Lecture 4

Frequency domain speech analysis (2)

Speech Signal ANALYSIS

Time domain:

- Average and maximum amplitude
- Amplitude density
- Average energy
- TEAGER energy
- Number of zero crossings
- Fundamental frequency (F0)
- TESPAR coding

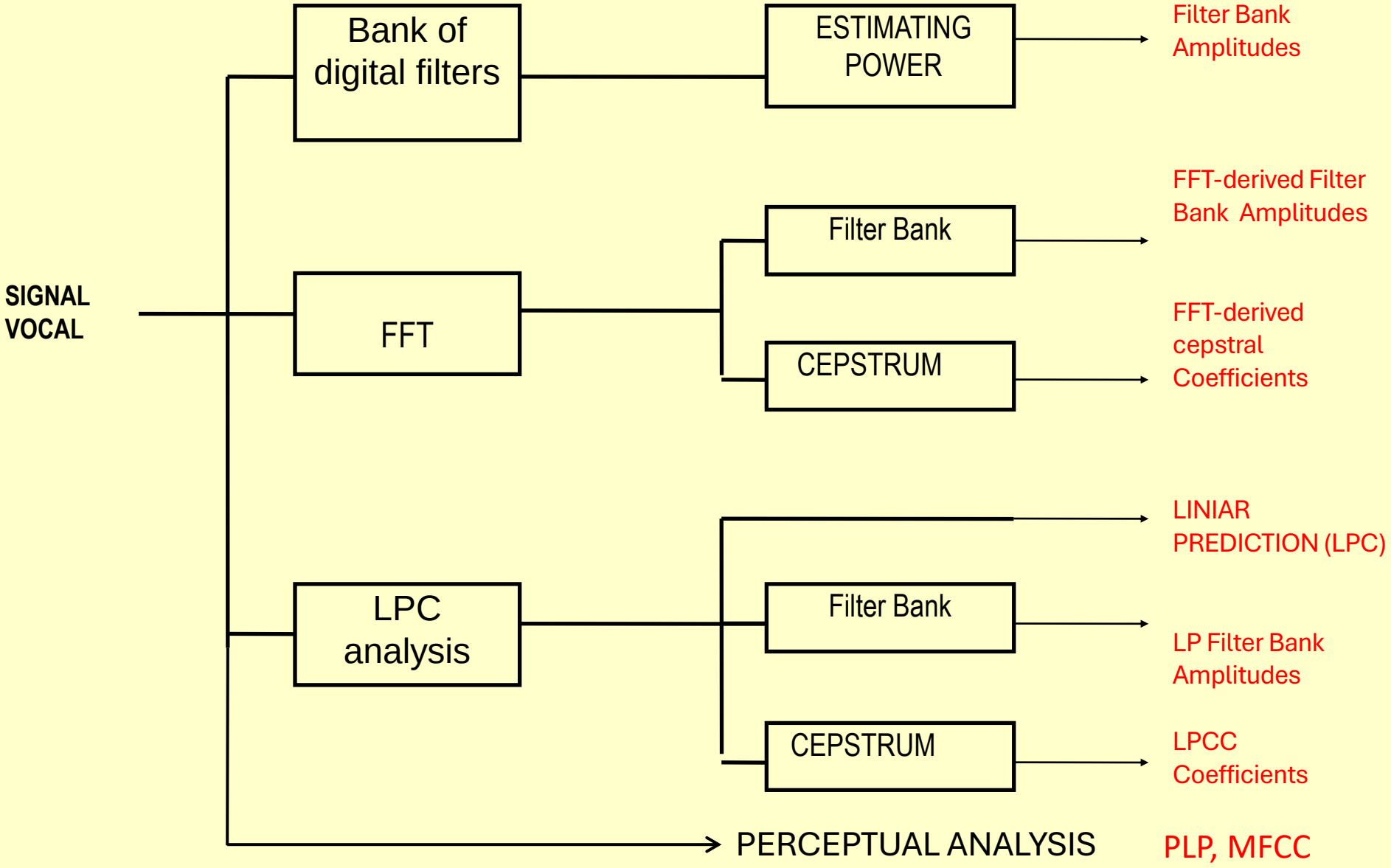
Frequency domain:

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

Time-frequency analysis

- 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

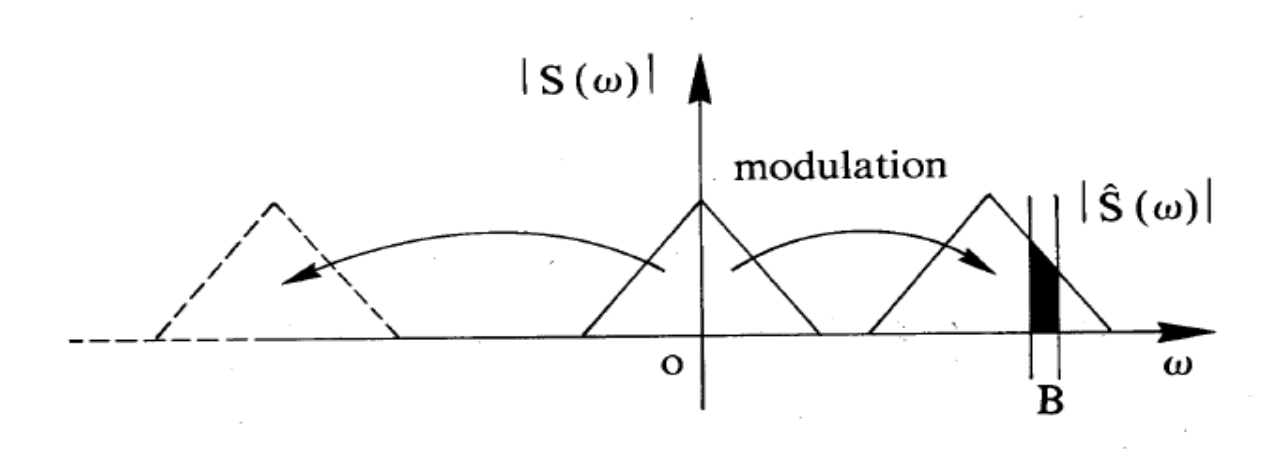


Algorithms for spectral analysis [Picone]

1. THE SPECTROGRAM (time-frequency-amplitude representation)

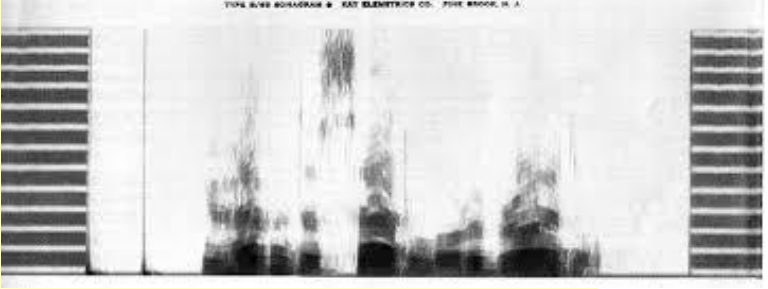
- Evolution of the SS spectral function over time - the spectrograph
- Spectrogram plotting can be: contour or bright/dark;
- The first instrument used by phoneticians - Key Elemetrics;
- Composed of: modulator, filter, sonogram plotting drum (f/t)
- The principle is superheterodyne filtering, if SV is s(t), the modulated signal is

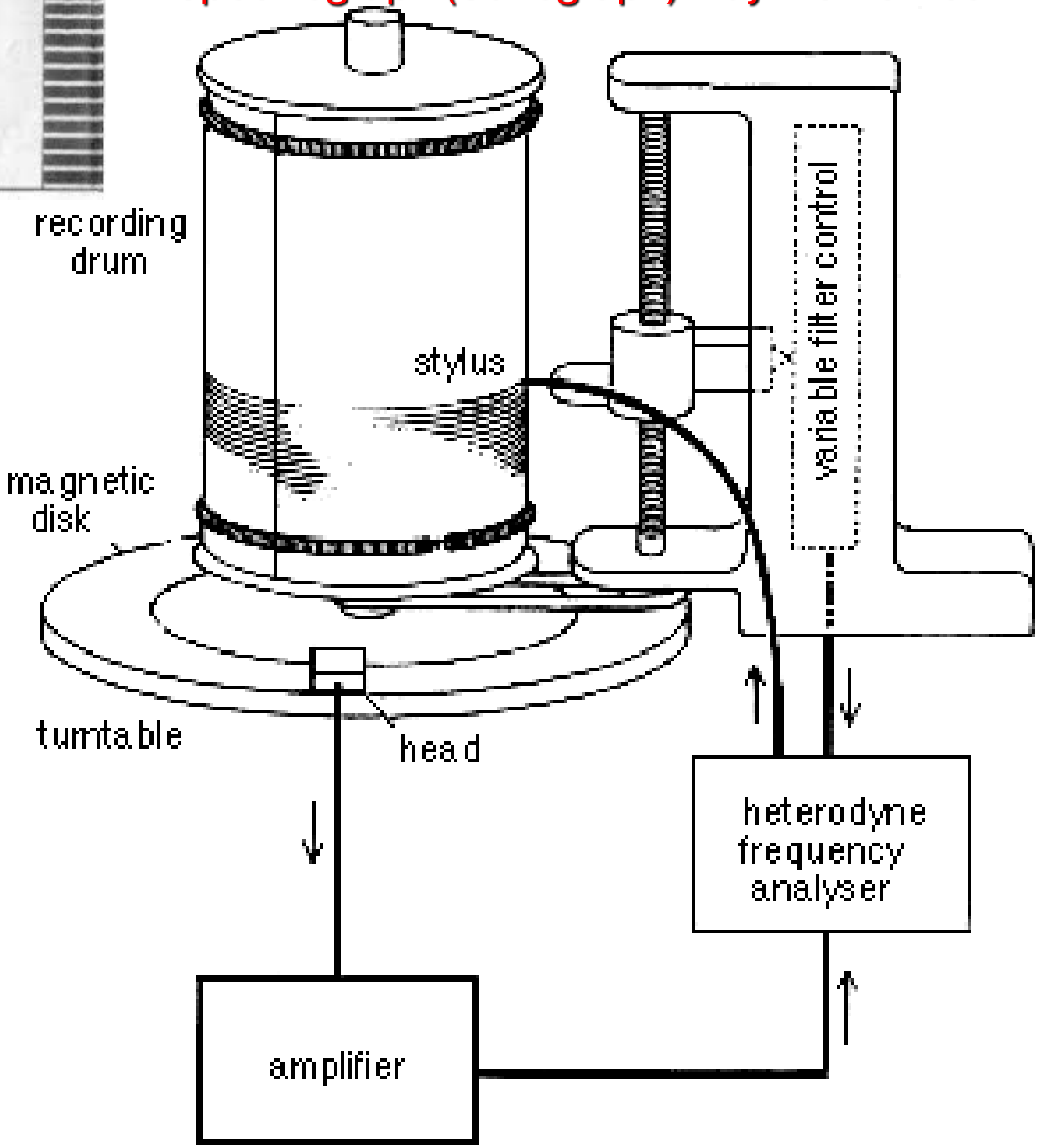
ŝ(t) = s(t)*cos 2πft and the spectrum is shifted to higher frequencies and sweeps the input of a BPF;

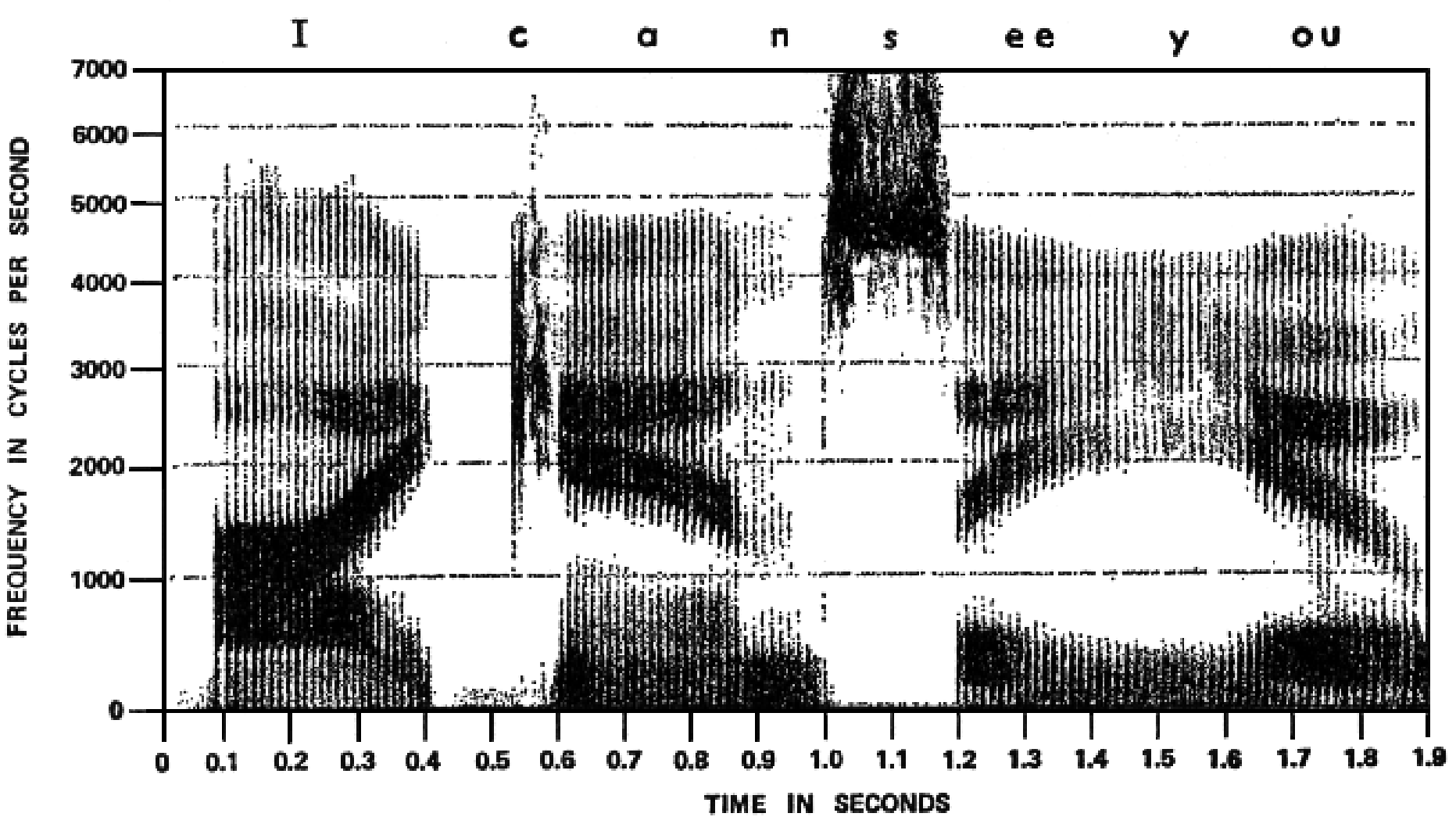


- The energy in this band is obtained by rectification-integration;
- Modify the carrier frequency to sweep the entire signal spectrum in the filter band;
- B=300Hz - **broadband spectrogram** - emphasizing the ***temporal changes of the signal***
- B=45Hz - **narrow band spectrogram** - emphasizes ***frequency changes in the signal***

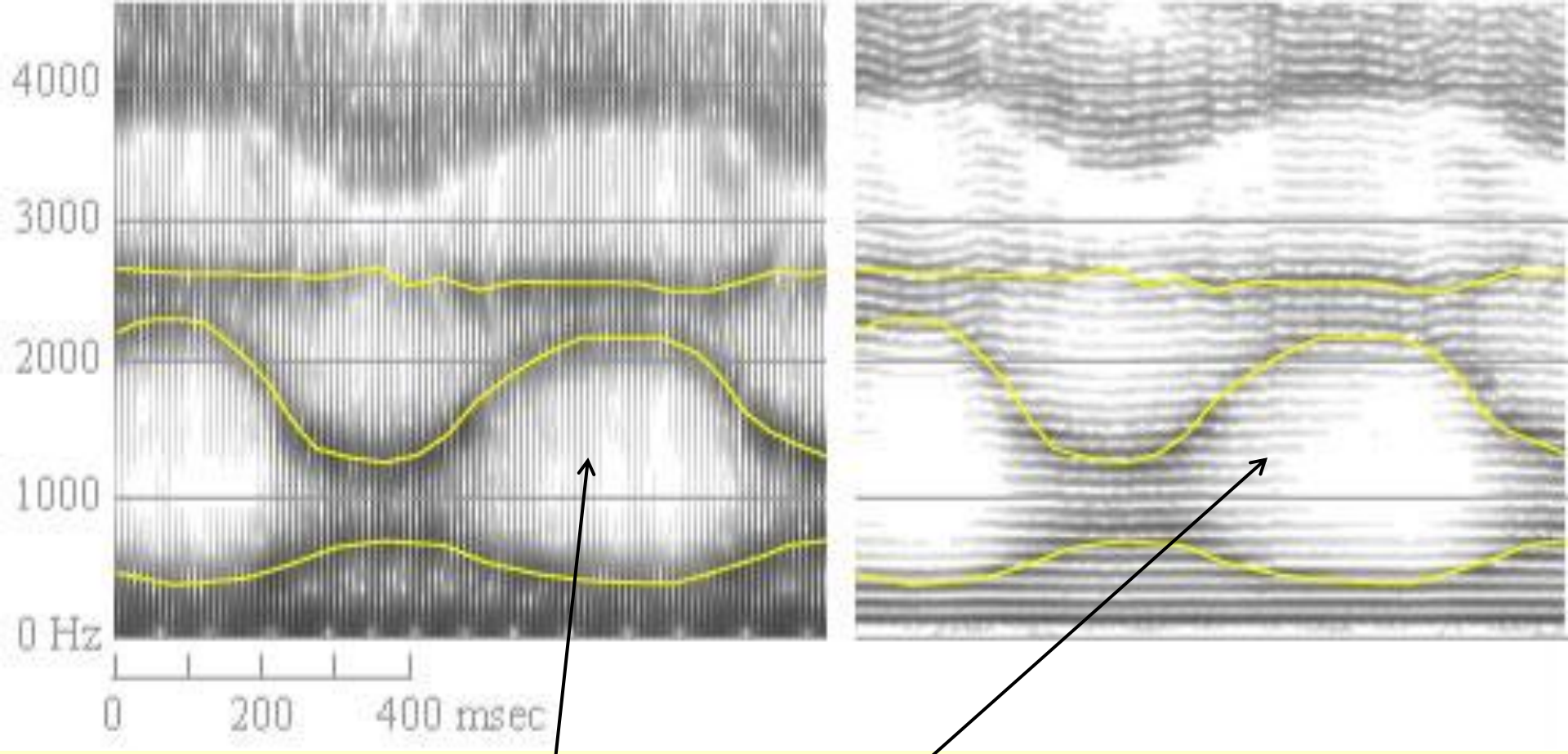
Spectrograph (Sonograph) Kay Elemetrics







THE SPECTROGRAM (sonagram)



Broadband or narrowband spectrogram

(depending on the length of the weighting window)

Obs. If B→0 the spectrum >>> Fourier spectrum

Broadband spectrogram - *has broad spectral spikes (formants) in time - highlights most of the individual periods of the FF as vertical grooves that the IR filter is comparable in time to a period of the FF.*

Narrow-band spectrogram

- *FF harmonics visible in sonorous regions*
- *formant frequencies still visible*
- *FF usually visible*
- *unvoiced regions do not show a well-profiled structure*

2. Spectral speech analysis - FFT

- the signal is a sum of sinusoids or complex exponentials, and it leads to practical solutions to problems (estimating formants, estimating F0, and analyzing the signal itself)

- Fourier representations provide - a convenient means of determining the response of linear systems to a sum of sinusoids

- clear evidence of signal properties hidden in the original signal

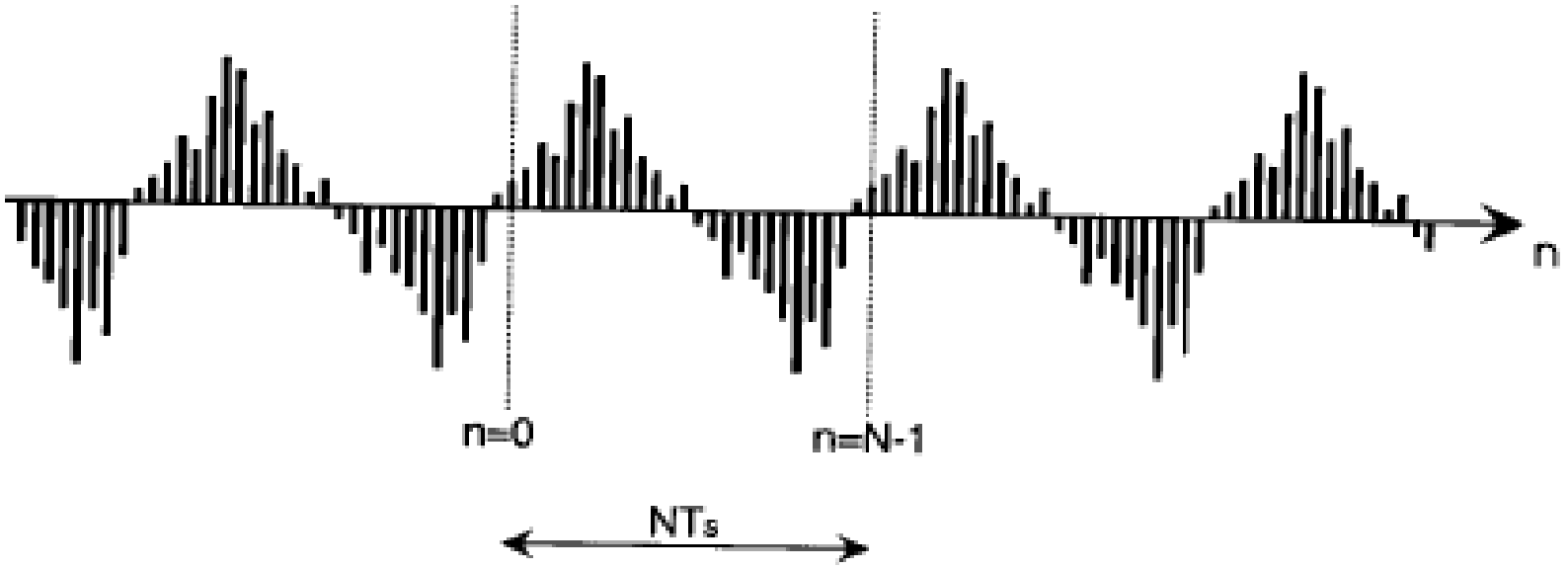
- the FFT algorithm (Cooley-Tukey 1965) allows real-time signal processing

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{kn}, \qquad 0 \le k \le N-1$$

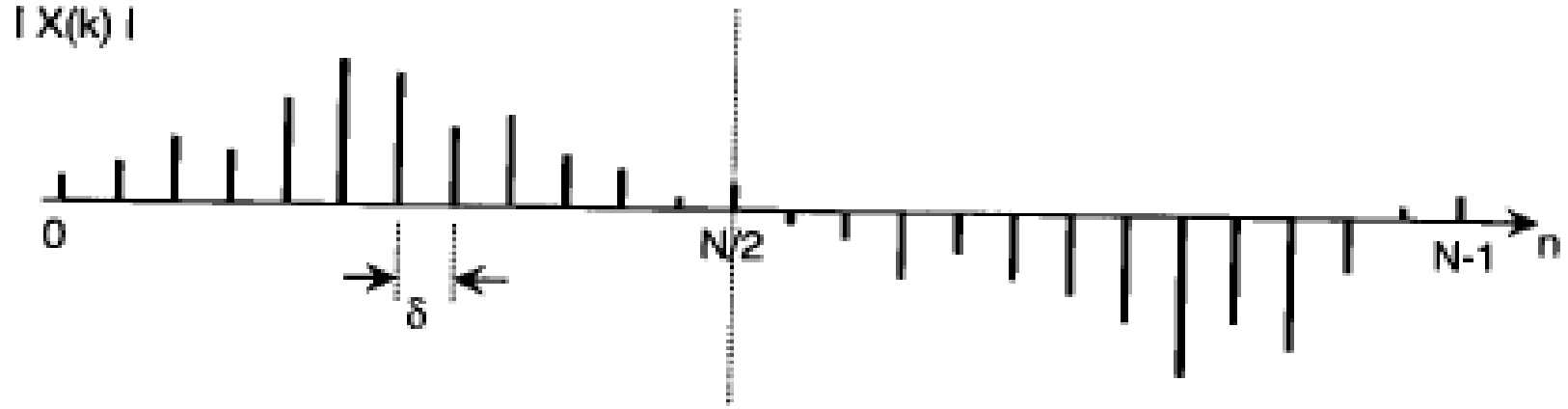
$$W_N = e^{-j2\pi/N}$$

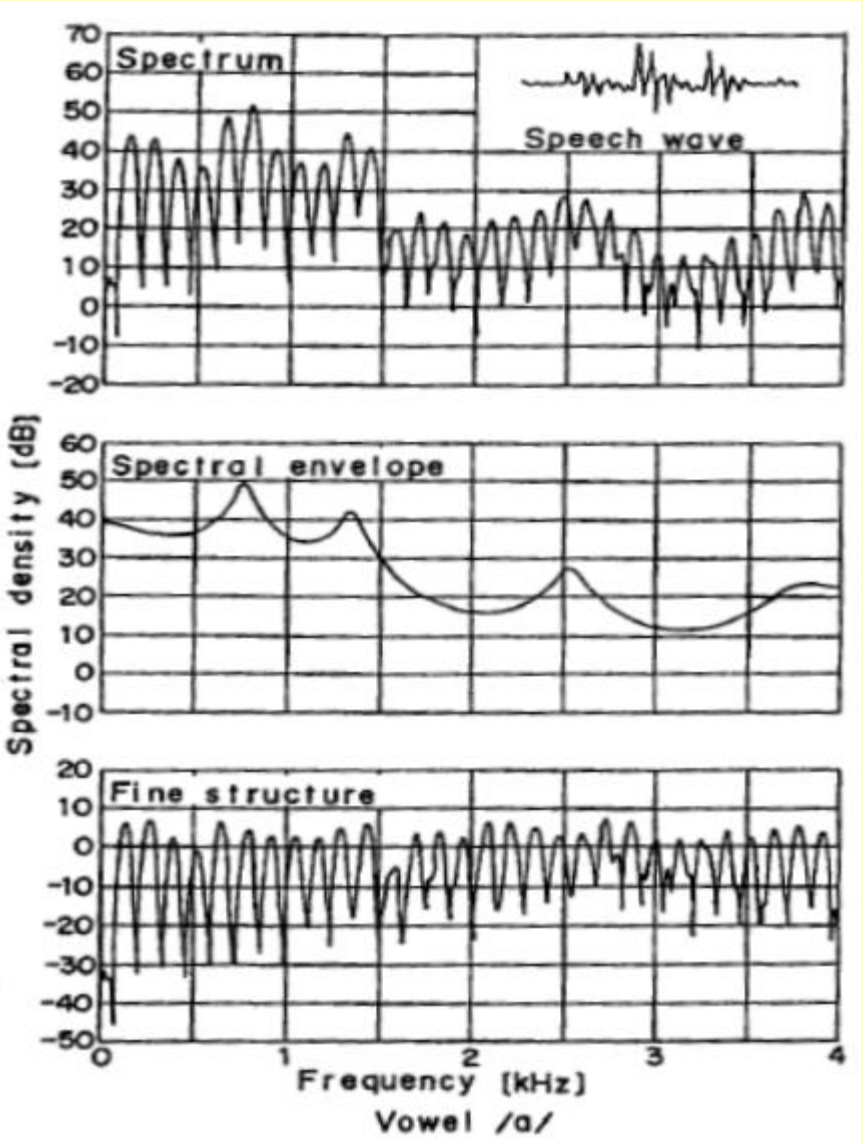
$$x(n) = \frac{1}{N} \sum_{n=0}^{N-1} X(k) W_N^{-nk}, \qquad 0 \le n \le N-1$$

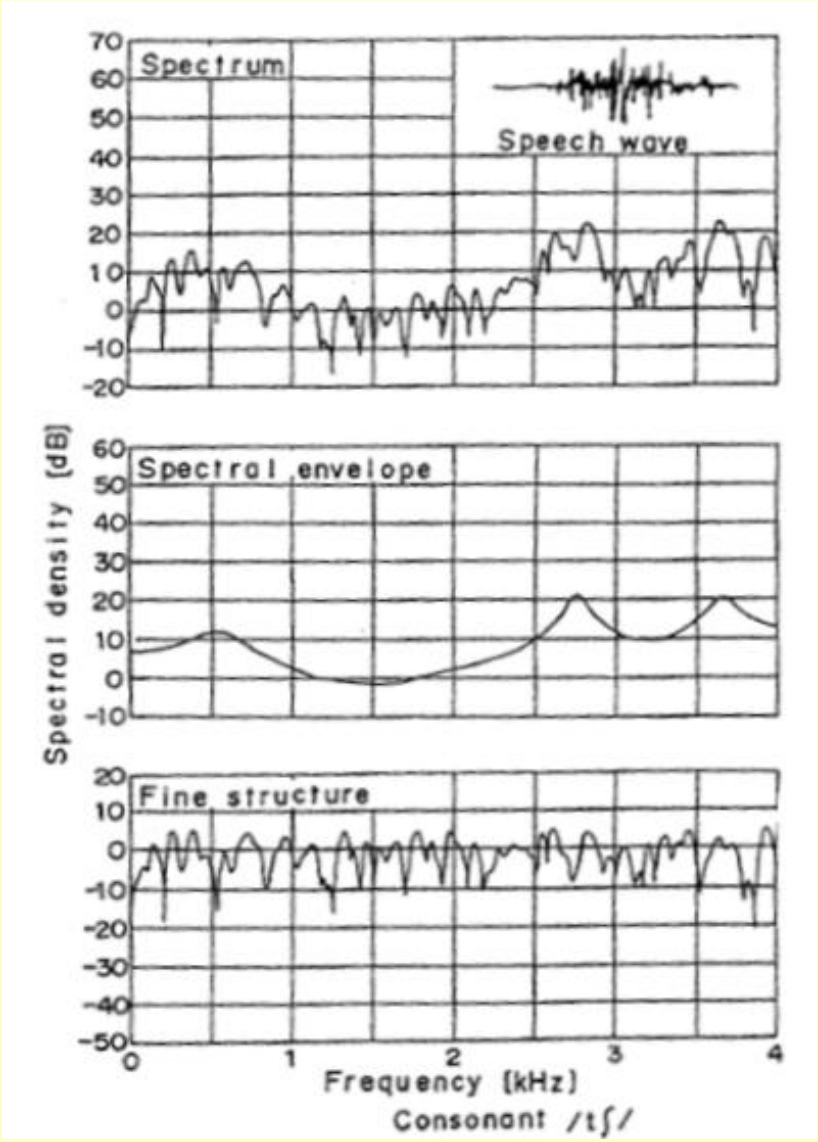
https://www.cis.rit.edu/class/simg716/Gauss_History_FFT.pdf



$$N\delta = \omega_s$$







- the short-term power spectrum is composed of the global spectral envelope and the fine structure

Short-Time Fourier Transform (STFT) - (time-frequency)

- Technique typique : Transformée de Fourier à Court Terme (TFCT)

$$X(t,\ f) = \int_R x(t+\tau)\, w(\tau)\, e^{-j2\pi f\tau}\, d\tau$$

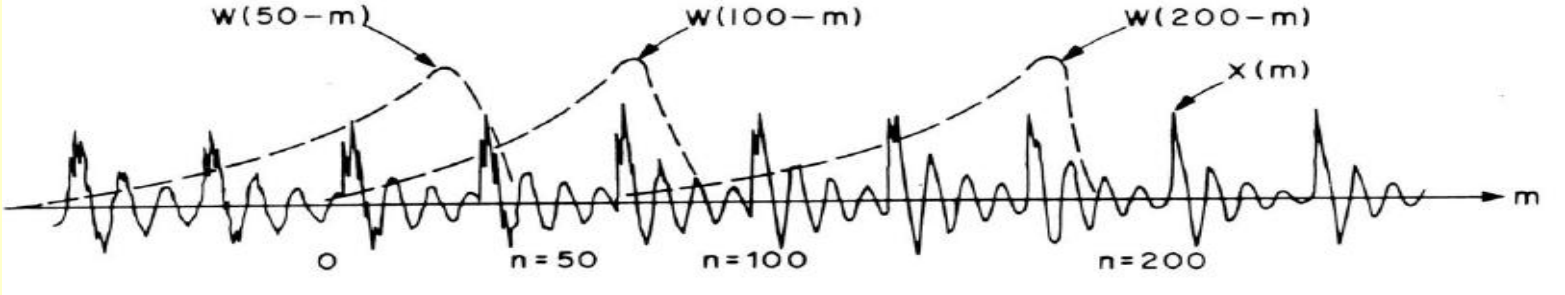
Position du bloc — Fenêtre à support fini

- En pratique : Transformée de Fourier Discrète (TFD)
  = implémentation numérique

$$X(k,m) = \sum_{n=0}^{N-1} x(k+n)\, w(n)\, e^{-j2\pi n\frac{m}{N}}$$

N échantillons ↔ *N* canaux fréquentiels

- Taille : 20 à 30 ms ↔ 320 à 480 échantillons à 16kHz
  (en pratique 256 ou 512)
- Module^2 = Densité Spectrale de Puissance (DSP) (à court terme)



- FT can't do simultaneous localization in time and frequency, not useful for time-varying signal analysis, non-stationarity

- Wide window → with good resolution in frequency and poor in time

- W (t)=1, infinitely long: → STFT transforms to FT, giving excellent frequency localization but no time localization.

- Narrow window → with good time resolution and poor frequency resolution

- W (t)=δ(t), infinitely short: → results in the signal in time (with a phase factor), giving excellent time localization, but no frequency localization.

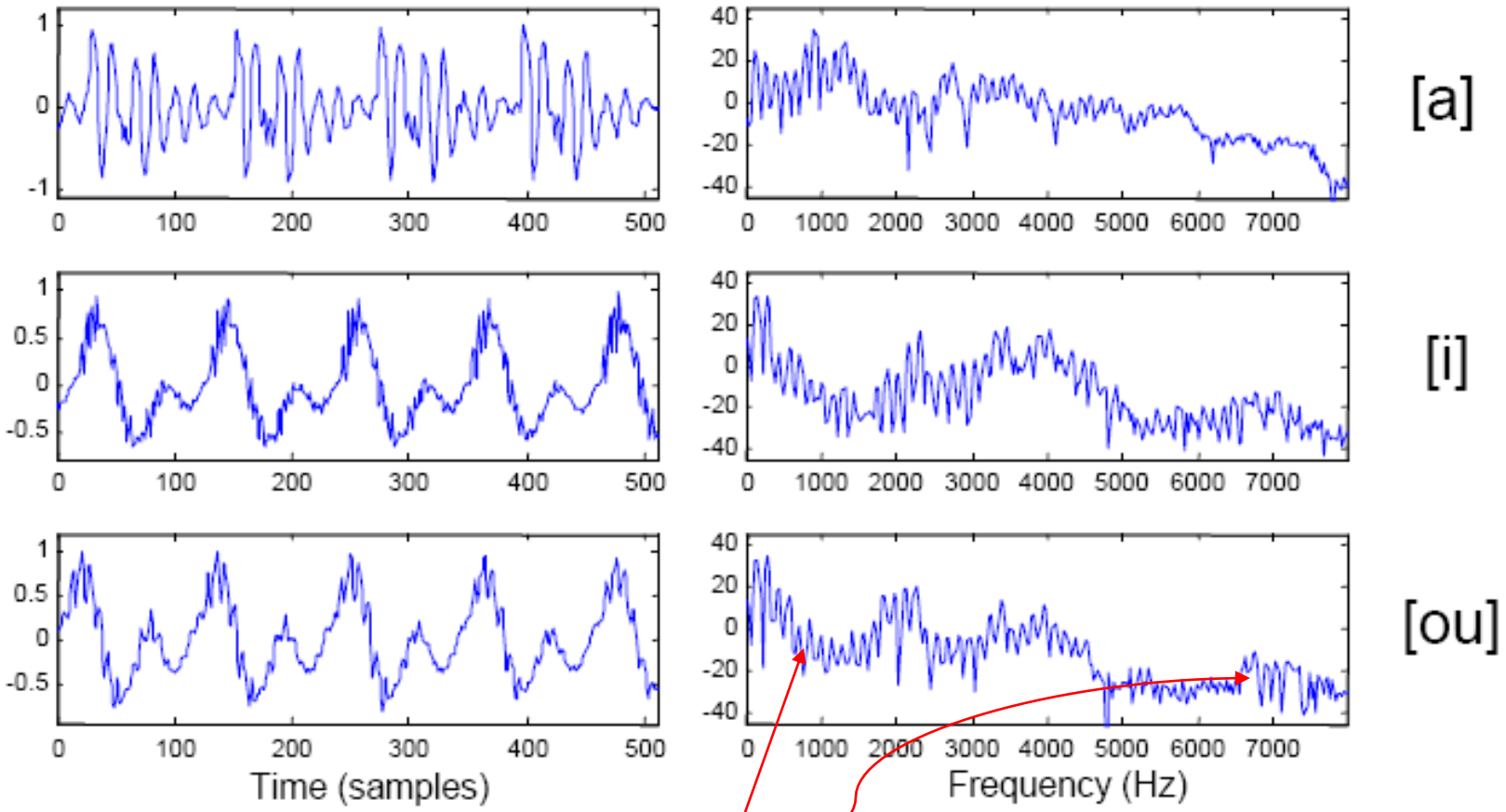
$$STFT_f^u(t',u) = \int_t \left[f(t) \cdot \delta(t-t') \right] \cdot e^{-j2\pi ut} dt = f(t') \cdot e^{-jut'}$$

Δt *and* Δf *can't be made arbitrarily small*!!

- Heisenberg's uncertainty principle

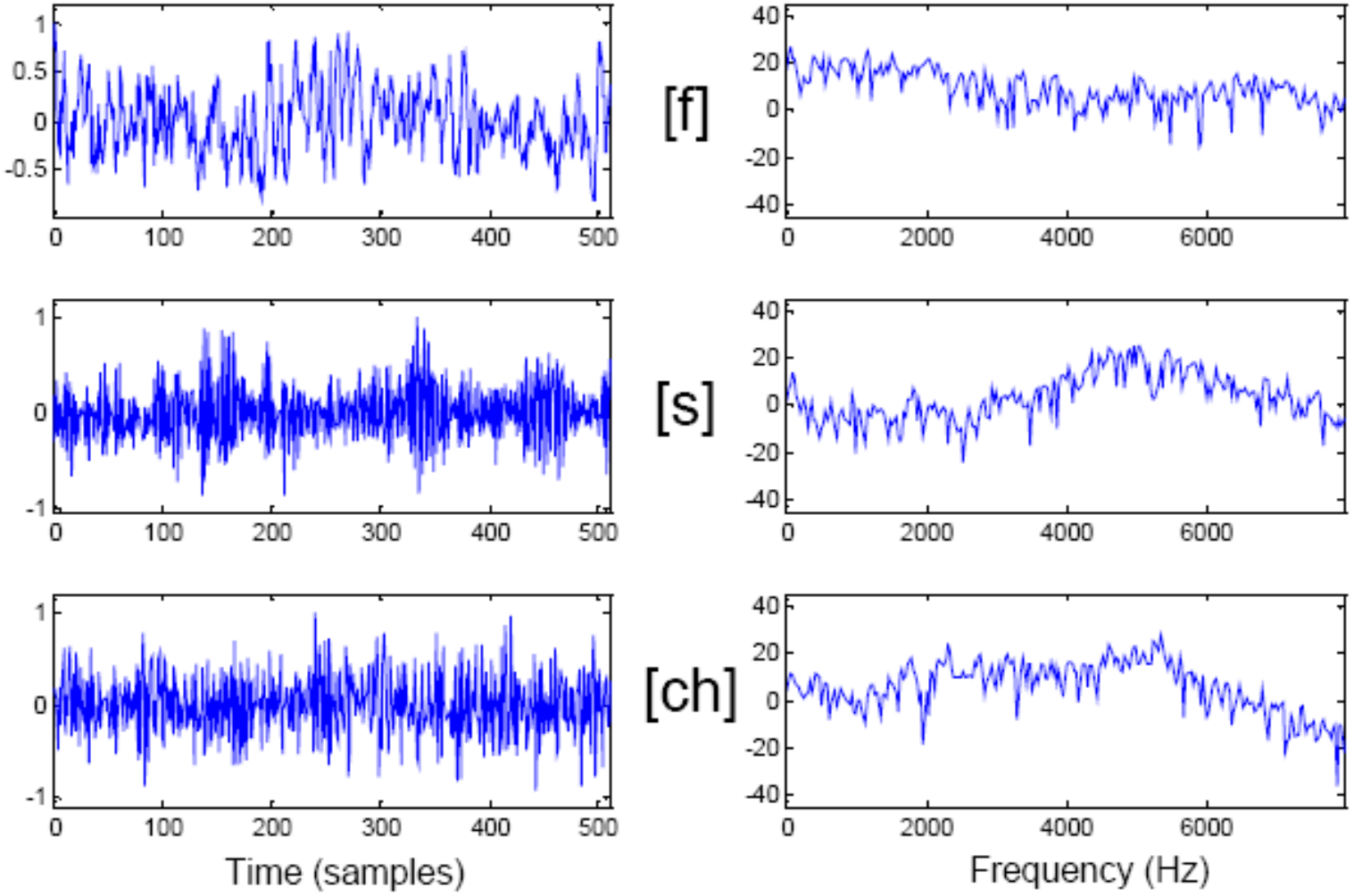
$$\Delta t \cdot \Delta f \geq \frac{1}{4\pi}$$

Exemples typiques : blocs cohérents de sons voisés (voyelles)

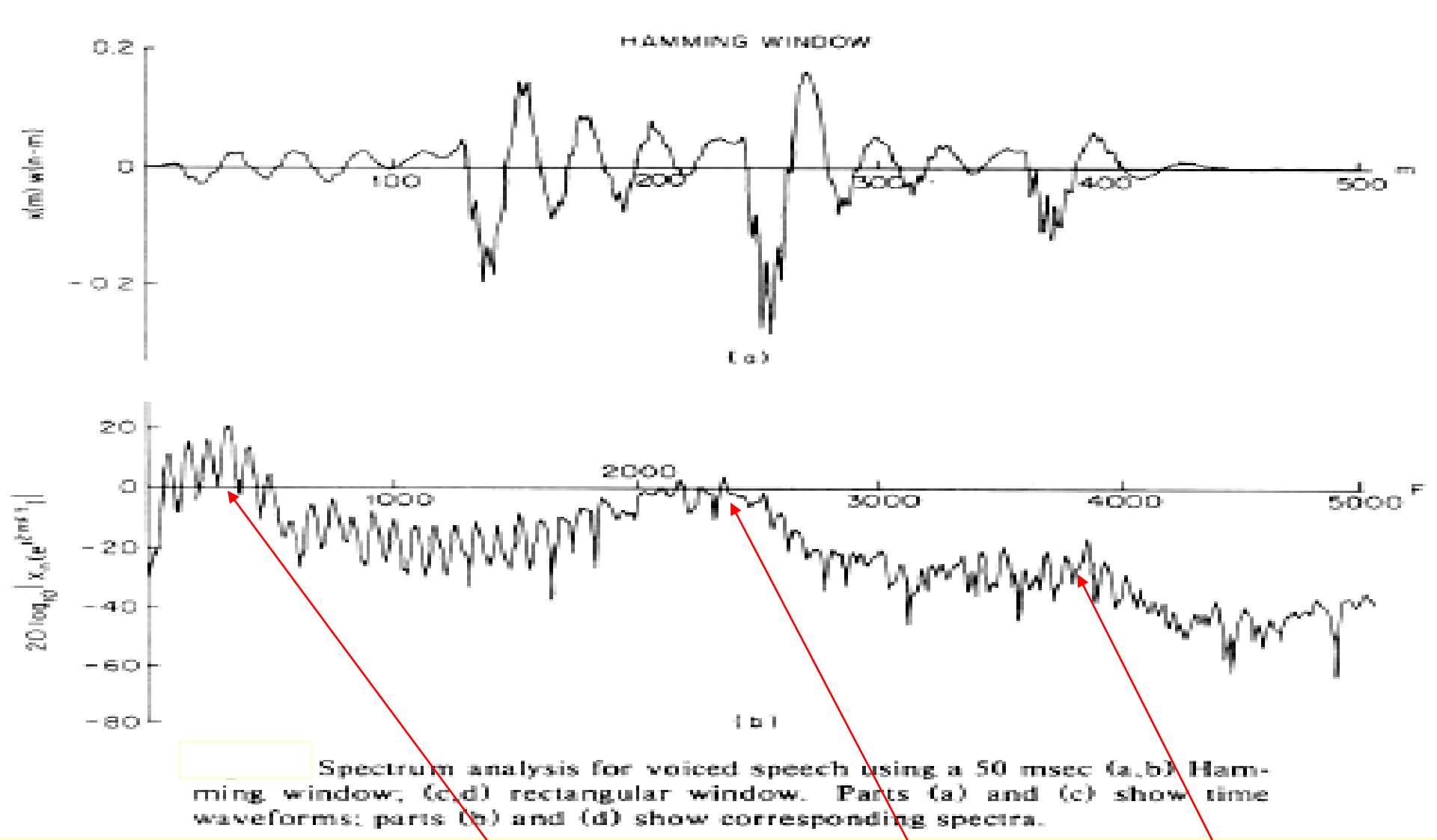


- Signaux quasi-périodiques → partie harmonique (raies) en BF + partie bruitée en HF
- Le relief de l'enveloppe spectrale caractérise les différents sons de la parole ; zones de fortes énergie = formants

Exemples typiques : blocs cohérents de sons non-voisés (fricatives)

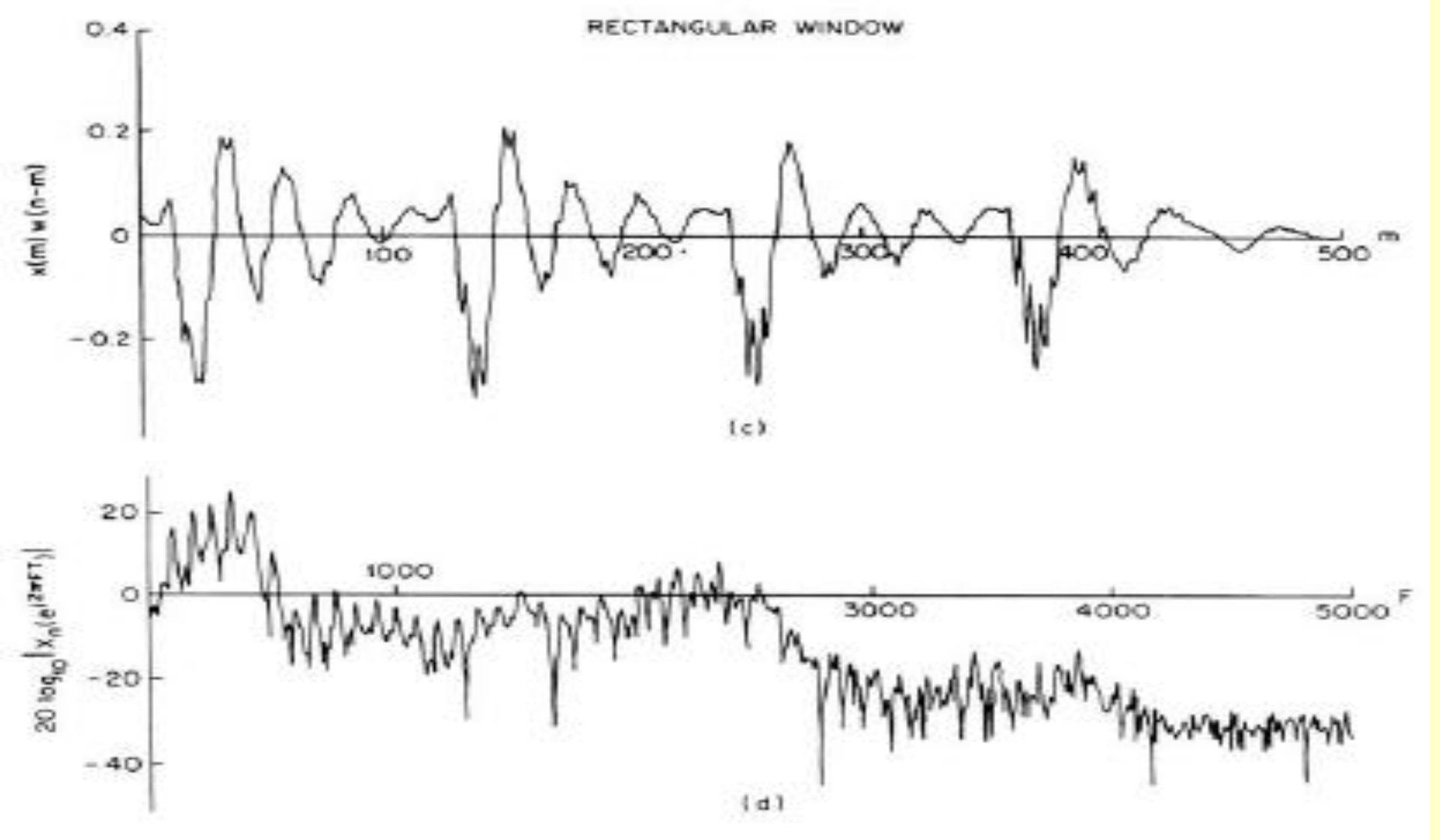


- NB : Pour des sons non-stationnaires, c'est plus délicat !



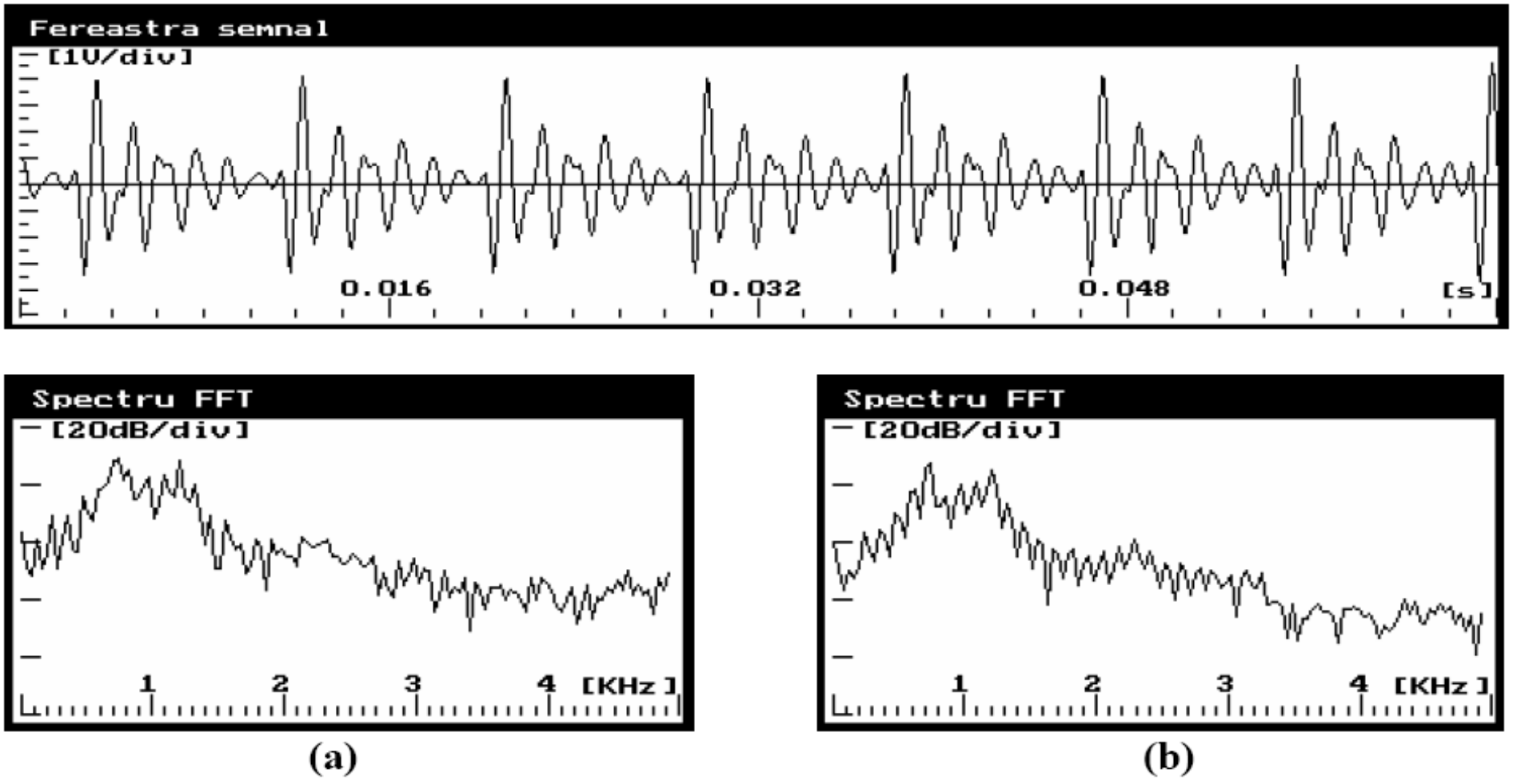
Spectrum analysis for voiced speech using a 50 msec (a,b) Hamming window; (c,d) rectangular window. Parts (a) and (c) show time waveforms; parts (b) and (d) show corresponding spectra.

- 500 samples/window (50 ms)
- a periodicity can be seen in time and frequency
- We can see the first formant (300-400 Hz), the second resonance at 2200 Hz, and the third at 3800 Hz.

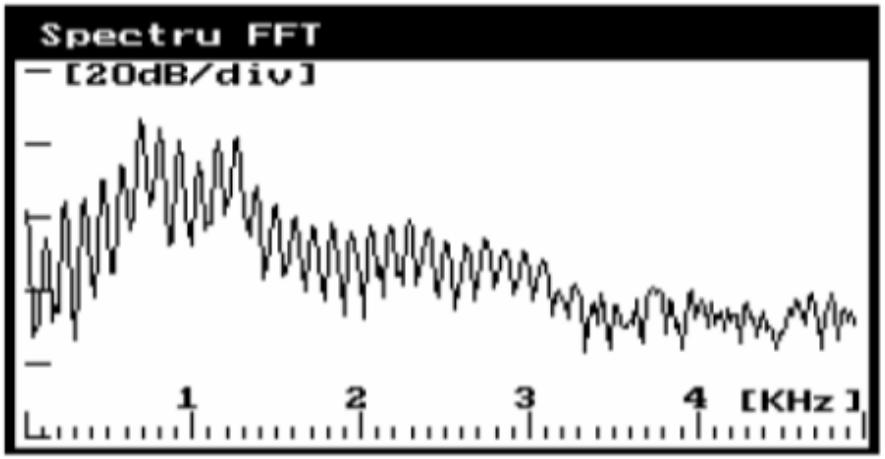


-clear evidence of FF harmonics can be observed in the case of the rectangular window (RW), due to the narrower main lobe

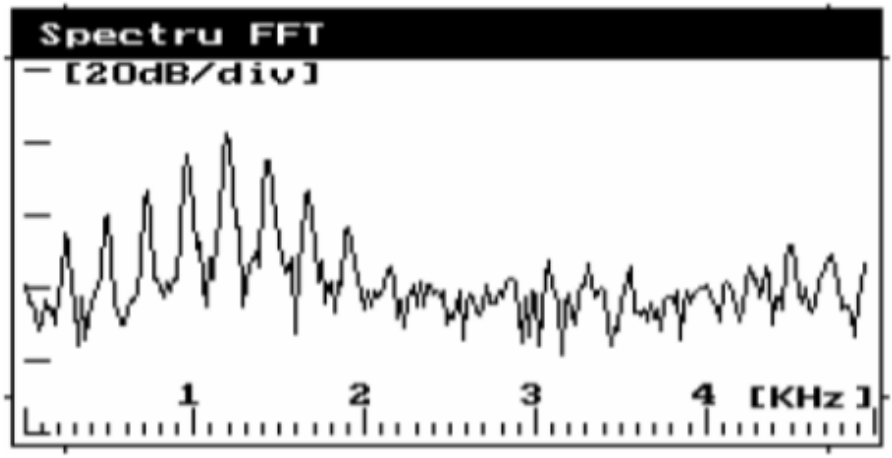
- the frequency spectrum is noisier (RW), due to inter-harmonic interference, because the R window has side lobes with only -14 dB attenuation



Frame of “a” vowel and its Fourier spectrum (256) using rectangular window (a) and Hamming window

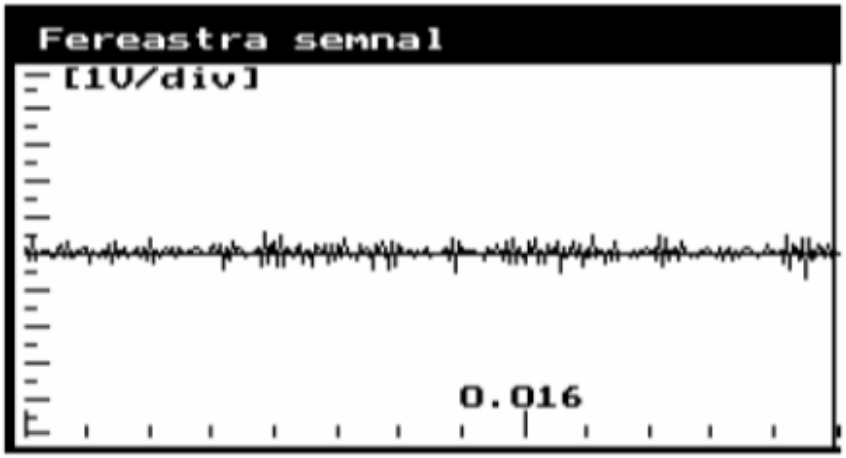


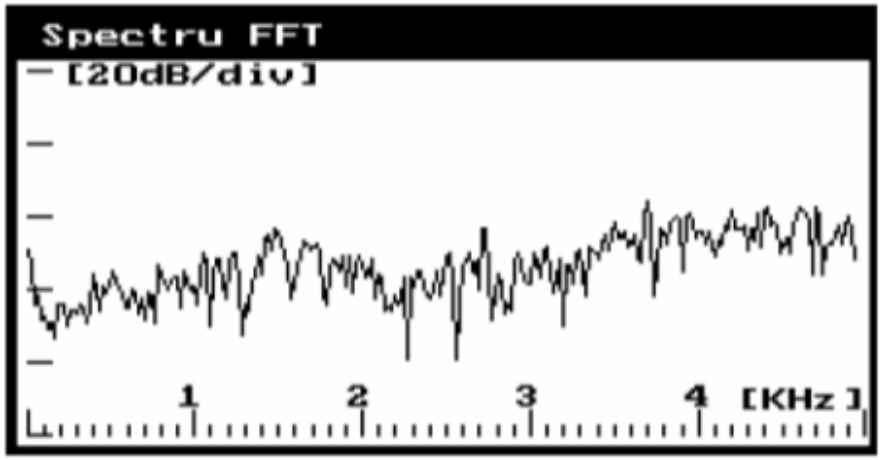
(a)



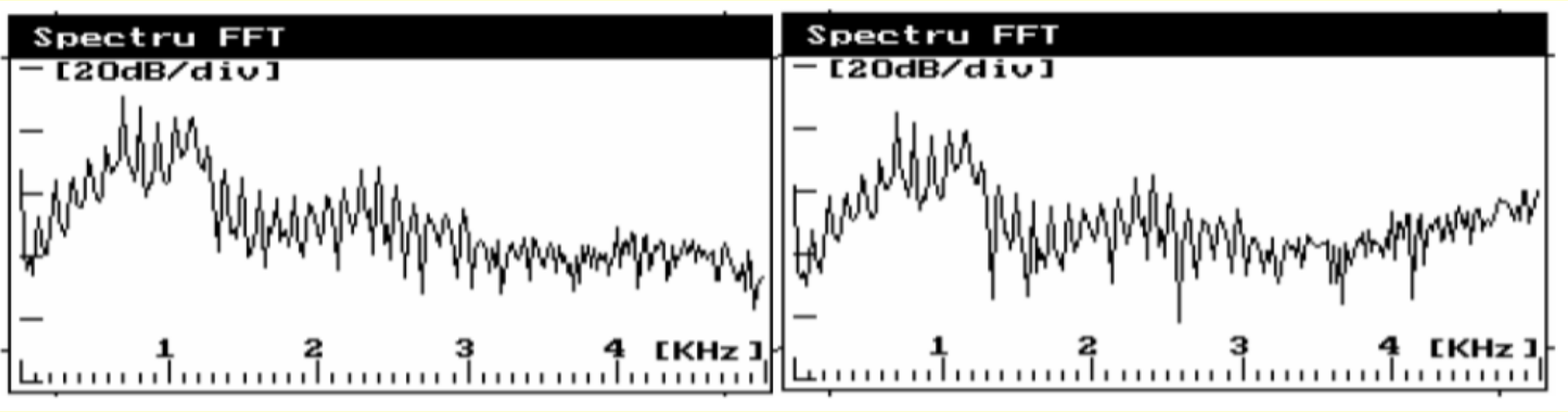
(b)

The “a” vowel Fourier spectrum (512) using Hamming window
for man-speaker (a) and women-speaker (b)



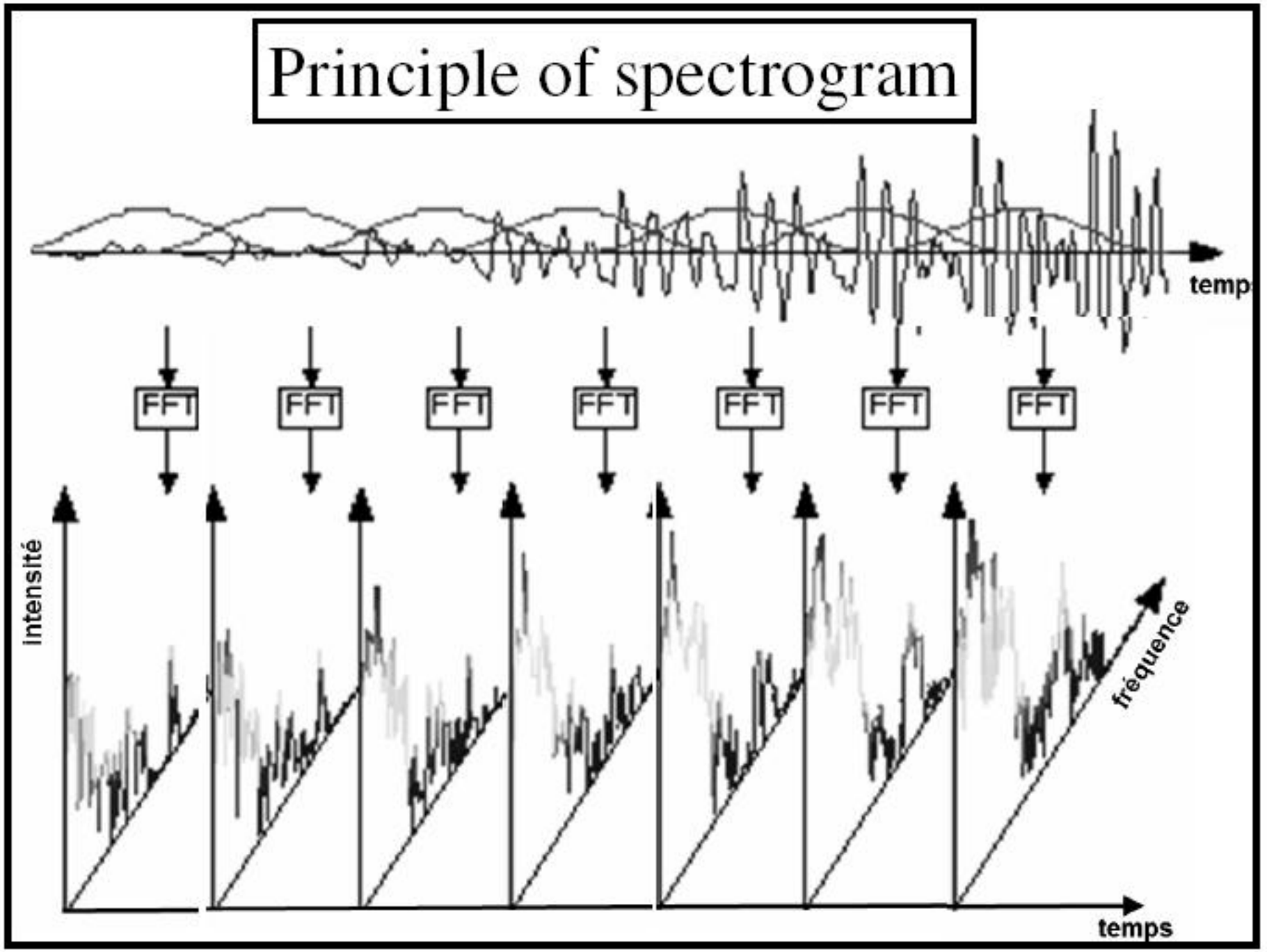


Frame of “s” consonant and its Fourier spectrum (512)



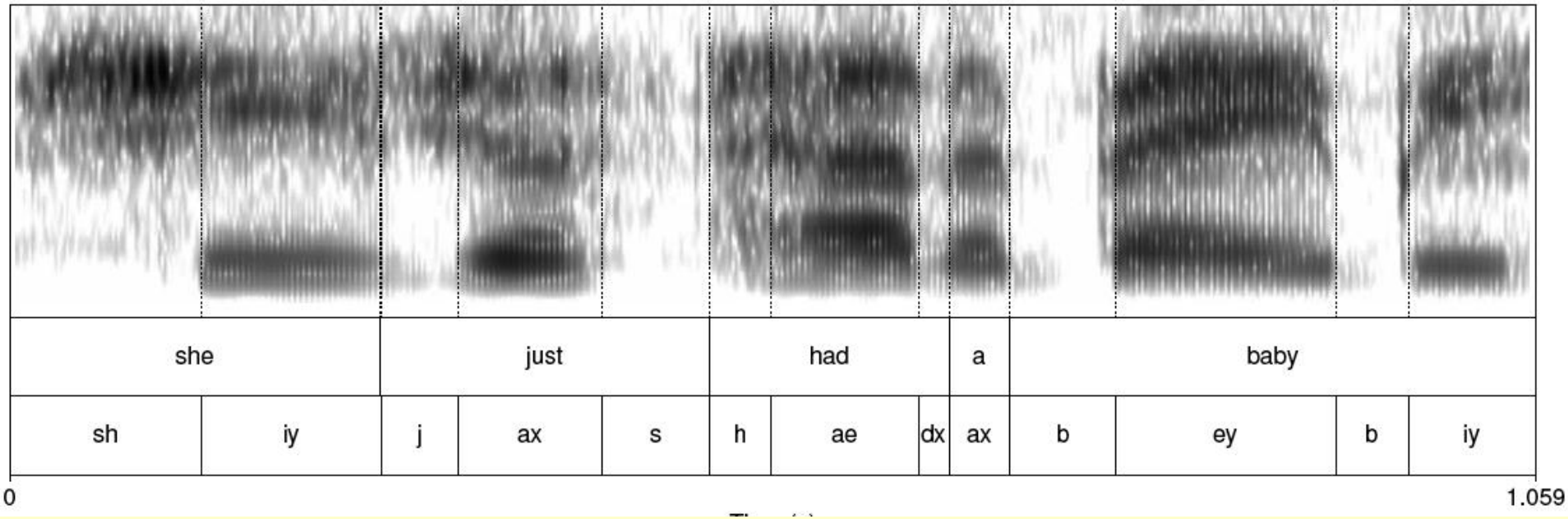
Fourier spectrum (512) of a frame of “a” vowel
with/without preemphasis

Principle of spectrogram



The Spectrogram

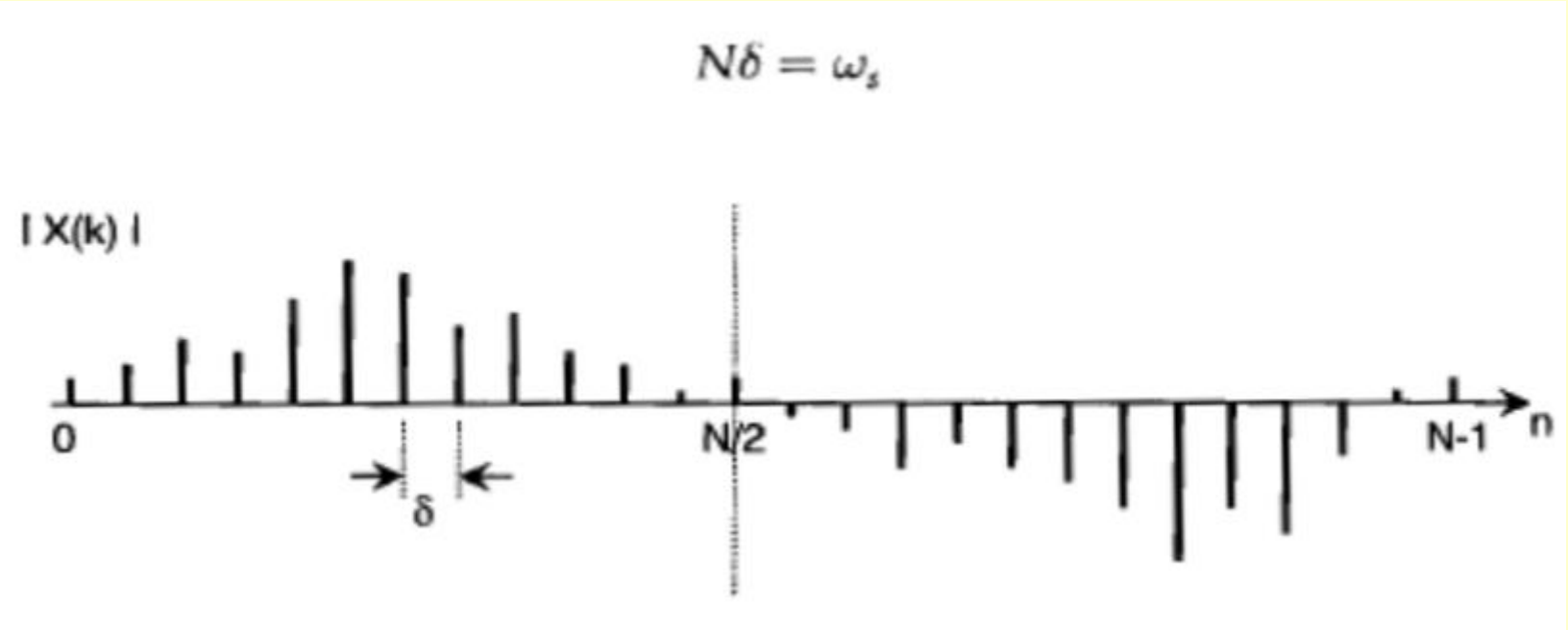
F[Hz]



t[s]

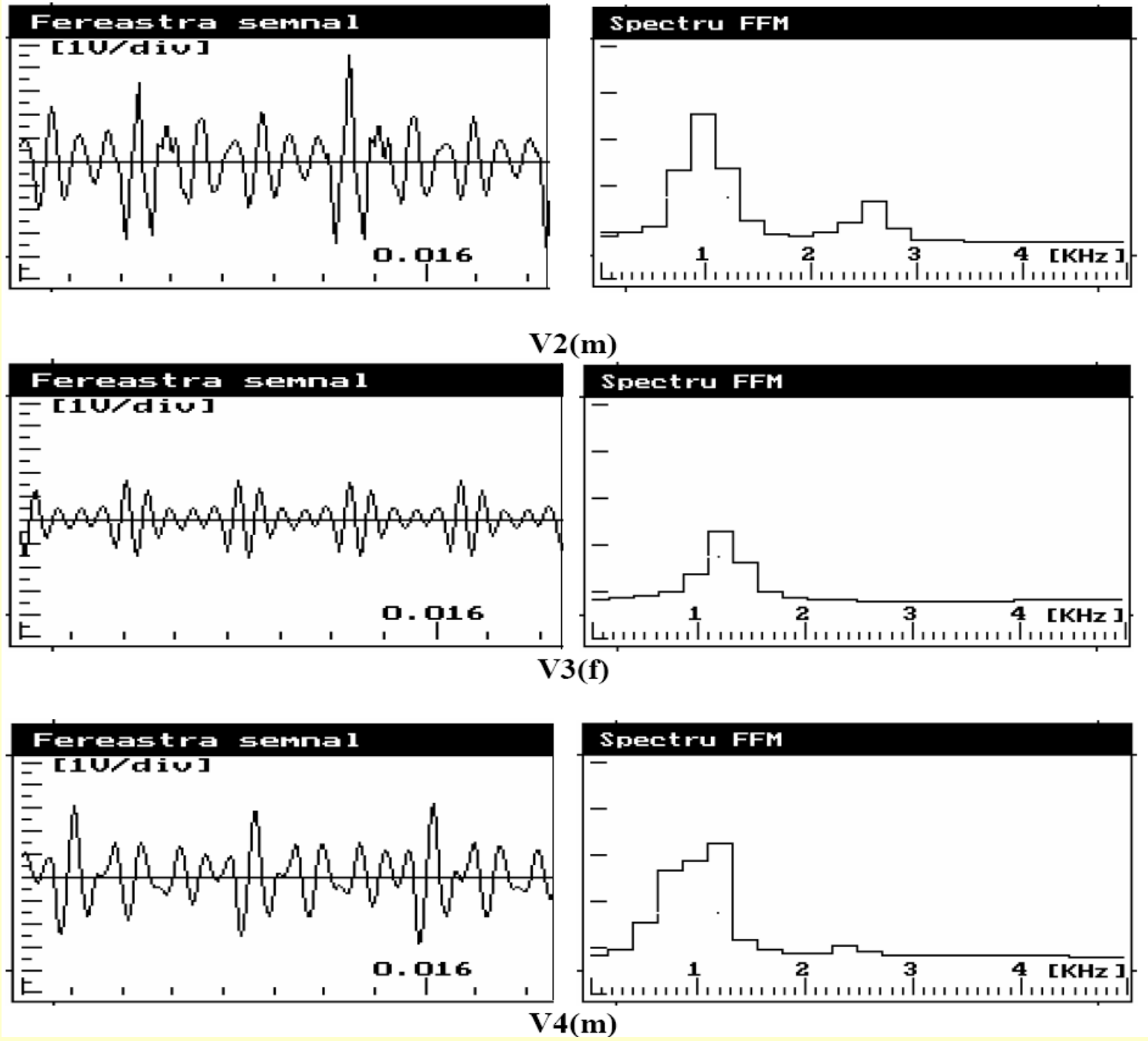
- Graphe 2D : spectre (TFCT) au cours du temps
  (x = temps, y = fréquence, couleur ou niveaux de gris = intensité)
- Limitation de la résolution temps-fréquence : spectro à large bande vs. spectro à bande étroite

3. The filter bank from FFT

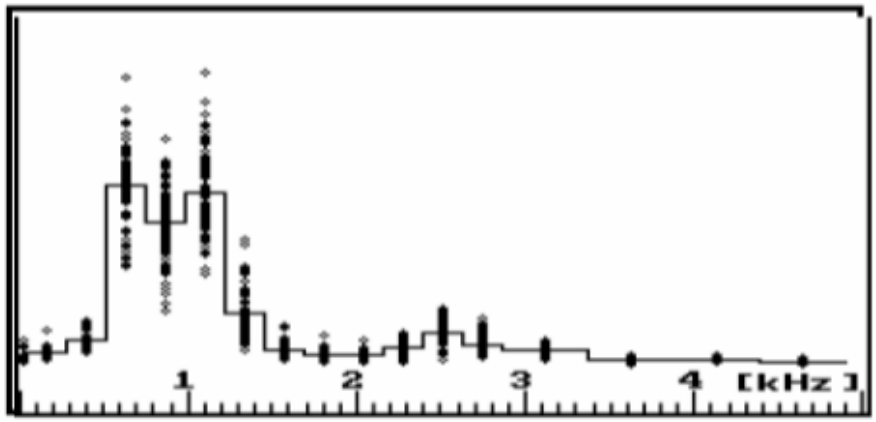


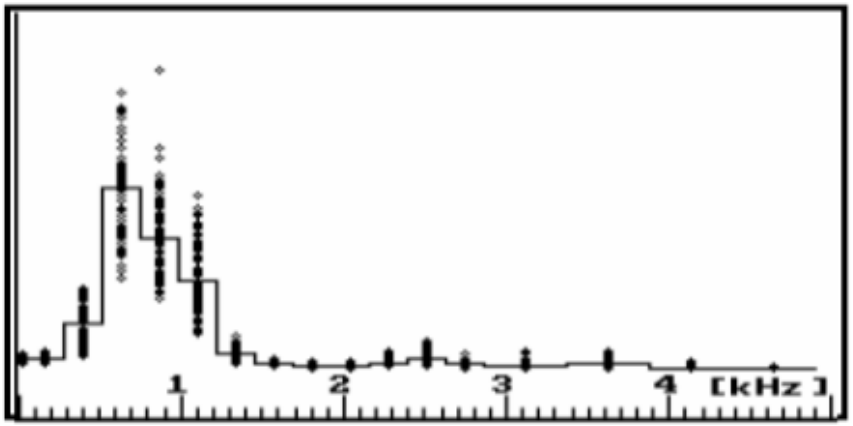
Band 1 (0-200Hz) - 4 Fourier coefficients
Bands 2-13 (200-3000Hz)- 6 Fourier coefficients/band
Bands 14-17 (3-5KHz) - 13 Fourier coefficients/band

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



"a" vowel frame and its Fourier average spectrum



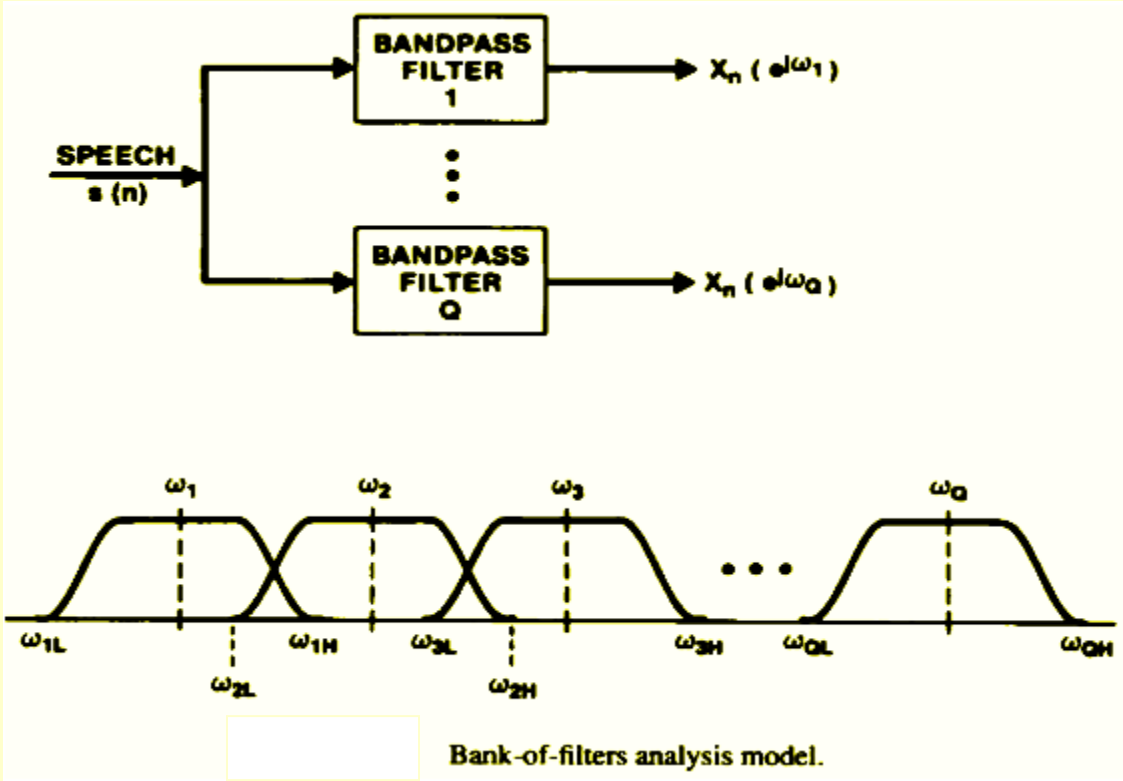


Intra-speaker variability reflected in the averaged spectrum for 100 frames obtained from the utterance of the vowel "a" by 2 male speakers

4. Speech analysis by digital filter bank

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

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



Bank-of-filters analysis model.

- SS energy is measured in certain bands
- Analog filters (Dudley 1939, Bell Labs) >>> digital filters

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

Shifts the spectrum from the band to the LF band and creates HF images

Retains the CC component and removes HF images created by non-linearities

Reduce data

Retains the CC component and removes HF images created by non-linearities

Compression Log, m-law

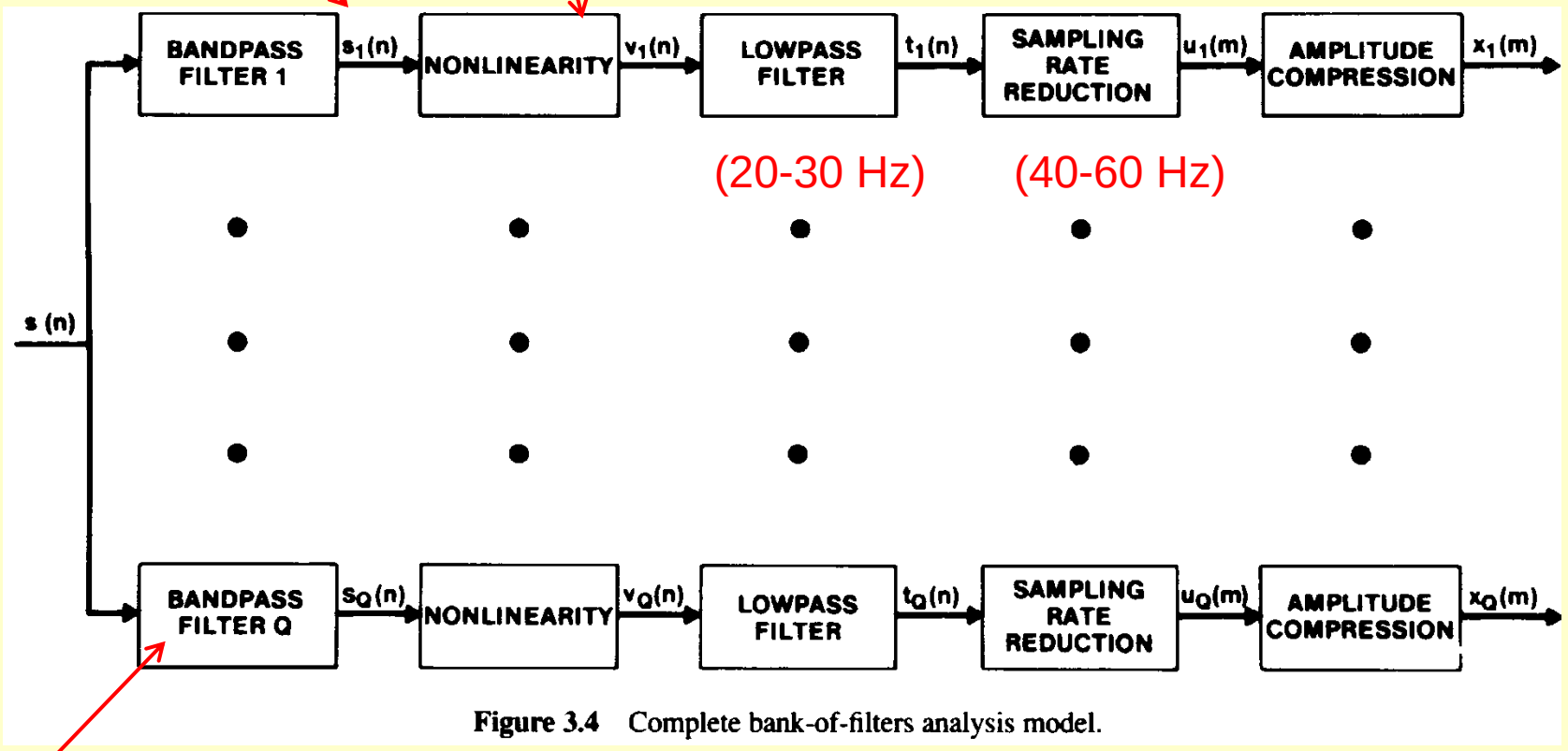
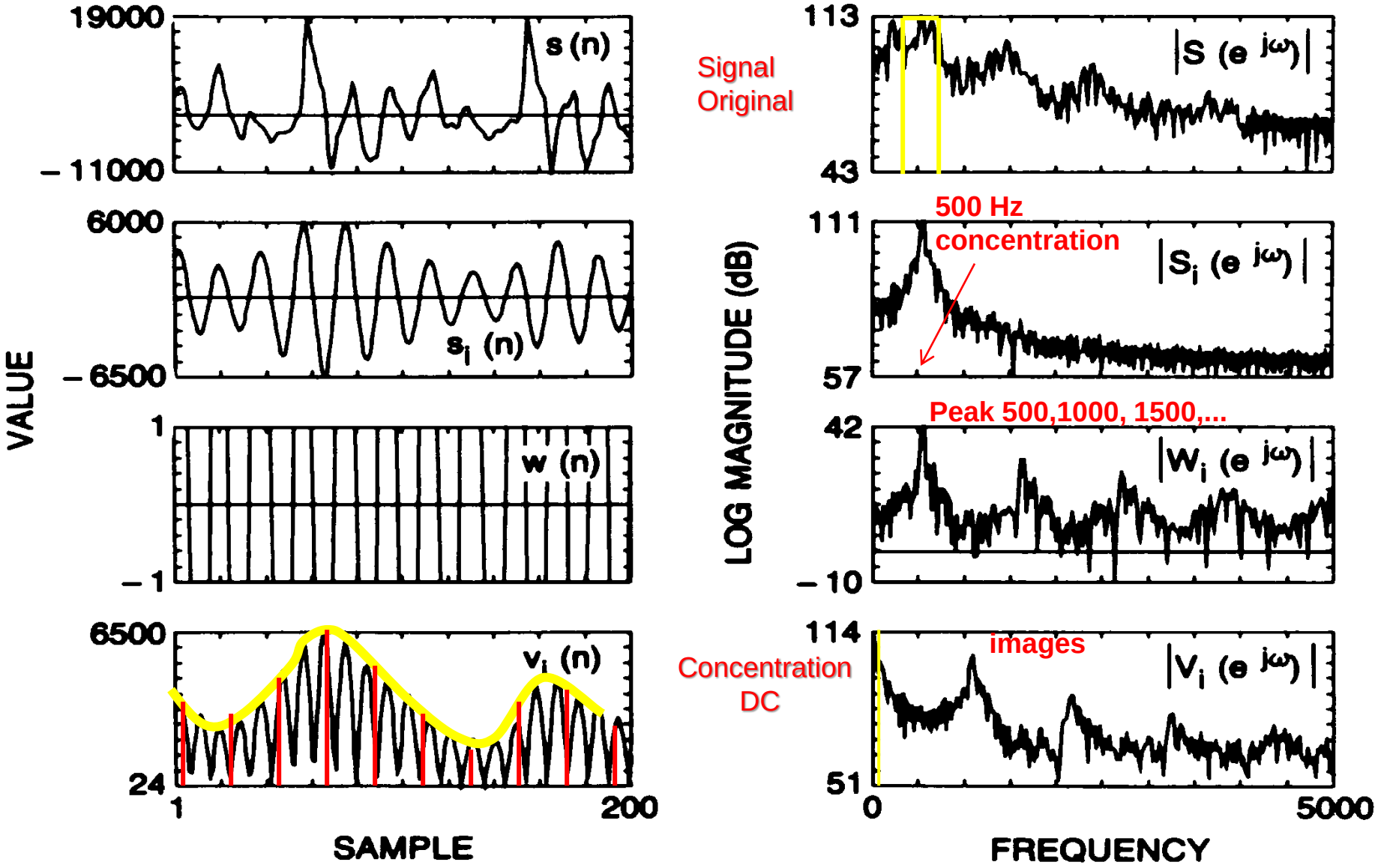


Figure 3.4 Complete bank-of-filters analysis model.

1. PBF -uniform-non/uniform (log, Mel, Bark)
2. signal rectification - FPB output >>Orig.
3. LPF – comp < 20-30Hz
4. Fes reduction(~50Hz)
5. Amplitude compression - laws A, μ



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

Uniform Filter Bank

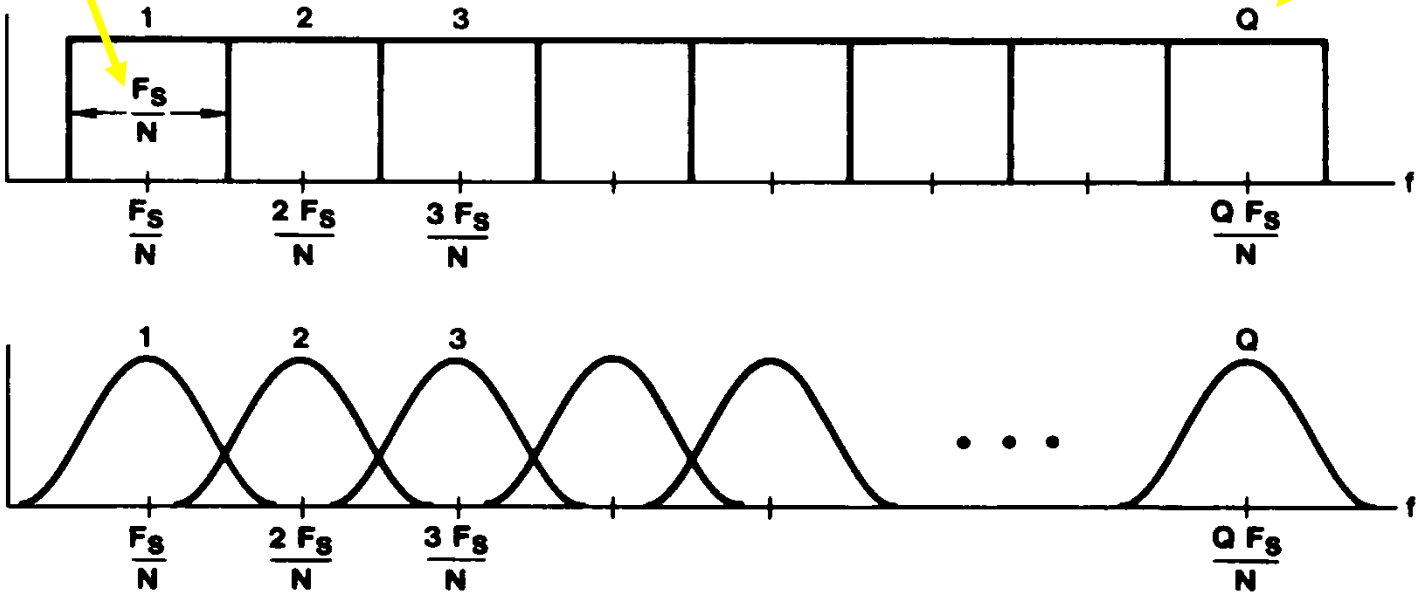
Filter Band
(no overlapping)

Central
Frequences

$$f_i = \left(F_s / N\right) i, \quad 1 \le i \le Q$$

No.
filters

$$Q = N/2$$



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 – filters are uniformly distributed over the SS frequency band

- **Non-uniform FB**
  - Frequency log scale
  - Critical band scale
  - Mel Scale
  - Bark Scale

- **Logarithmic frequency scale**

- For Q BPF, central freq., f_i and band b_i:

$$b_1 = C$$

$$b_i = \alpha b_{i-1}, \qquad 2 \le i \le Q$$

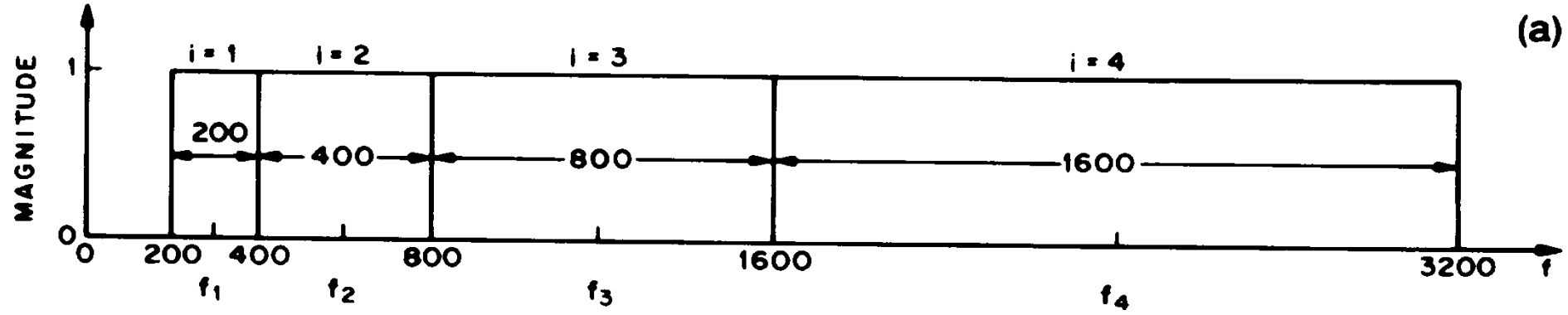
$$f_i = f_1 + \sum_{j=1}^{i-1} b_j + \frac{(b_i - b_1)}{2}$$

Arbitrary band for 1st filter

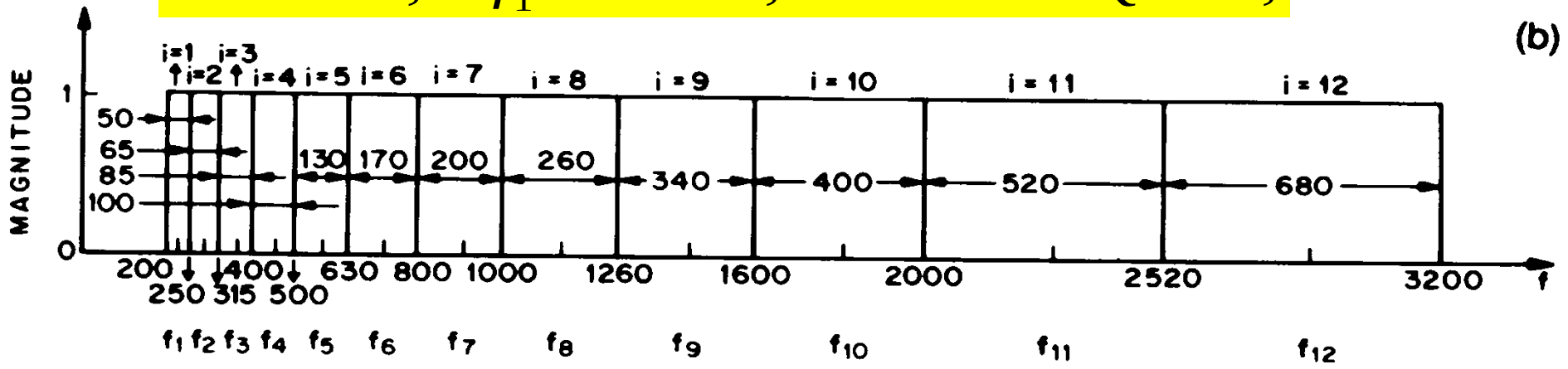
log. growth Factor usual 2

Central Frequency Arbitrarily for the 1st filter

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

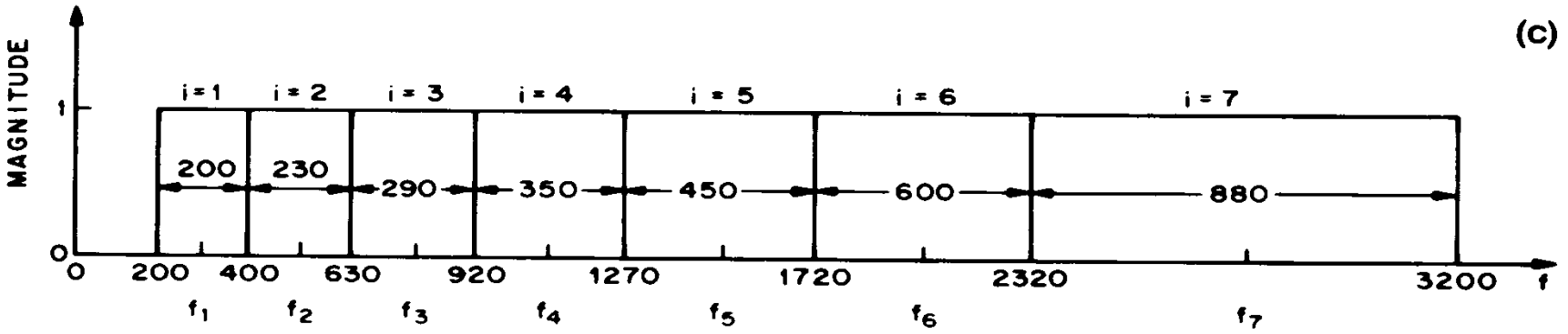


$C = 50Hz; \quad f_1 = 225Hz; \quad \alpha = 1.33 \quad Q = 12;$

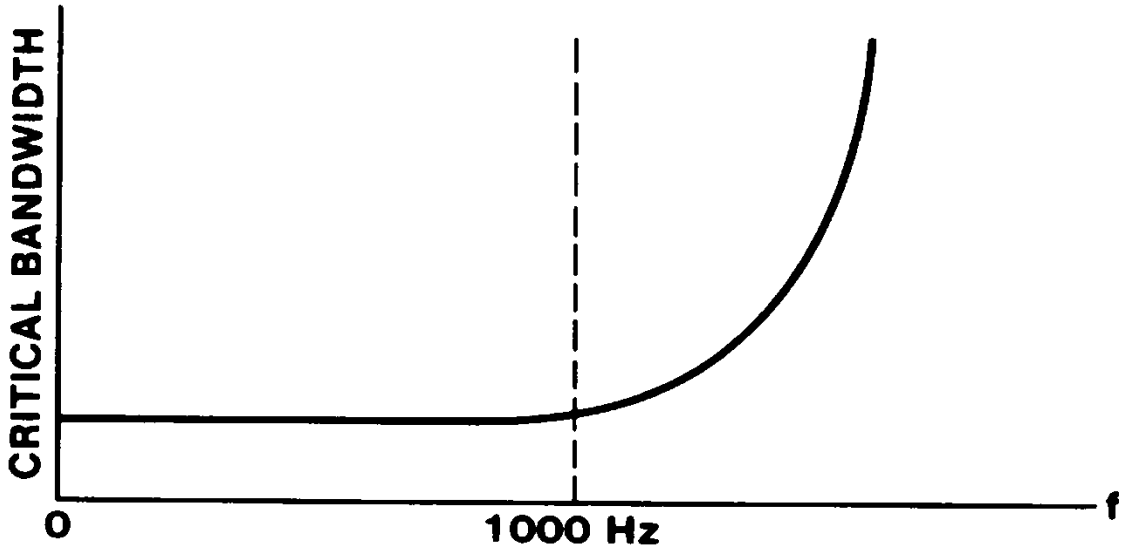


Ideal specifications of a 4-channel octave band-filter bank (a), a 12-channel third-octave band filter bank (b), and a 7-channel critical band scale filter bank (c) covering the telephone bandwidth range (200–3200 Hz).

The critical (perceptual) band scale



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).



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

33

$S_i(n)$

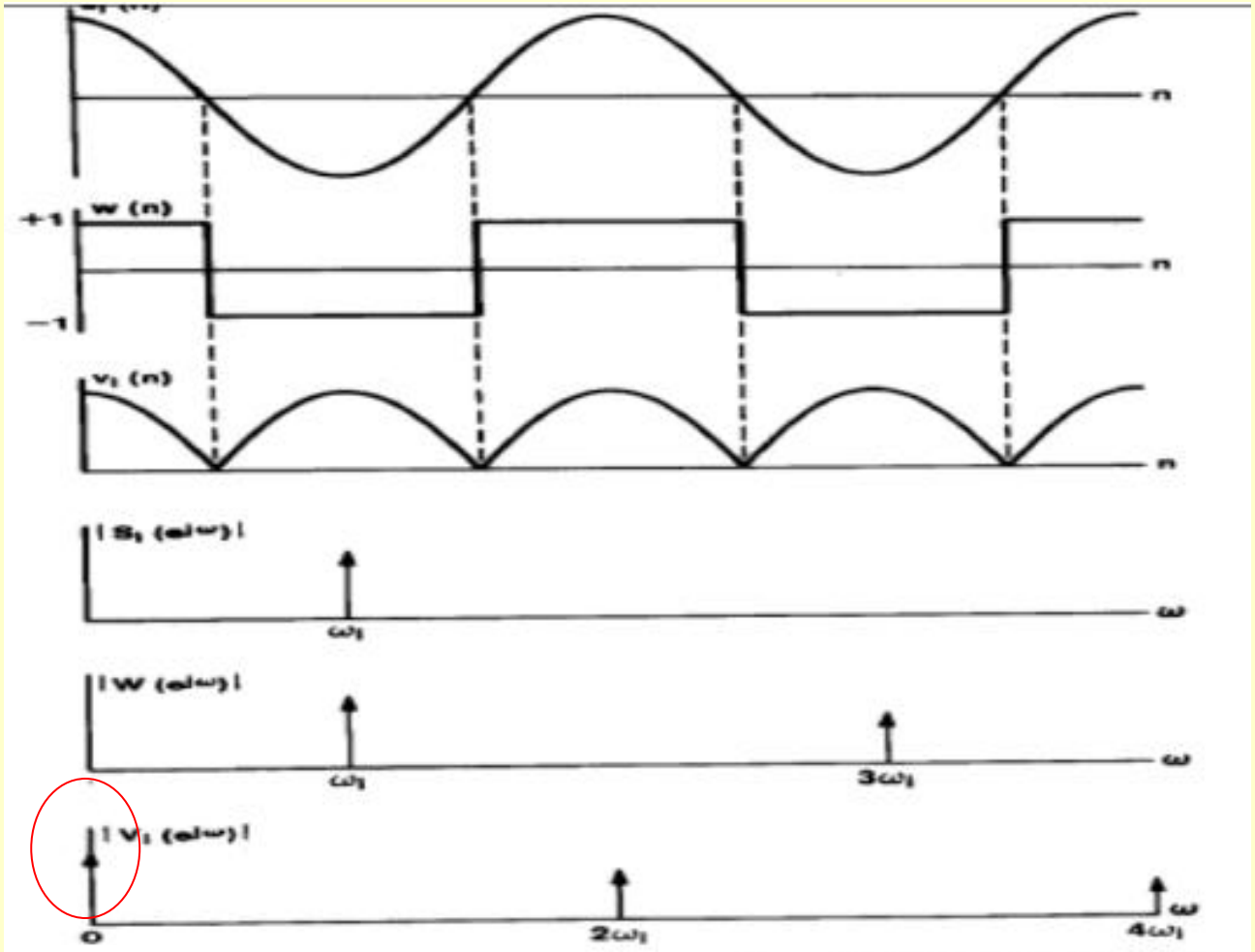
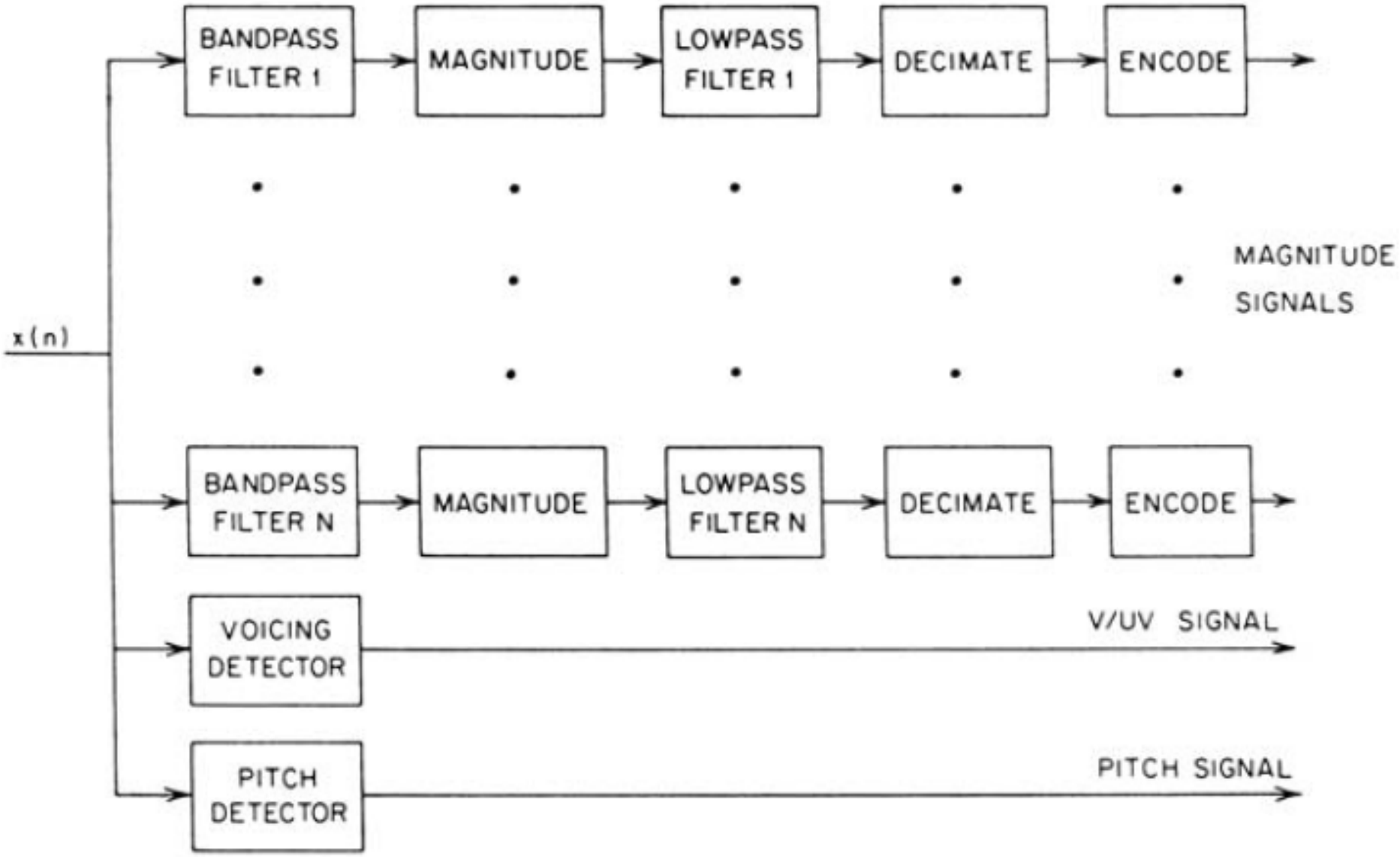
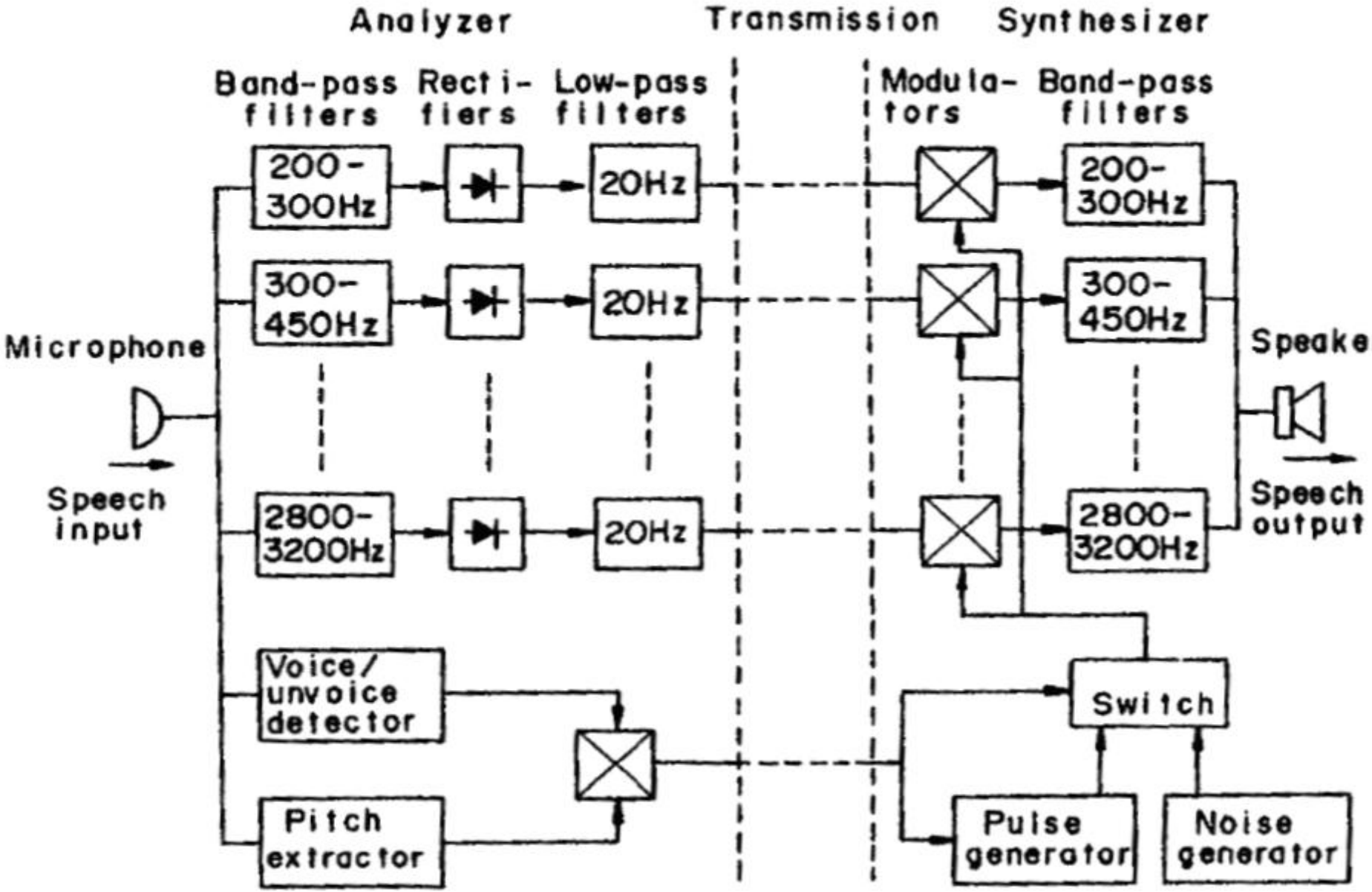


Figure 3.5 Typical waveforms and spectra for analysis of a pure sinusoid in the filter-bank model.

Ex. Simplified, for BPF$_i$



Block diagram of channel vocoder analyzer.



Structure of the (channel) vocoder.

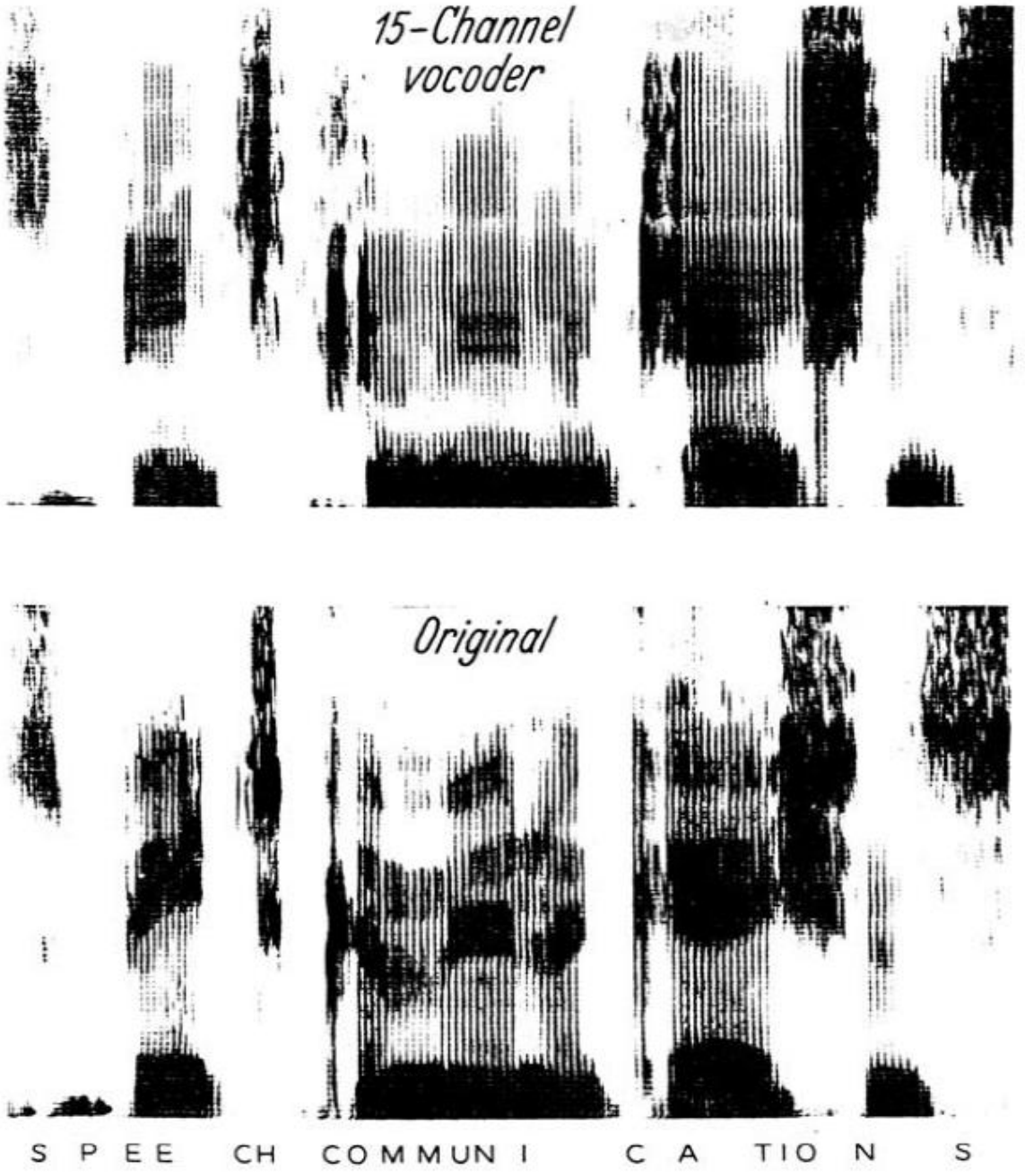


Fig. 6.63 An example of a 15 channel vocoder. (After Flanagan [31]).

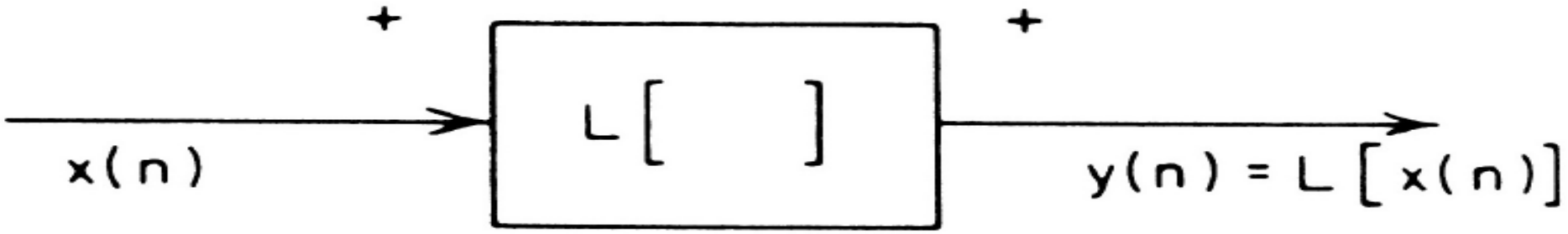
Ex. What is the compression ratio of a Q=16 PBF channel vocoder processing SS at B-band<8kHz compared to PCM coding? Let's assume that f_{es} =20kHz, and the ADC resolution is 12 bits.

5. Cepstral analysis

⇒ Speech analysis => estimate the parameters of a speech production model and measure their variations

SV = excitation * system response

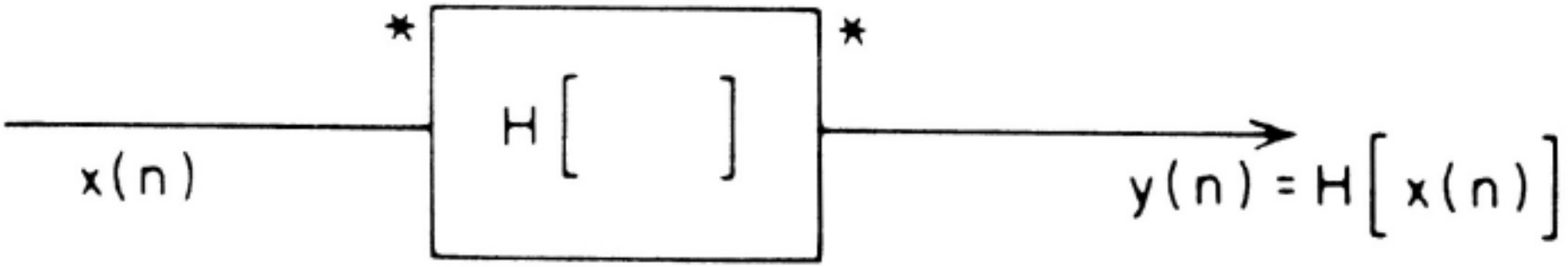
- If you want to separate the SV excitation from the vocal tract response, homomorphic filtering methods are used Linear systems respect the principle of superposition:



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

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

$\Rightarrow$ Homomorphic systems respect the principle of generalized superposition (convolution):



- For a LTI system :

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

The principle of "generalized" superposition replaces + by convolution * :

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

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

- Homomorphic filtering => homomorphic system [H], which allows the desired signal to pass unaltered and stops the unwanted signal.

$$x(n) = x_1(n) * x_2(n)$$ *x1 - unwanted*

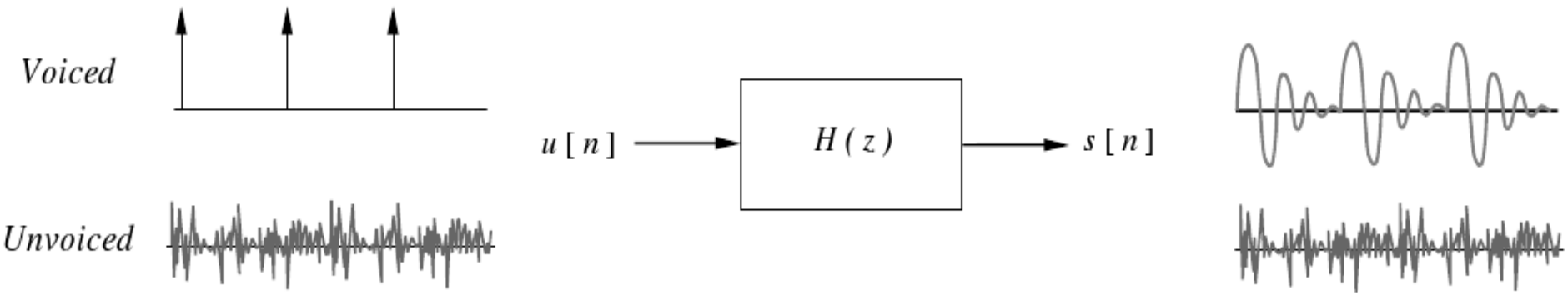
$$H\left[x(n)\right] = H\left[x_1(n)\right] * H\left[x_2(n)\right]$$

$$H\left[x_1(n)\right] \rightarrow \delta(n)$$ - removal of $x_1(n)$

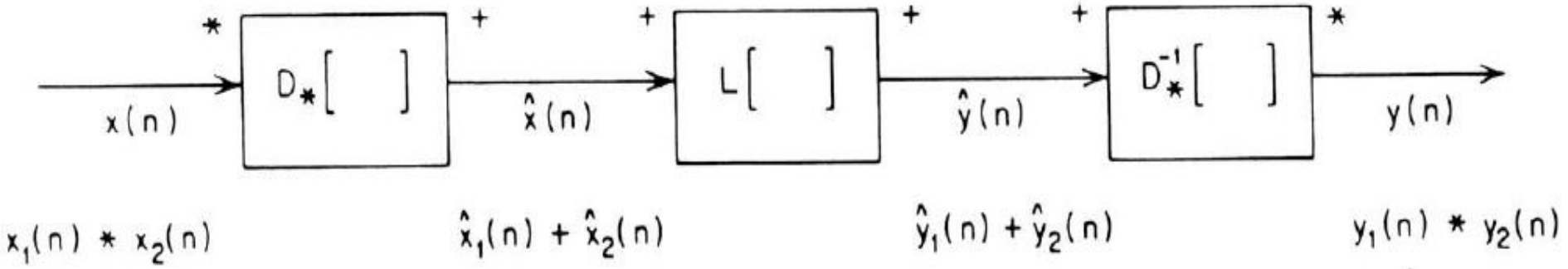
$$H\left[x_2(n)\right] \rightarrow x_2(n)$$

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

- For linear systems, we can draw an analogy with additive noise elimination



The canonical form of homomorphic convolution



- Any homomorphic system can be represented as three cascaded systems for convolution

1. The system takes the convolution input and combines it to make additive outputs
2. System is a classical linear system.
3. 3rd is the inverse of the first system - additive inputs into convolution outputs

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

- Convolution

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

- Additive relationship

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

- Linear system

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

- Inverse convolution relation

The frequency canonical form

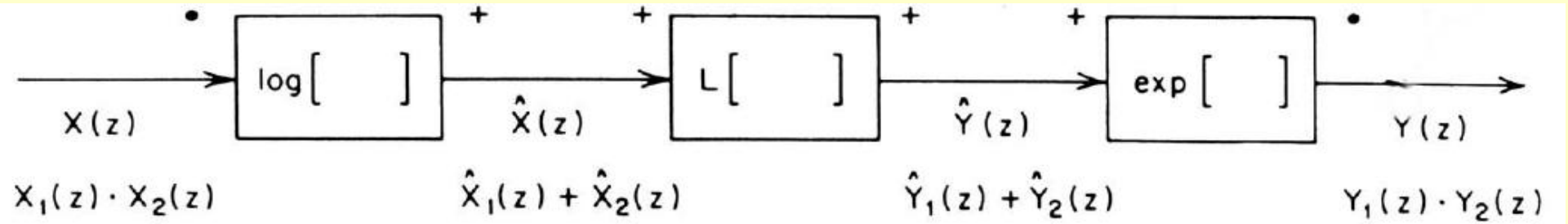
- It is found a system that transforms the convolution into a sum

$$x(n) = x_1(n) * x_2(n)$$
$$X(z) = X_1(z) \cdot X_2(z)$$

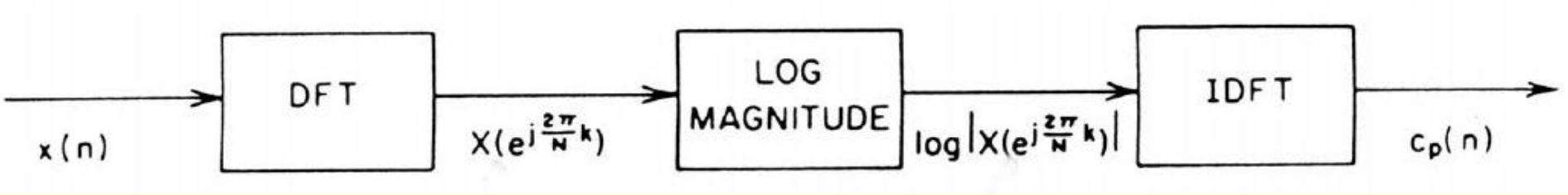
$$D_*\left[x(n)\right] = \hat{x}_1(n) + \hat{x}_2(n) = \hat{x}(n)$$
$$D_*\left[X(z)\right] = \hat{X}_1(z) + \hat{X}_2(z) = \hat{X}(z)$$

- The logarithm function transforms the product into a sum

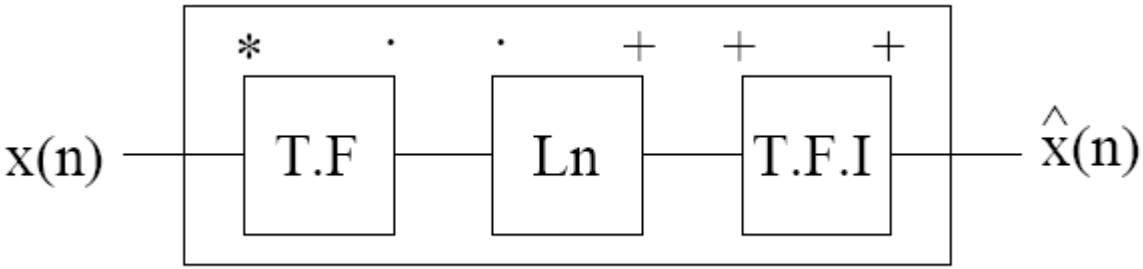
$$\hat{X}(z) = \log\left[X(z)\right] = \log\left[X_1(z) \cdot X_2(z)\right]$$
$$= \log\left[X_1(z)\right] + \log\left[X_2(z)\right] = \hat{X}_1(z) + \hat{X}_2(z)$$
$$\hat{Y}(z) = L\left[\hat{X}_1(z) + \hat{X}_2(z)\right] = \hat{Y}_1(z) + \hat{Y}_2(z)$$
$$Y(z) = \exp\left[\hat{Y}_1(z) + \hat{Y}_2(z)\right] = Y_1(z) \cdot Y_2(z)$$

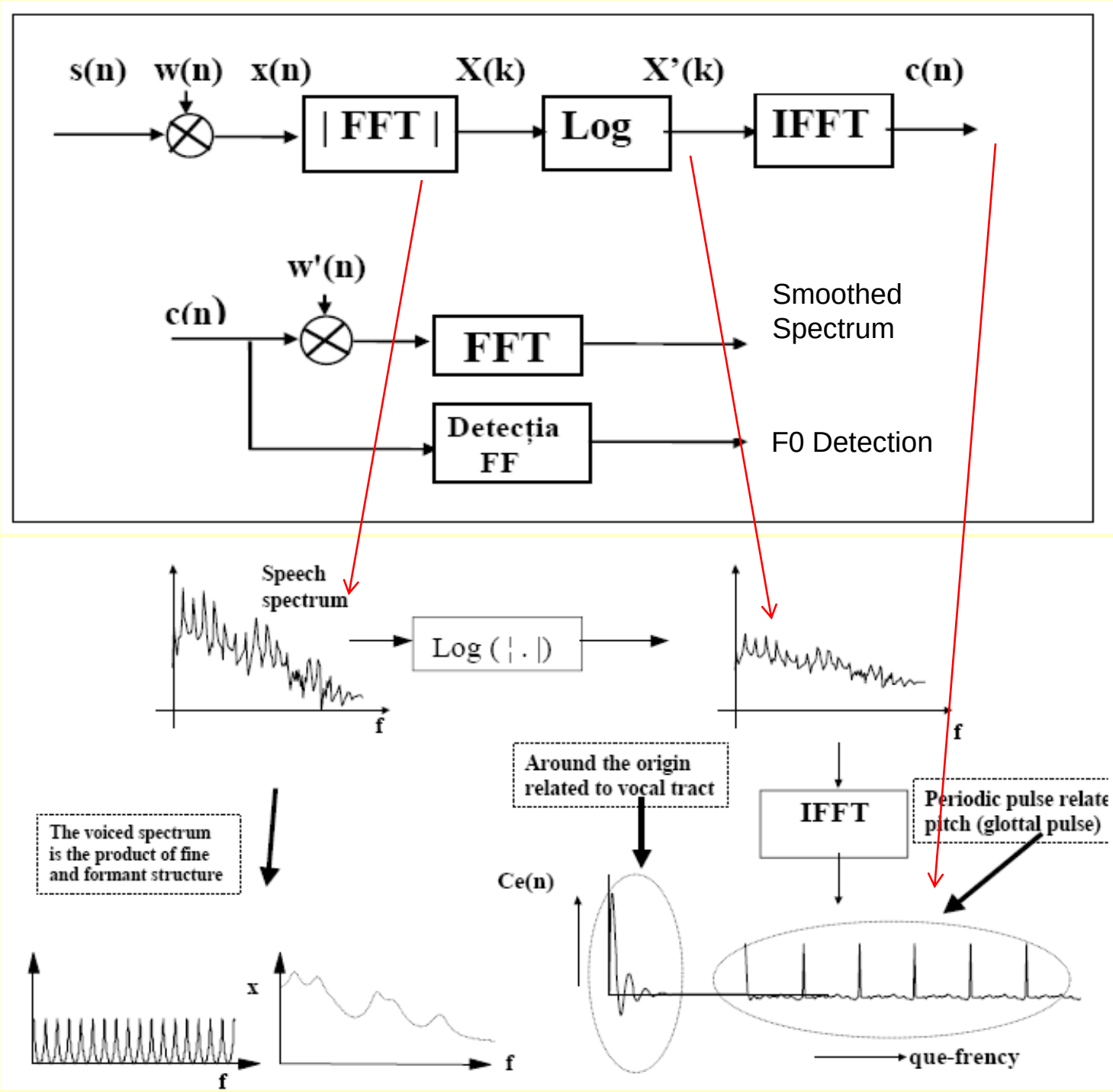


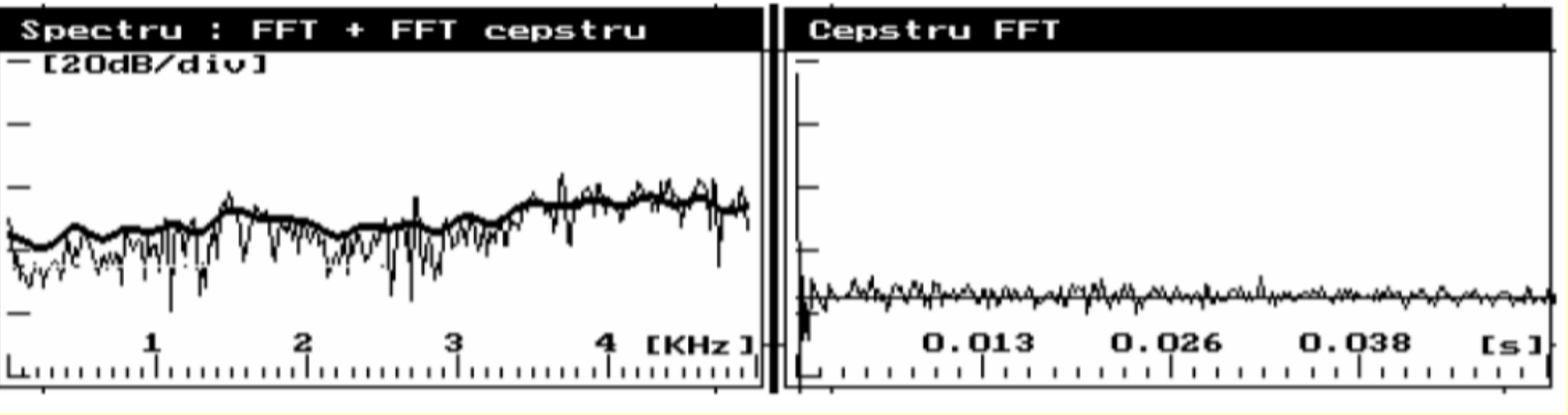
Cepstrum for SS



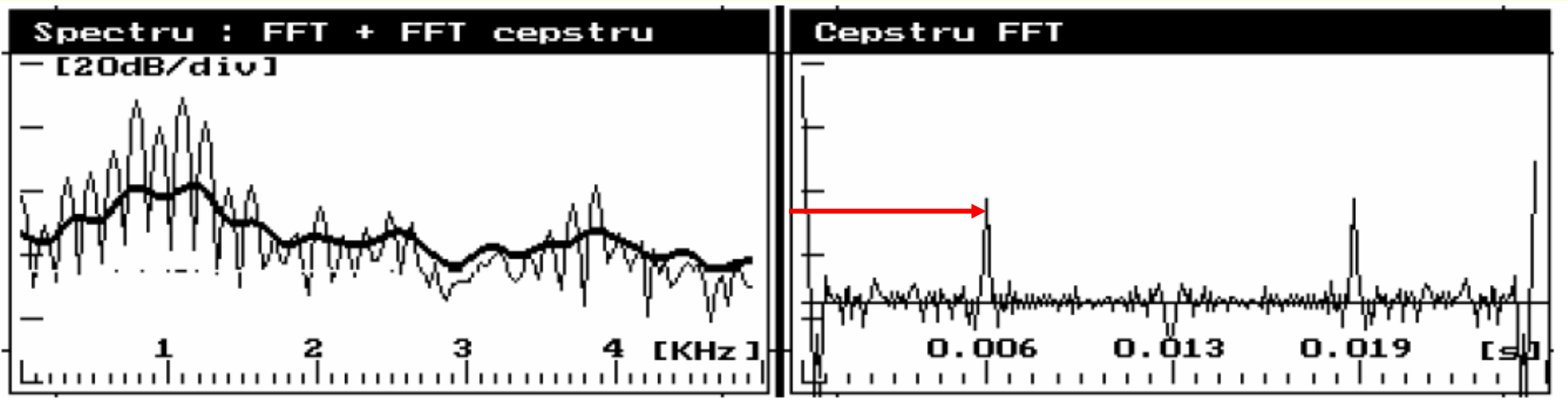
- Signal $x(n) = x_1(n) * x_2(n)$
- Transformée de Fourier (pour passer de la convolution à une multiplication) $X(\omega) = X_1(\omega)X_2(\omega)$
- Logarithme
  $\hat{X}(\omega) = ln[X(\omega)] = ln[X_1(\omega)] + ln[X_2(\omega)] = \hat{X}_1(\omega) + \hat{X}_2(\omega)$
- Transformée de Fourier inverse. Le signal revient dans le domaine temporel mais il reste additif. $\hat{x}(n) = \hat{x}_1(n) + \hat{x}_2(n)$



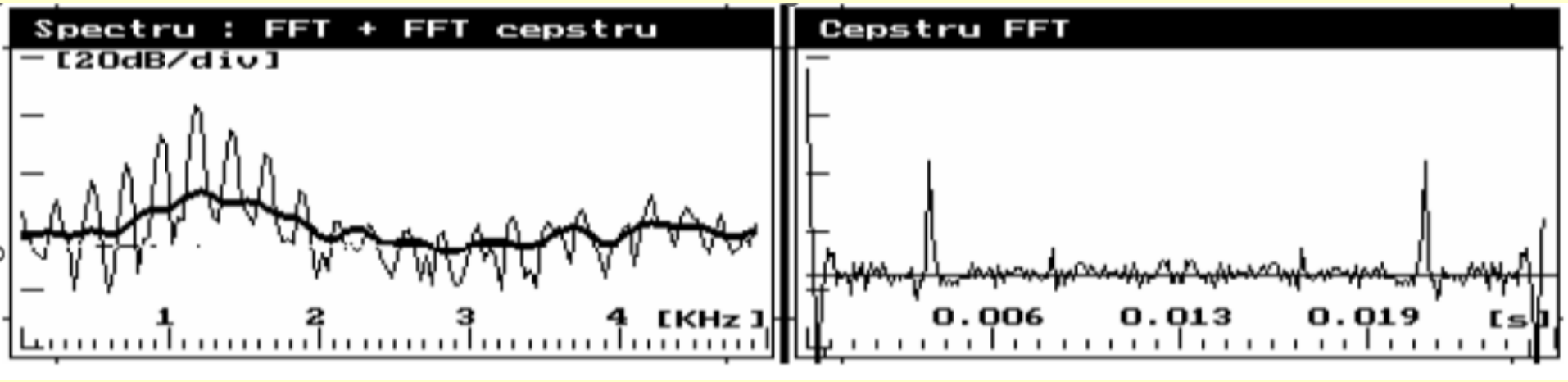

s(n)
w(n)
x(n)
X(k)
X'(k)
c(n)
| FFT |
Log
IFFT
w'(n)
c(n)
FFT
Smoothed Spectrum
Detecția FF
F0 Detection
Speech spectrum
Log (| . |)
f
f
Around the origin related to vocal tract
IFFT
Periodic pulse relate pitch (glottal pulse)
The voiced spectrum is the product of fine and formant structure
Ce(n)
x
f
f
que-frency



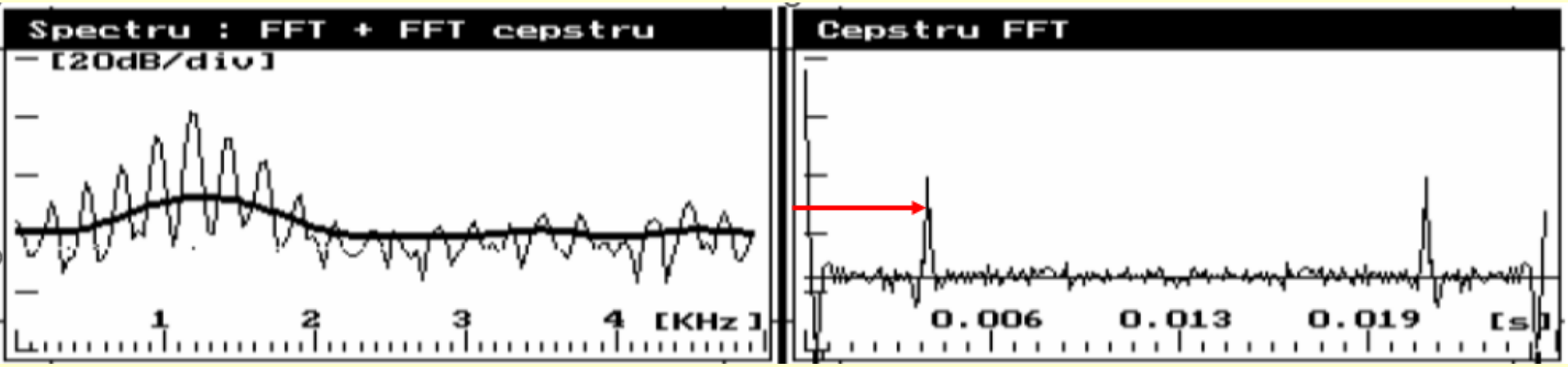
The FFT cepstrum and the FFT spectrum for an unvoiced frame (“s”)



The FFT cepstrum and the FFT spectrum for a voiced frame (“a”)



FFT cepstrum and its Fourier and smoothed spectra for a speech frame ("a") (256) uttered by a woman, using a 3.2ms liftering window



FFT cepstrum and its Fourier and smoothed spectra for a speech frame ("a") (256) uttered by a woman, using a 1.2ms liftering window

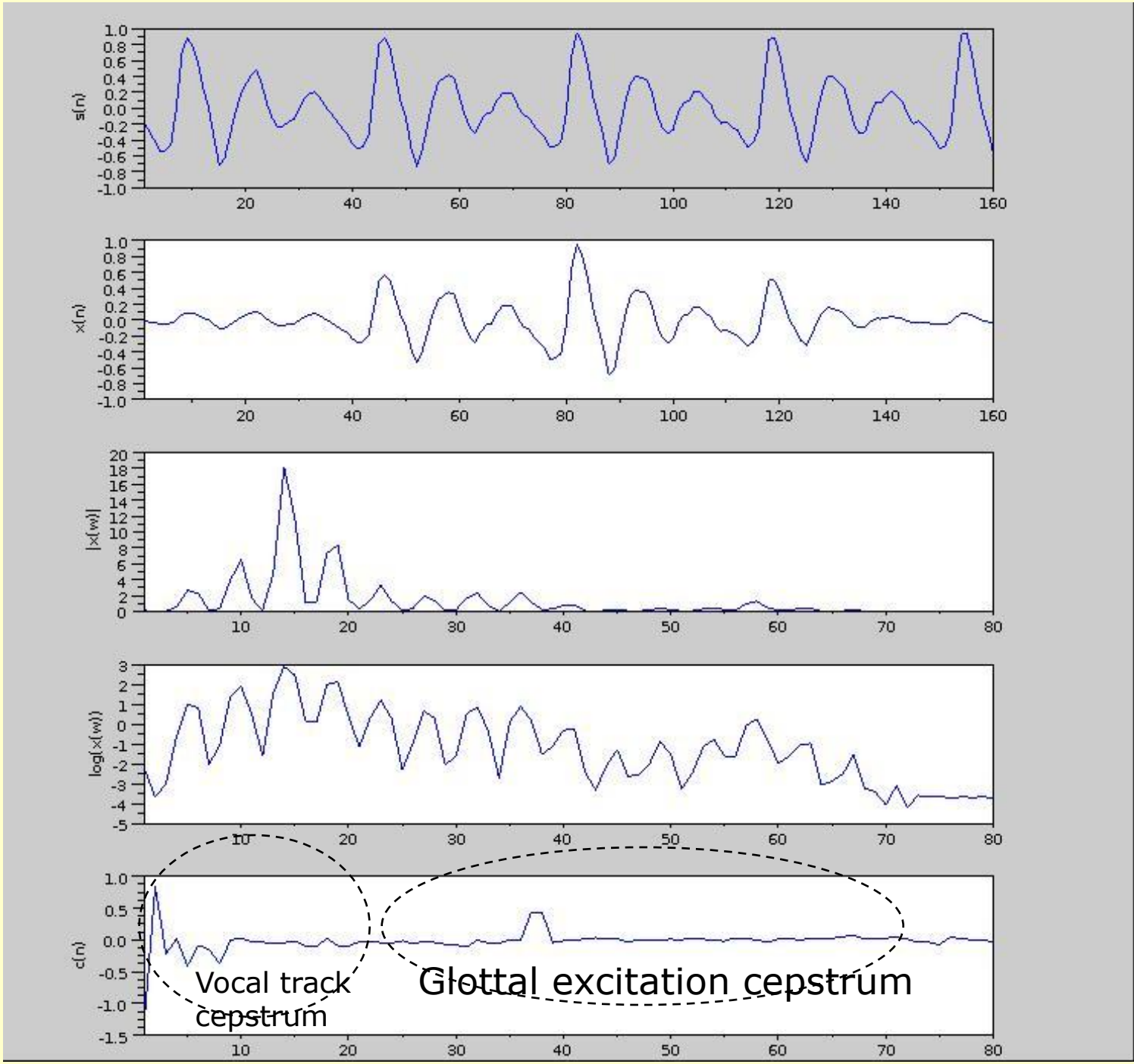
x(n) time domain signal

x(w)=dft(x(n)),frequency signal

|x(w)|

Log (|x(w)|)

C(n)=iDft(Log (|x(w)|)) =>Cepstrum



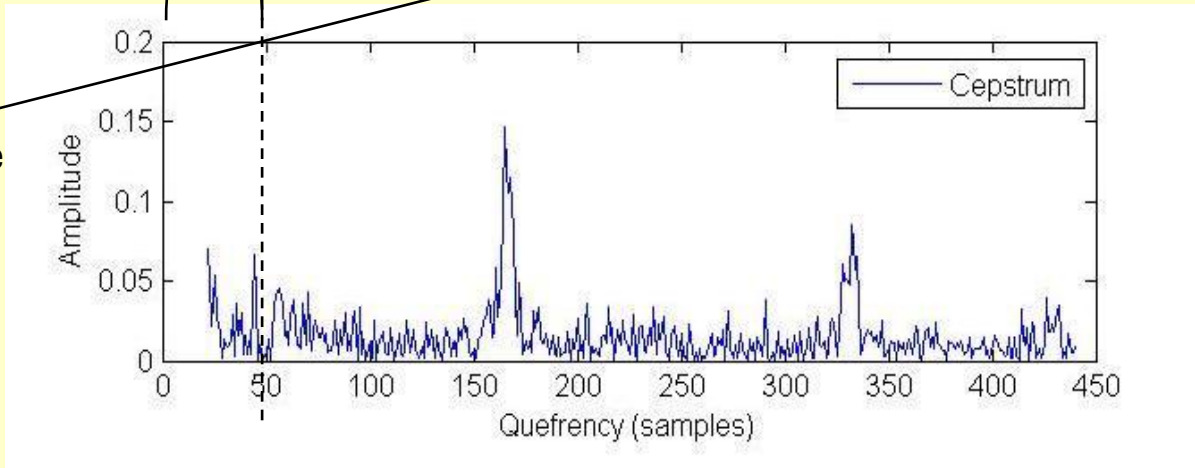
Liftering

- Low time liftering:
  - Magnify (or Inspect) the low time to find the vocal tract filter cepstrum

Vocal tract
Cepstrum
Used for
Speech
recognition

Glottal excitation
Cepstrum

- High time liftering:
  - Magnify (or Inspect) the high time to find the glottal excitation cepstrum (remove this part for speech recognition.



Cut-off Found
by experiment

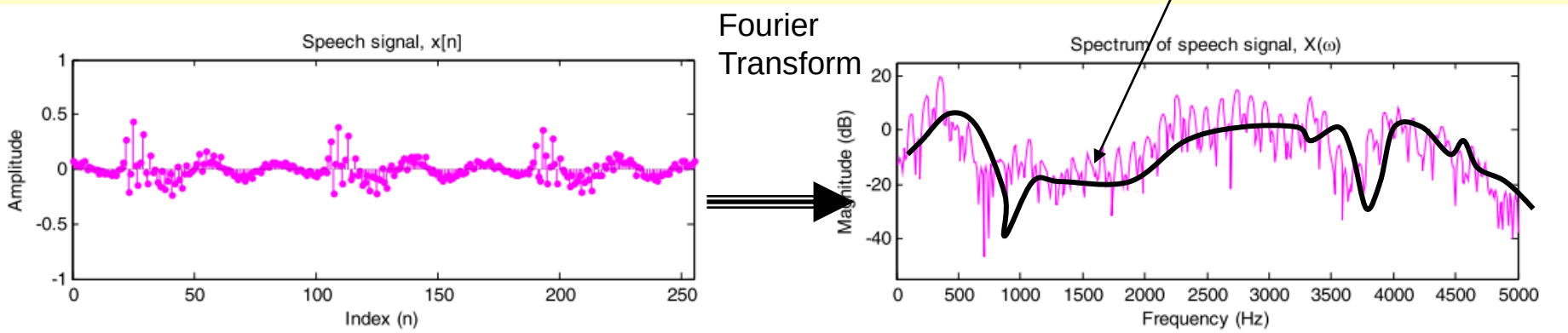
Frequency =FS/ quefrency coef.

Reasons for liftering Cepstrum of speech

- Why we need this?

-remove the ripples of the spectrum caused by glottal excitation.

Too many ripples in the spectrum are caused by vocal cord vibrations.
But we are more interested in the speech envelope for recognition and synthesis



Speech signal x(n)

Spectrum of x(n)

Homework

SS is sampled at 20kHz. For short-term spectral analysis, a sliding window of 20 ms is used, which moves by 10 ms for the analysis of consecutive frames. The radix-2 FFT method is used to calculate the DFT.

1. How many samples are used for each analysis frame?

2. What is the frame analysis rate for short-term spectral analysis?

3. What is the size required for DFT and FFT to guarantee the absence of temporal aliasing?

4. What is the frequency resolution (Hz) between 2 consecutive samples?