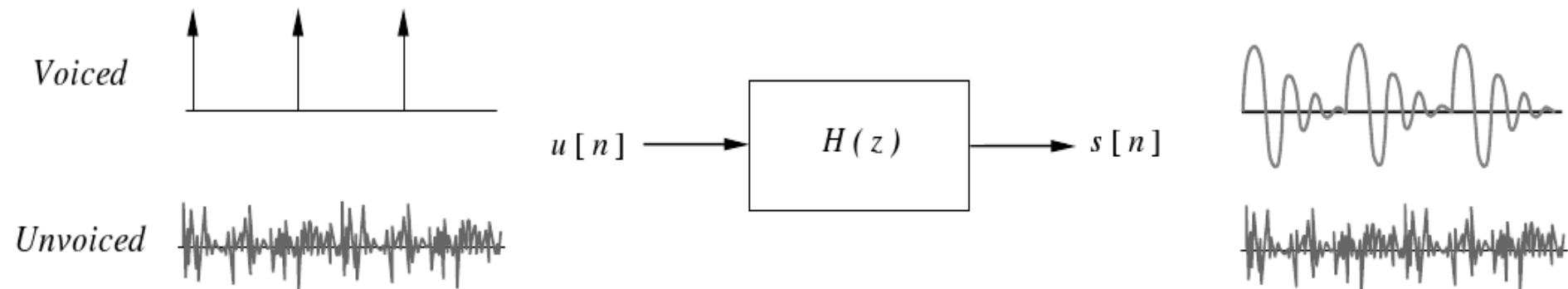
$$H[x(n)] = H[x_1(n)] * H[x_2(n)]$$

$$H[x_1(n)] \rightarrow \delta(n) \quad ; \text{Opresc } X1$$

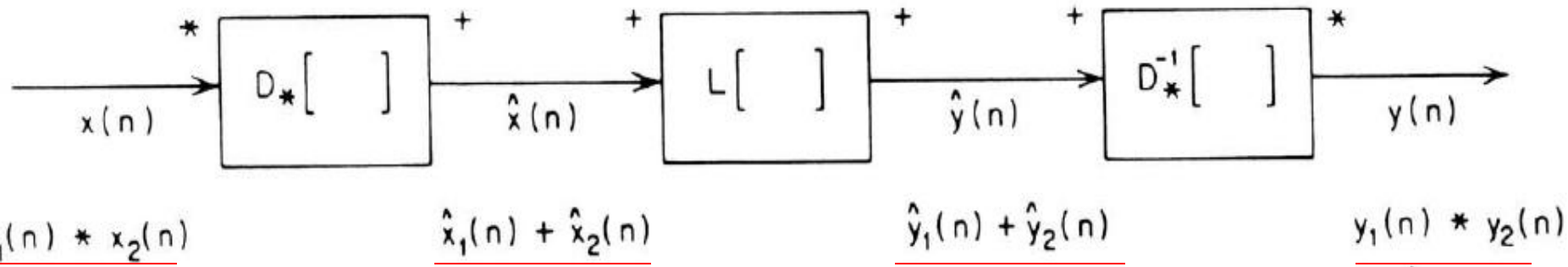
$$H[x_2(n)] \rightarrow x_2(n)$$

$$H[x(n)] = \delta(n) * x_2(n) = x_2(n)$$

- Pentru sistemele liniare putem face analogia cu eliminarea zgomotelor aditive



Forma canonica pentru convolutia homomorfica



-orice sistem homomorfic poate fi reprezentat ca si 3 sisteme/blocuri cascade ptr convolutie

1. Sistemul ia intrarile combinate prin convolutie si le transforma in iesiri aditive
2. Sistemul este unul liniar conventional
3. Sistemul este inversul primului sistem – ia intrarile aditive si le transforma in iesiri prin convolutie

$$x(n) = x_1(n) * x_2(n)$$

$$\hat{x}(n) = D_*[x(n)] = \hat{x}_1(n) + \hat{x}_2(n)$$

$$\hat{y}(n) = L[\hat{x}_1(n) + \hat{x}_2(n)] = \hat{y}_1(n) + \hat{y}_2(n)$$

$$y(n) = D_*^{-1}[\hat{y}_1(n) + \hat{y}_2(n)] = y_1(n) * y_2(n)$$

-relatia de convolutie

-Relatia aditiva

-Sistem linear

- inversa rel. de conv.

Forma canonica pentru domeniul frecventa

- Trebuie gasit un sistem care transforma convolutia in adunare

$$x(n) = x_1(n) * x_2(n)$$

$$X(z) = X_1(z) \cdot X_2(z)$$

$$D_*[x(n)] = \hat{x}_1(n) + \hat{x}_2(n) = \hat{x}(n)$$

$$D_*[X(z)] = \hat{X}_1(z) + \hat{X}_2(z) = \hat{X}(z)$$

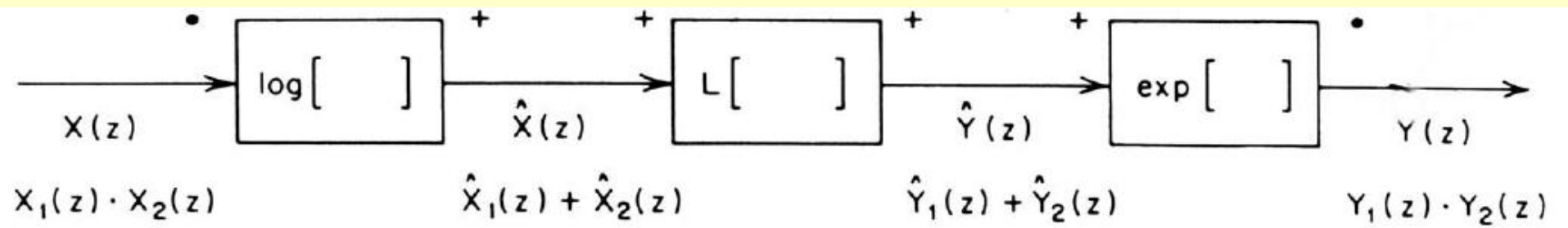
- Functia logaritm transforma produsul in suma :

$$\hat{X}(z) = \log[X(z)] = \log[X_1(z) \cdot X_2(z)]$$

$$= \log[X_1(z)] + \log[X_2(z)] = \hat{X}_1(z) + \hat{X}_2(z)$$

$$\hat{Y}(z) = L[\hat{X}_1(z) + \hat{X}_2(z)] = \hat{Y}_1(z) + \hat{Y}_2(z)$$

$$Y(z) = \exp[\hat{Y}_1(z) + \hat{Y}_2(z)] = Y_1(z) \cdot Y_2(z)$$





Modelul acustic liniar de producere a vorbirii

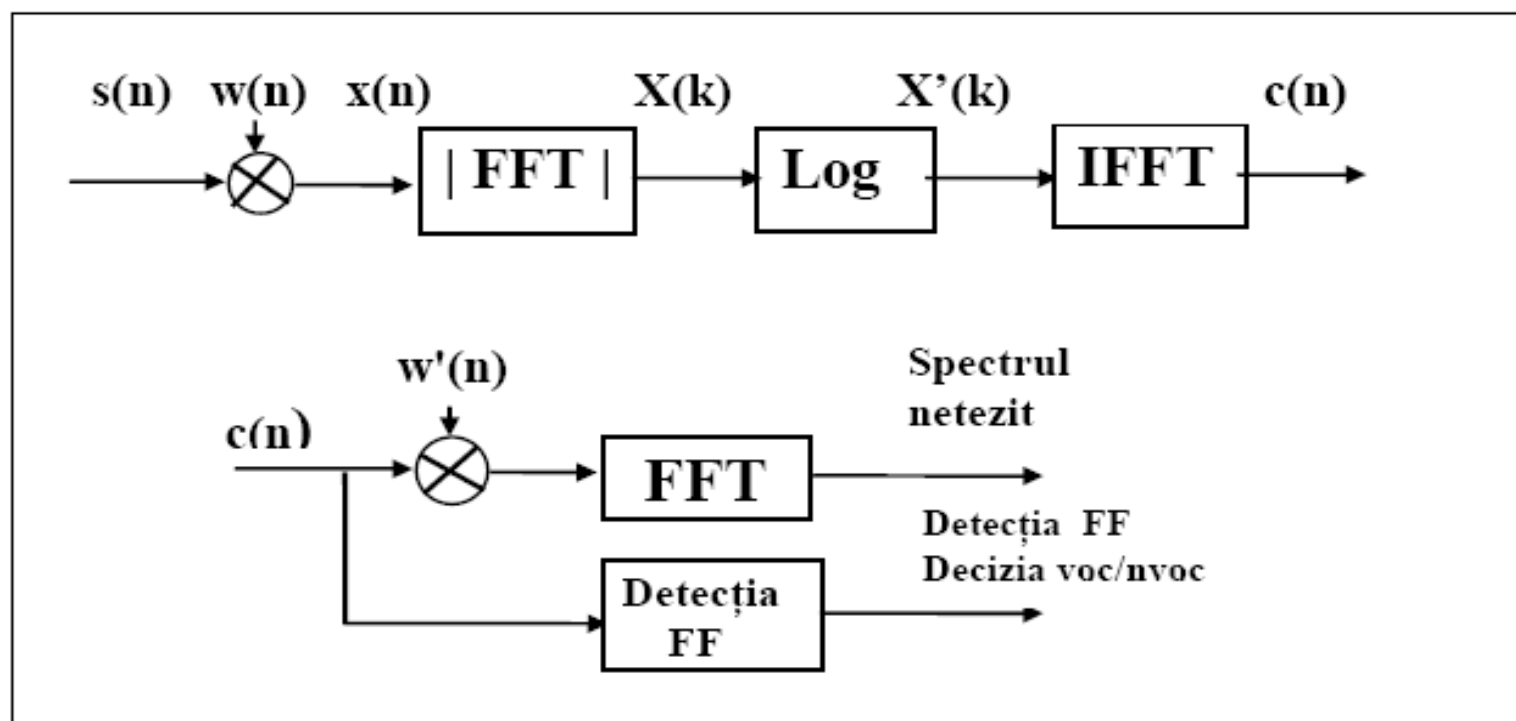
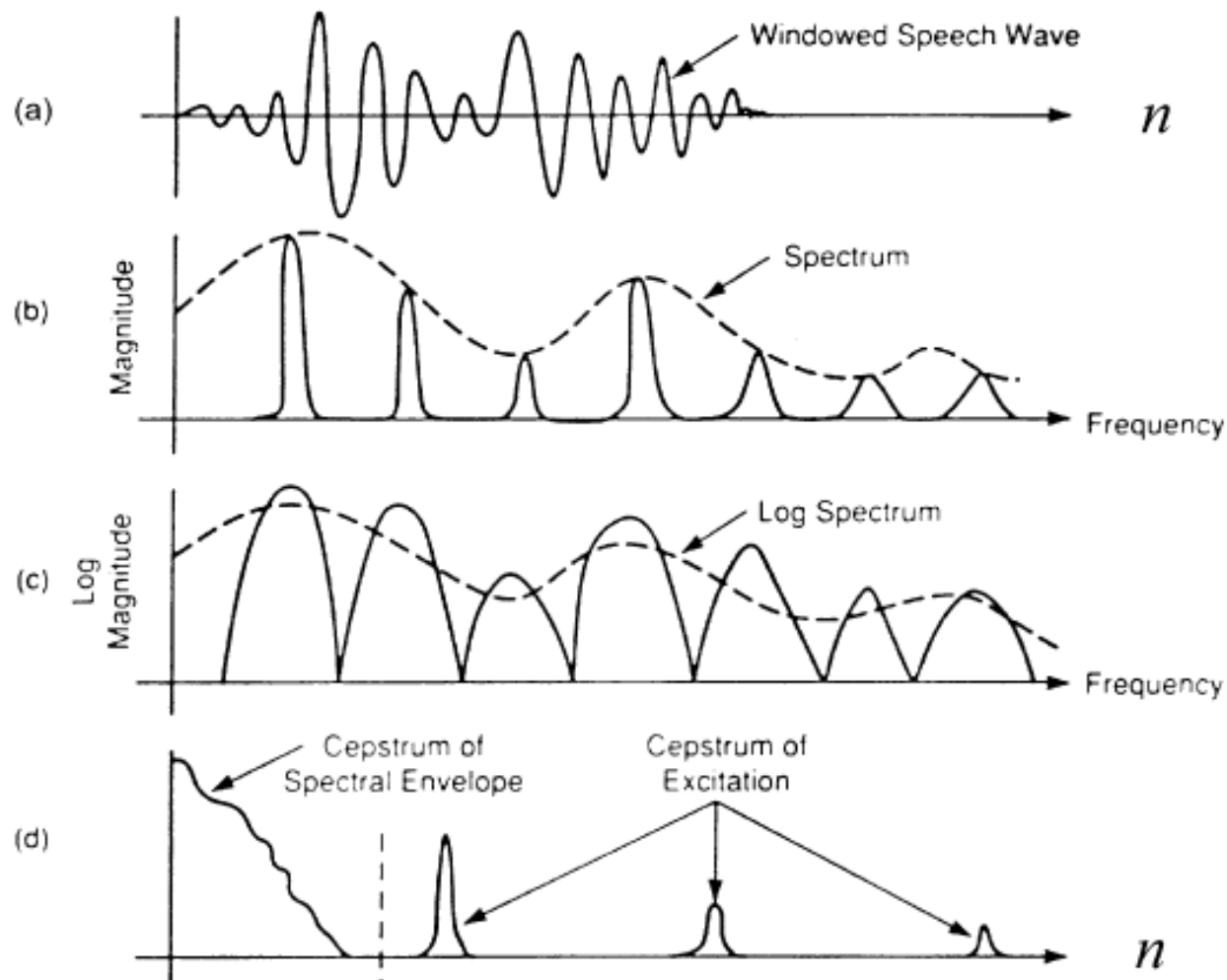
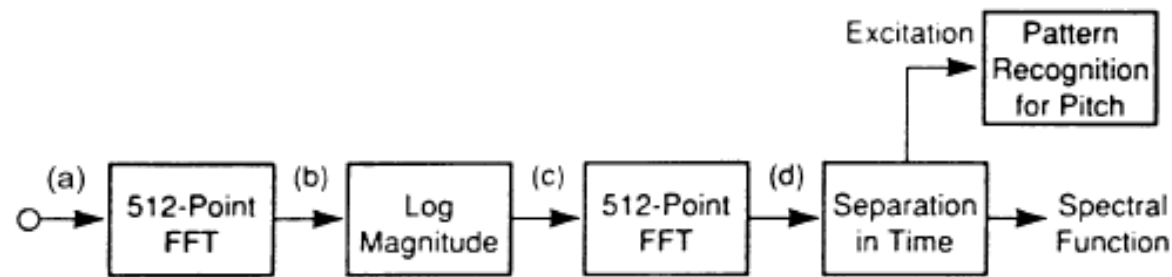
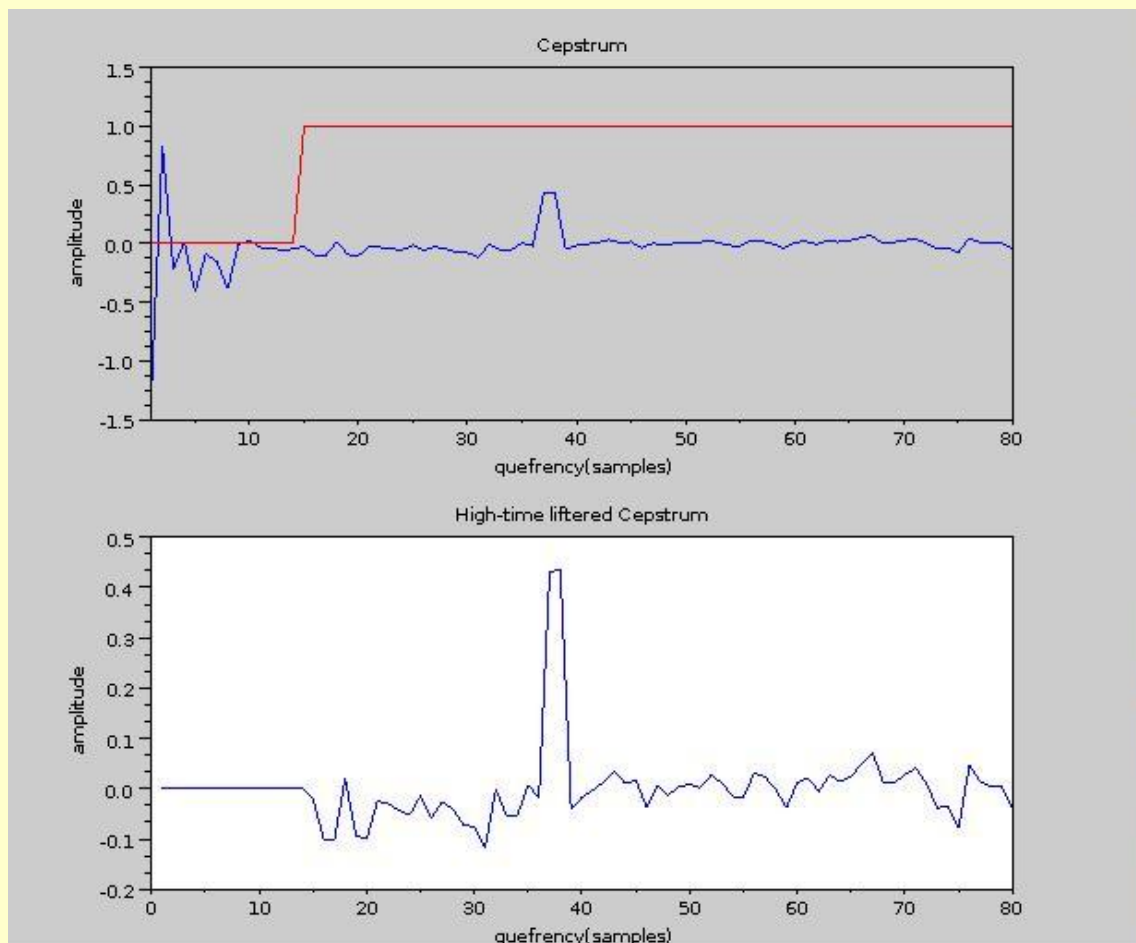


Diagrama analizei cepstrale pentru extragerea anvelopei spectrale netezite și a frecvenței fundamentale.



- Operațiunea de **liftare** este similară operațiunii de filtrare în domeniul frecvenței, în care o regiune de **quefrență** dorită pentru analiză este selectată prin înmulțirea întregului cepstru cu o fereastră dreptunghiulară în poziția dorită.
- Există două tipuri de operații de lifterizare, liftarea în timp redus și liftarea în timp ridicat.
- Operația de liftare în timp redus este efectuată pentru a extrage caracteristicile tractului vocal în domeniul quefrență și
- liftarea în timp mare este efectuată pentru a obține caracteristicile de excitație ale cadrului de vorbire analizat.

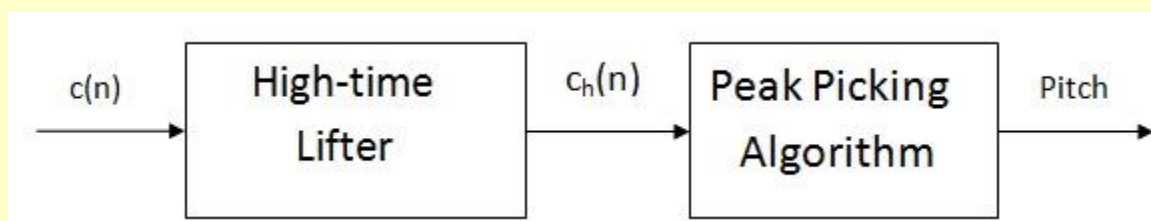


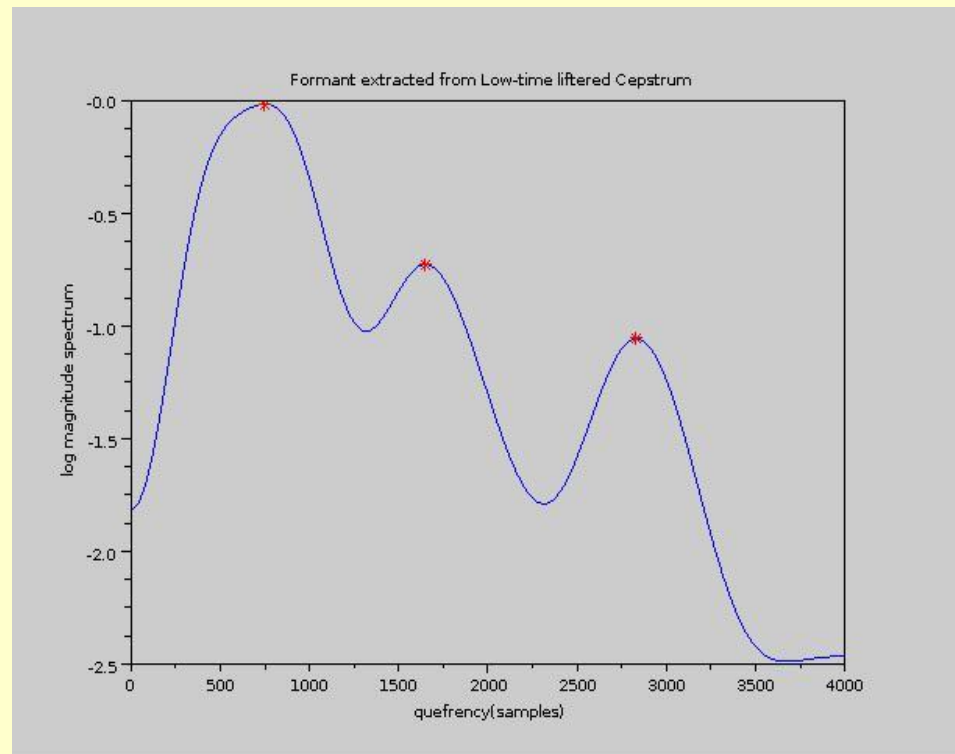
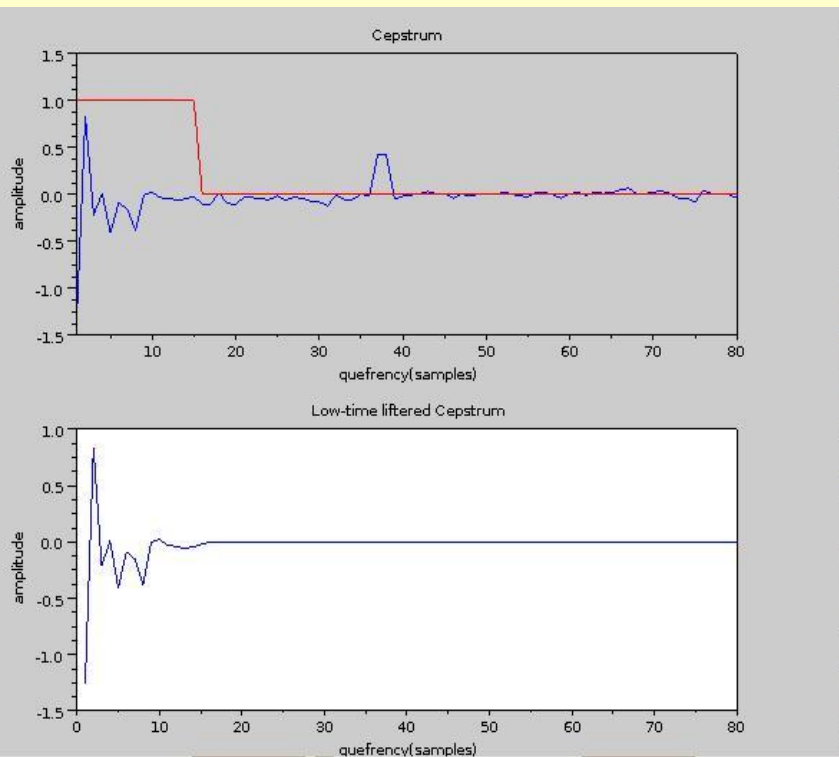
$$w_h[n] = \begin{cases} 1, & L_c \leq n < \frac{N}{2} \\ 0, & \text{else} \end{cases}$$

unde L_c este lungimea de tăiere a ferestrei de liftare, iar $N/2$ este jumătate din lungimea totală a cepstrului. De obicei, L_c este utilizat ca 15 sau 20.

Liftare de timp mare:

Cepstrul unui segment vocal și fereastra de liftare (în culoare roșie) și partea tractului vocal a cepstrului obținut prin liftarea în timp mare

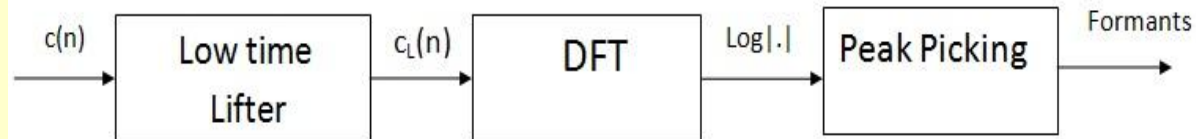




Liftarea în timp mic:

Cepstrul unui segment vocal și fereastra de liftare în timp mic (în culoare roșie) și caracteristicile tractului vocal ale cepstrului obținut prin liftarea în timp redus

$$w_e[n] = \begin{cases} 1, 0 \leq n \leq L_c \\ 0, L_c \leq n \leq \frac{N}{2} \end{cases}$$



unde L_c este lungimea de tăiere a ferestrei de liftare, iar $N/2$ este jumătate din lungimea totală a cepstrului. De obicei, L_c utilizat este 15 sau 20. Caracteristicile tractului vocal pot fi obținute prin înmulțirea cepstrului $c(n)$ cu fereastra de liftare cu timp mic

O intuiție despre caracteristicile cepstrale poate ajuta la a afla ce ar trebui să căutăm atunci când le folosim într-un sistem de vorbire.

- Întrucât caracteristicile cepstrale sunt calculate prin preluarea transformatei Fourier a spectrului aliniat logaritmic, acestea conțin informații despre modificările ratei în diferitele benzi ale spectrului. Caracteristicile Cepstrale sunt favorabile datorită capacității lor de a separa impactul sursei și de a filtra într-un semnal de vorbire. Cu alte cuvinte, în domeniul cepstral, influența corzilor vocale (sursă) și a tractului vocal (filtru) într-un semnal poate fi separată, deoarece excitația cu frecvență joasă și filtrarea formantică a tractului vocal, sunt situate în diferite regiuni în domeniul cepstral.

- Dacă un coeficient cepstral are o valoare pozitivă, acesta reprezintă un sunet sonor, deoarece majoritatea energiei spectrale din sunetele sonore este concentrată în regiunile cu frecvență joasă.

- Pe de altă parte, dacă un coeficient cepstral are o valoare negativă, acesta reprezintă un sunet fricativ, deoarece majoritatea energiilor spectrale din sunetele fricative sunt concentrate la frecvențe înalte.

- Coeficienții de ordin inferior conțin majoritatea informațiilor despre forma spectrală generală a funcției de transfer sursă-filtru.

- Coeficientul de ordin zero indică puterea medie a semnalului de intrare.

- Coeficientul de ordinul întâi reprezintă energia spectrală de distribuție între frecvențele joase și înalte.

- Chiar dacă coeficienții de ordin mare reprezintă niveluri crescânde de detalii spectrale, în funcție de rata de eșantionare și de metoda de estimare, 12-20 coeficienți cepstral sunt de obicei optimi pentru analiza vorbirii. Selectarea unui număr mare de coeficienți cepstral duce la o mai mare complexitate a modelelor.

De exemplu, dacă intenționăm să modelăm un semnal de vorbire printr-un model de mixturi gaussiene (GMM), dacă se utilizează un număr mare de coeficienți cepstrali, avem de obicei nevoie de mai multe date pentru a estima cu exactitate parametrii GMM.

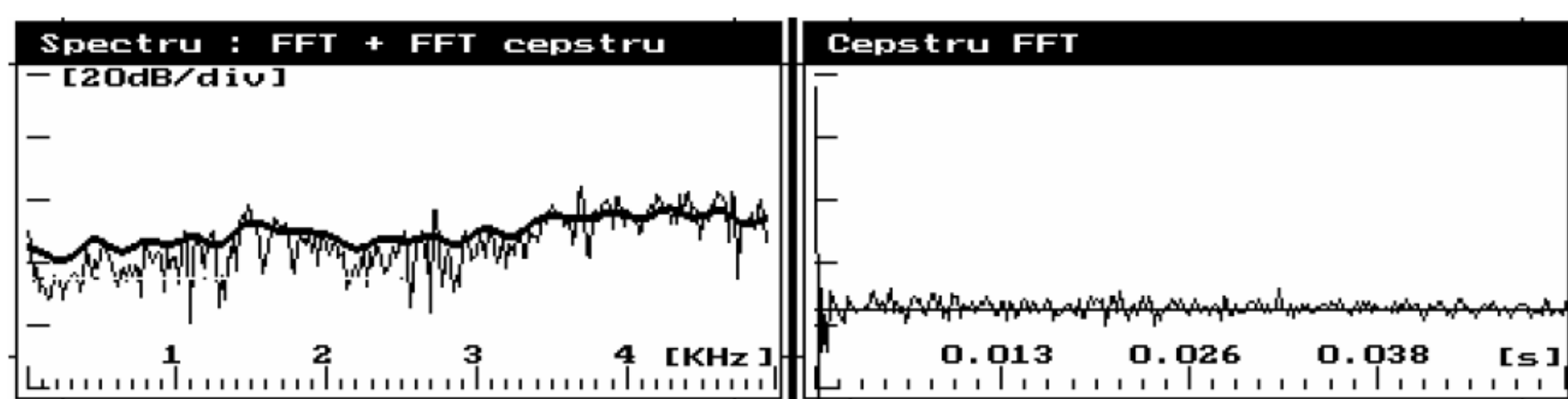
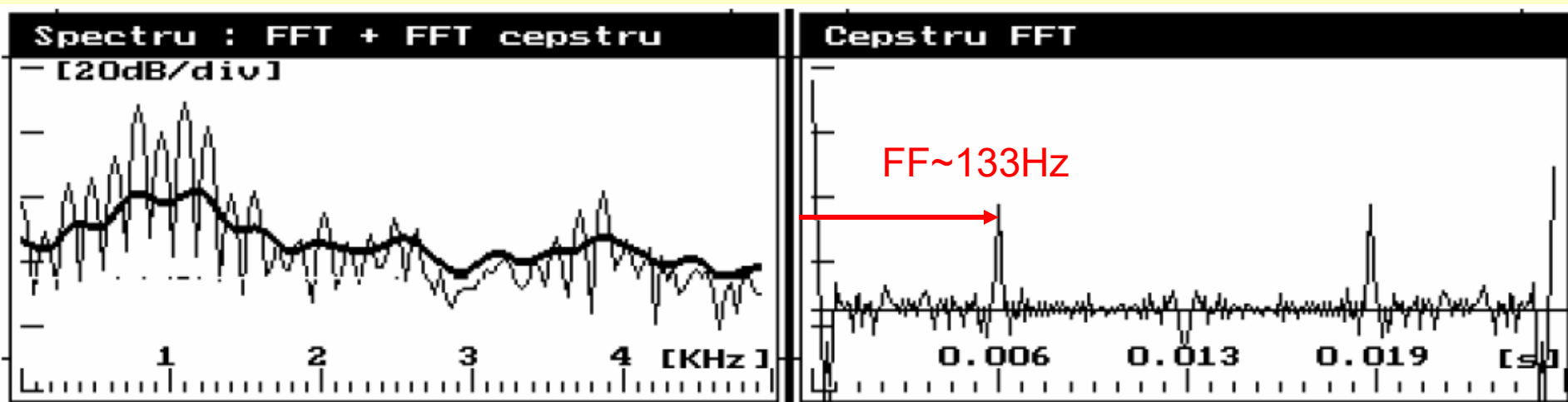


Fig. 3.7.3. Cepstrul și spectrul FFT corespunzător pentru un segment nevocalizat ("s")



Cepstrul și spectrul corespunzător pentru un segment vocalizat ("a") rostit de un locutor masculin

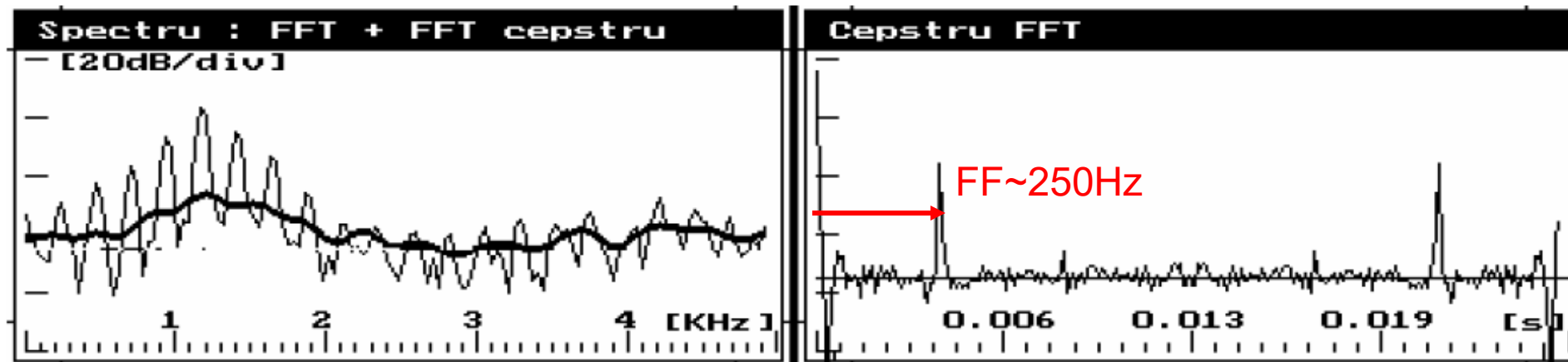


Fig. 3.7.5. Cepstrul și spectrul corespunzător pentru un segment vocalizat ("a") rostit de un locutor feminin (n=256), folosind o fereastră temporală de 3,2ms

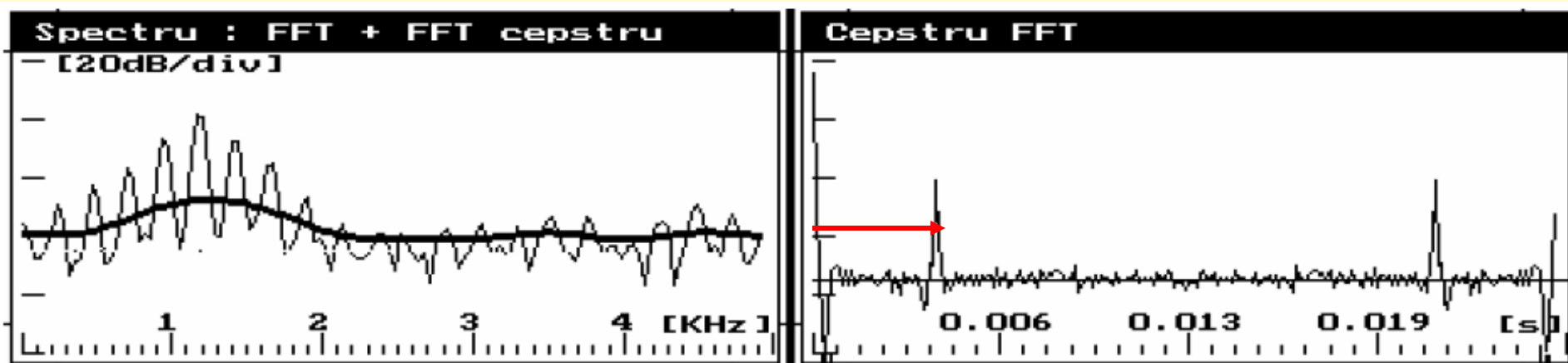
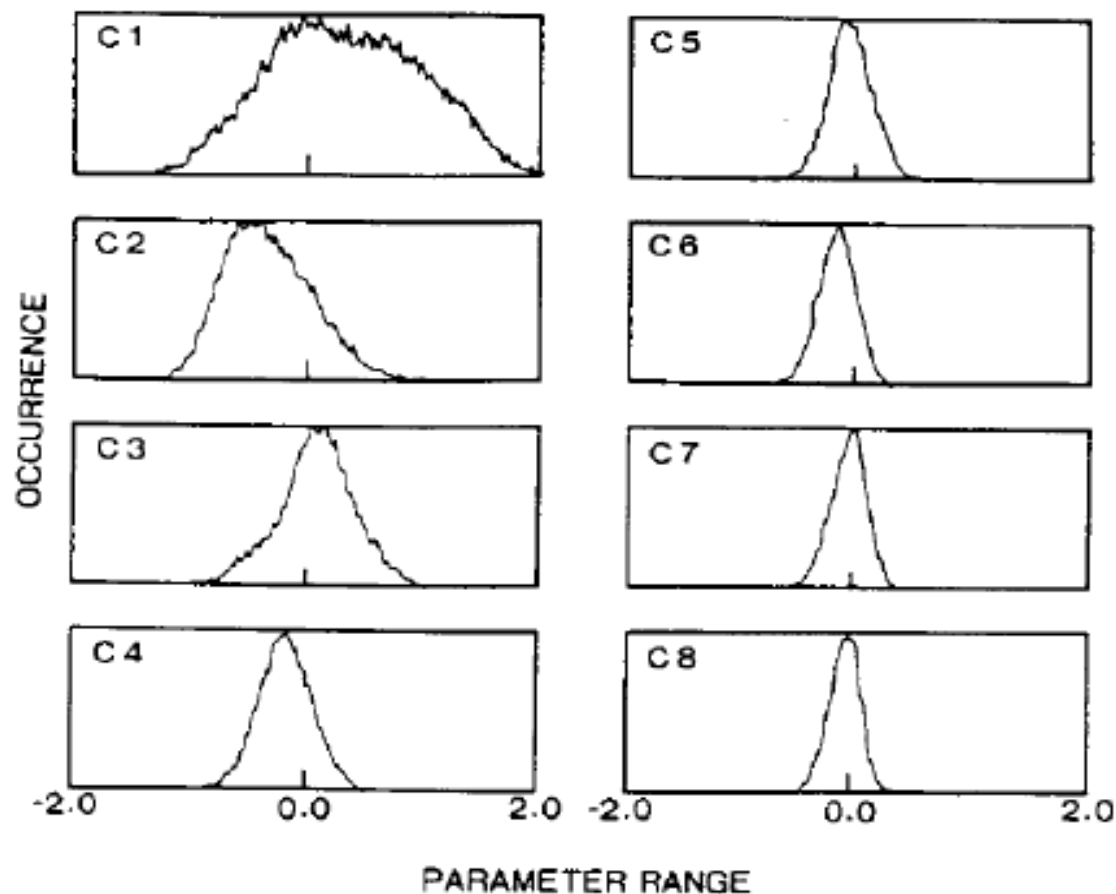


Fig. 3.7.6. Cepstrul și spectrul corespunzător pentru un segment vocalizat ("a") rostit de un locutor feminin (n=256), folosind o fereastră temporală, $w'(n)$, de 1,2ms

Statistical properties of cepstral coefficients

From a digit database (100 speakers) over dial-up telephone lines.



TDP: Cepstral Analysis

TEMA.

Un semnal vocal este esantionat la 20kHz. Pentru analiza spectrala pe timp scurt se utilizeaza o fereastră glisanta de 20ms care se deplaseaza cu 10ms pentru analiza cadrelor consecutive. Pentru calculul DFT se utilizeaza metoda FFT radix-2.

1. Cate esantioane se folosesc pentru fiecare cadru de analiza?
2. Care este rata de analiza a cadrelor pentru analiza spectrala pe timp scurt?
3. Ce dimensiune are este necesara pentru DFT si FFT pentru a garanta ca nu apare alias temporal?
4. Care este rezolutia in frecventa (Hz) intre 2 esantioane consecutive?

