

XỬ LÝ TIẾNG NÓI

Trịnh Văn Loan

FIT-HUT

Tài liệu tham khảo

1. La parole et son traitement automatique

Calliope, Masson, 1989

2. Traitement de la parole

Rene Boite et Murat Kunt, Presse Polytechniques
Romandes, 1987

3. Fundamentals of Speech Signal Processing

Saito S., Nakata K. , Academic Press, 1985



Nội dung

- 1. Một số khái niệm cơ bản**
- 2. Xử lý tín hiệu tiếng nói**
- 3. Mã hoá tiếng nói**
- 4. Tổng hợp tiếng nói**
- 5. Nhận dạng tiếng nói**



Xử lý tiếng nói ?

→ Xử lý thông tin chứa trong tín hiệu tiếng nói nhằm **truyền, lưu trữ** tín hiệu này hoặc **tổng hợp, nhận dạng** tiếng nói.

→ Các nghiên cứu được tiến hành để xử lý tiếng nói yêu cầu những hiểu biết trên nhiều lĩnh vực ngày càng đa dạng: từ **ngữ âm** và **ngôn ngữ học** cho đến **xử lý tín hiệu...**



Mục đích

- Mã hoá một cách có hiệu quả tín hiệu tiếng nói để truyền và lưu trữ tiếng nói.
- Tổng hợp và nhận dạng tiếng nói tiến tới giao tiếp người-máy bằng tiếng nói.

→ Tất cả các ứng dụng của xử lý tiếng nói đều cần phải dựa trên các kết quả của phân tích tiếng nói



1. Một số khái niệm cơ bản

Phân biệt tiếng nói và âm thanh:

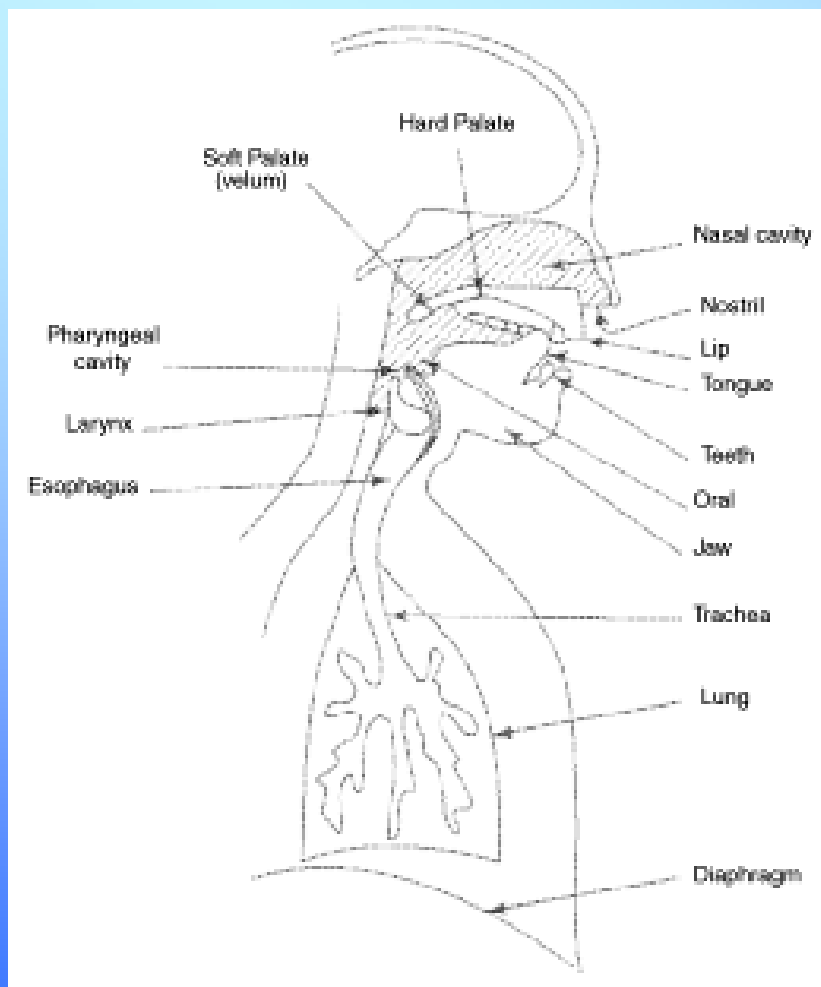
● Tiếng nói được phân biệt với các âm thanh khác bởi các đặc tính âm học có nguồn gốc từ cơ chế tạo tiếng nói.

Có 2 loại nguồn âm

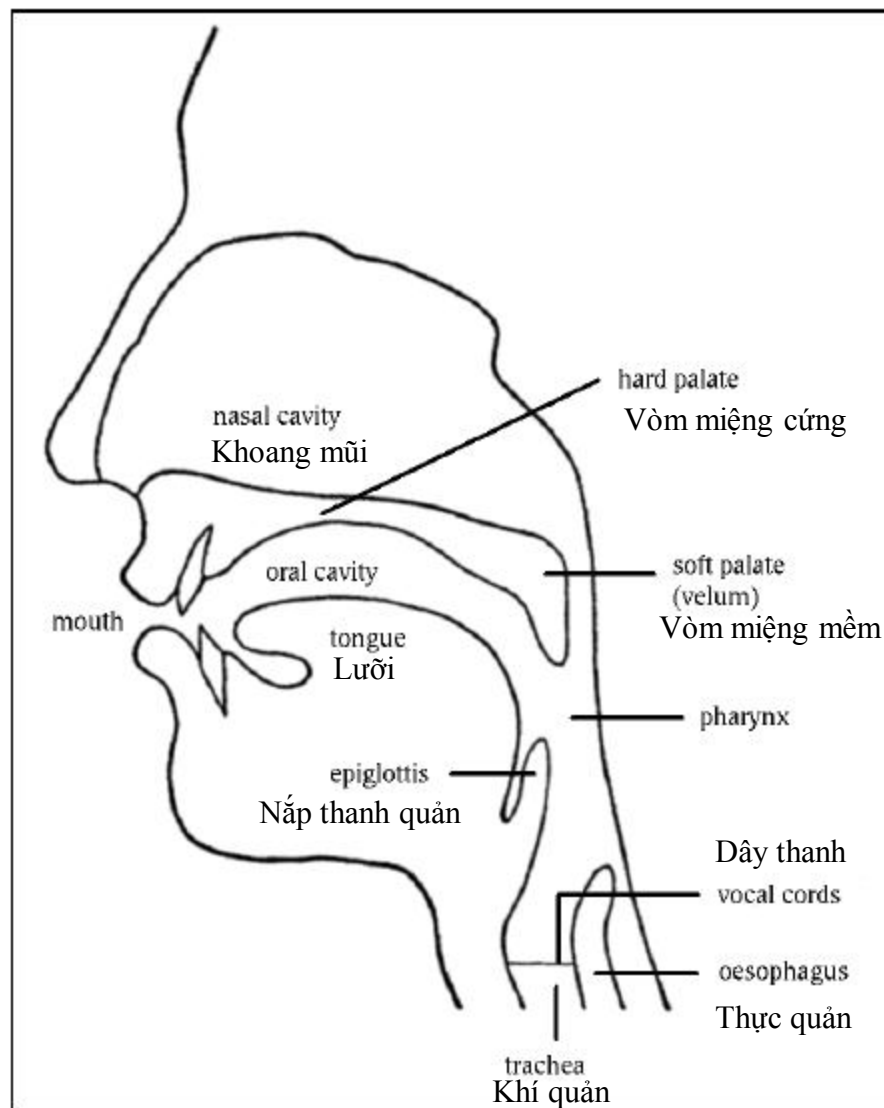
- tuần hoàn (dây thanh rung)
- tạp âm (dây thanh không rung)



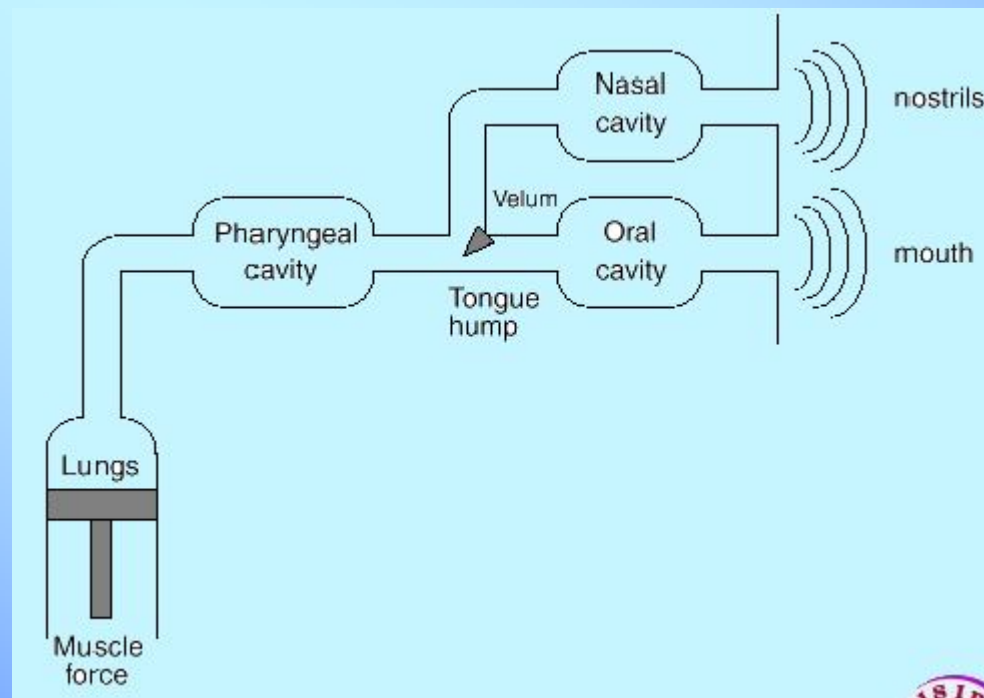
Bộ máy phát âm



1. Một số khái niệm cơ bản

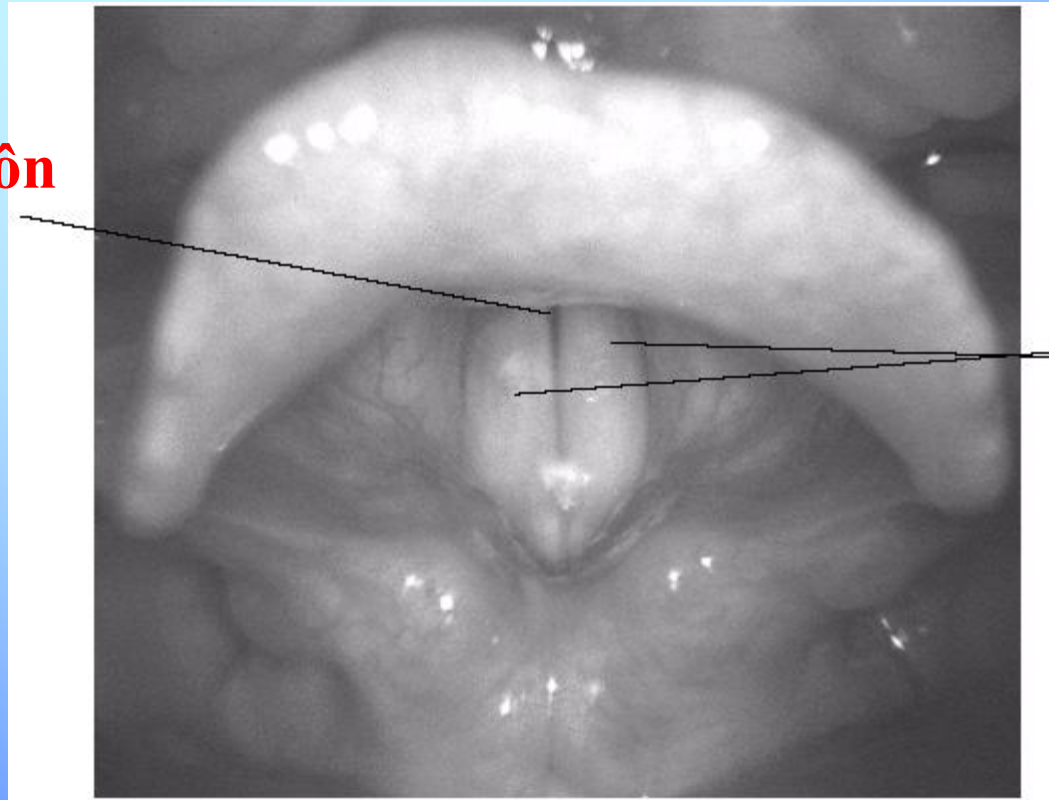


Sơ đồ khối bộ máy phát âm



Thanh môn (1)

Thanh môn

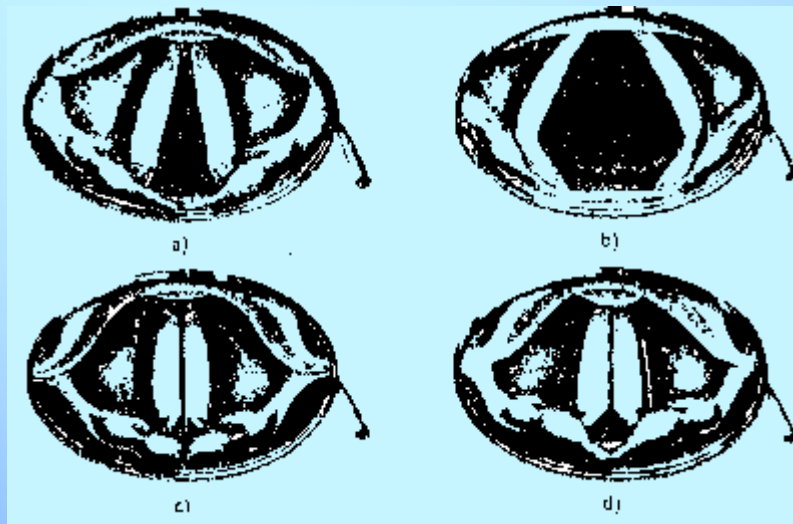


Dây thanh

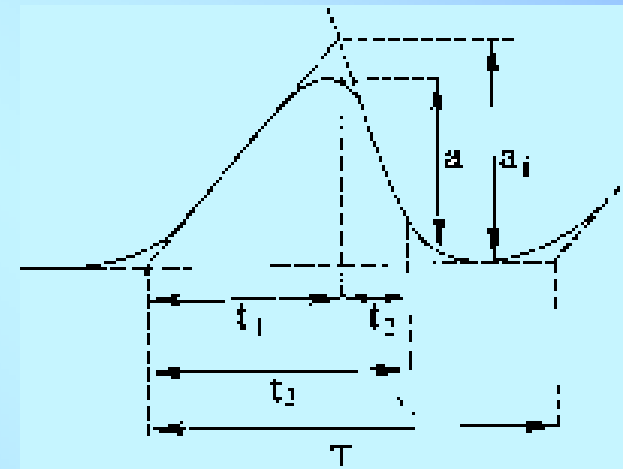
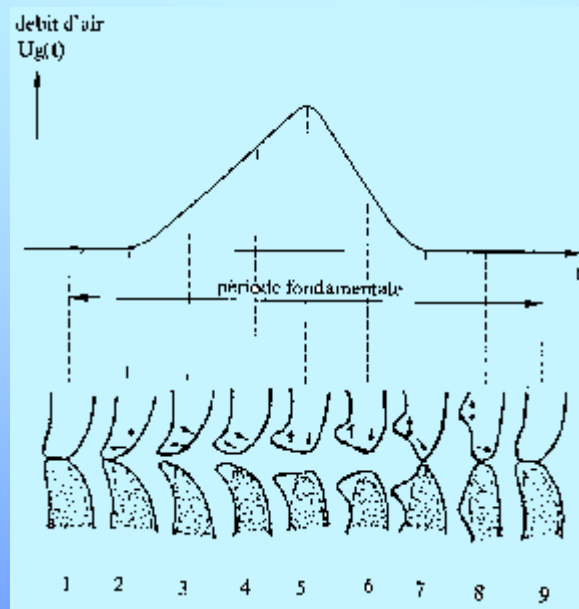


Thanh môn (2)

● Ở các vị trí hít, thở, phát âm, nói thì trào

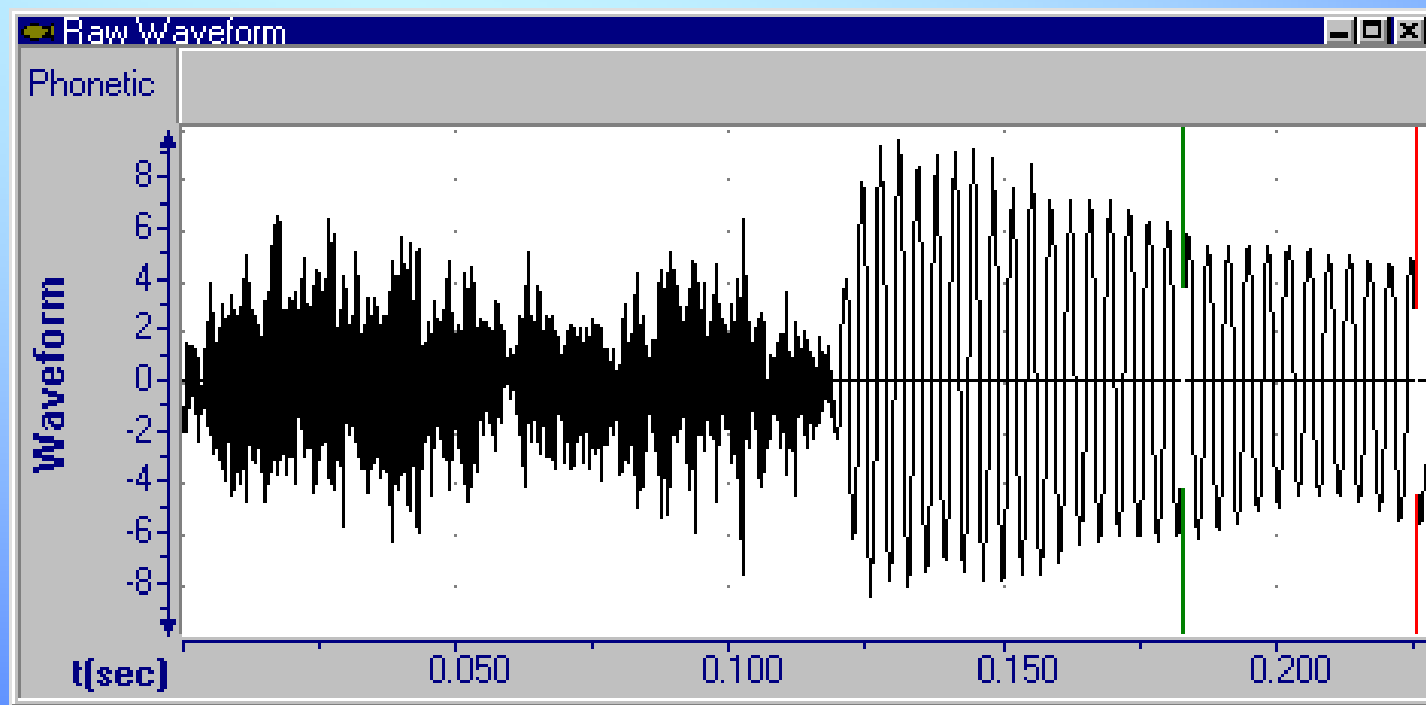


Dây thanh trong một chu kỳ dao động



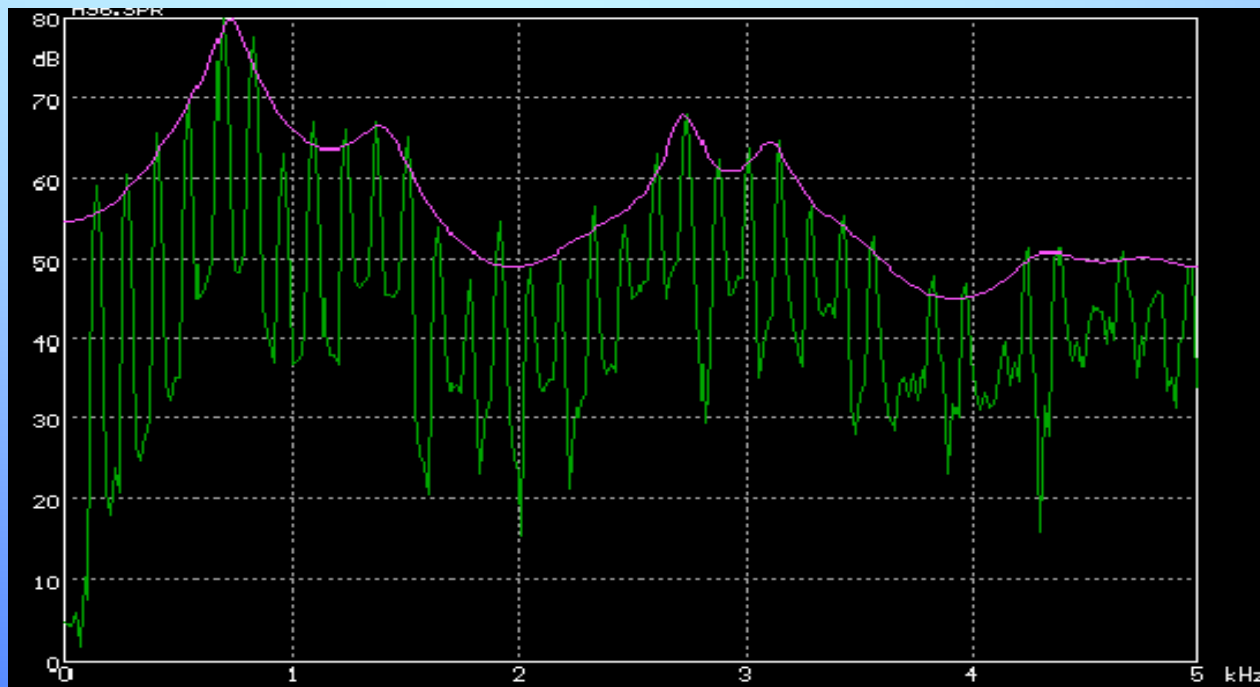
Biểu diễn tín hiệu tiếng nói

→ Dạng sóng theo thời gian



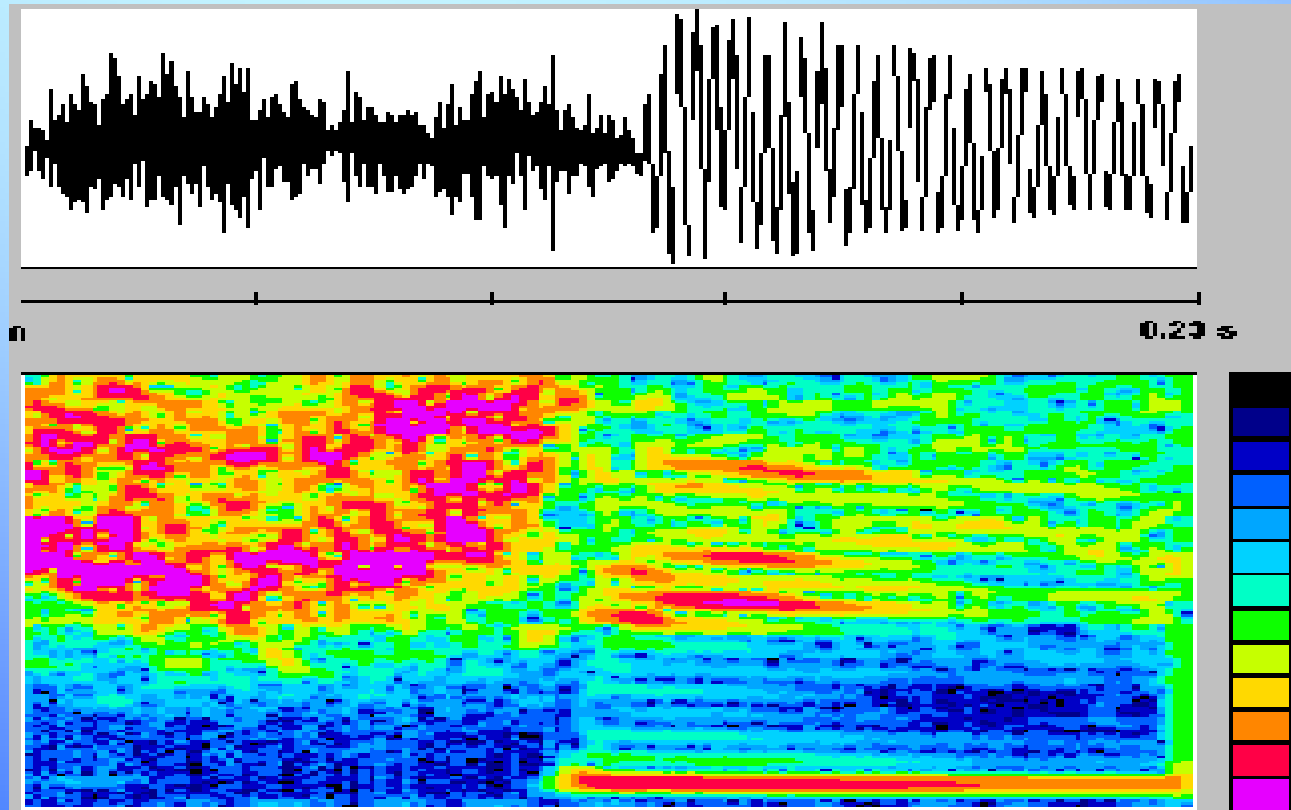
Biểu diễn tín hiệu tiếng nói

→ Phổ tín hiệu tiếng nói

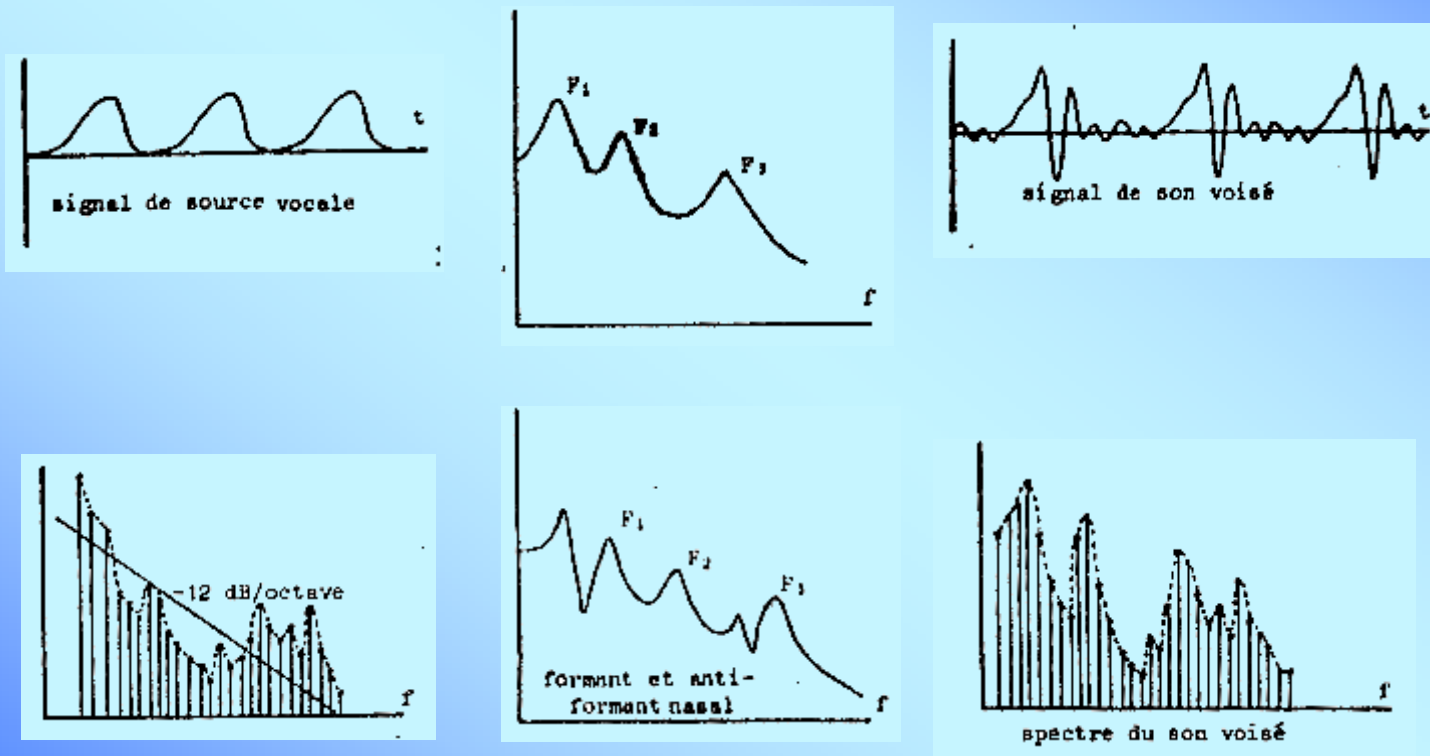


Biểu diễn tín hiệu tiếng nói

→ Spectrogram (Sonagram)

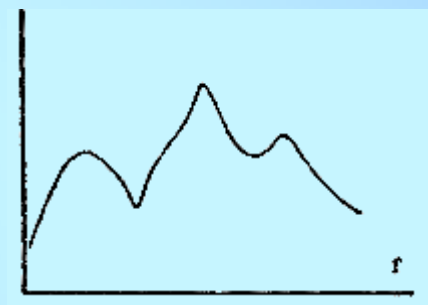
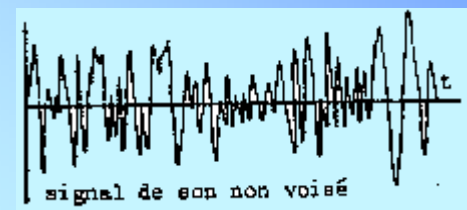
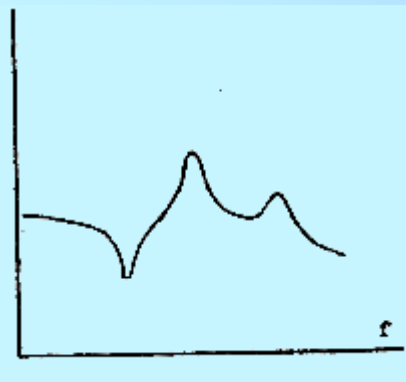
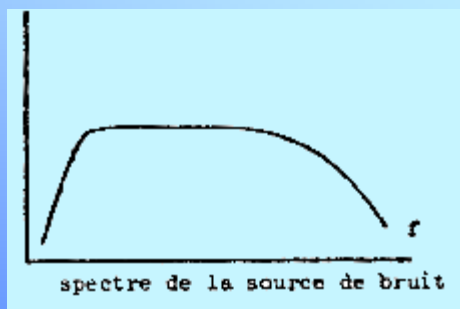
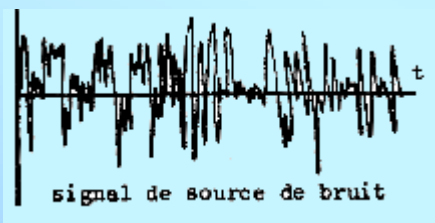


Tạo âm hữu thanh . Formant và antiformant



Tạo âm vô thanh

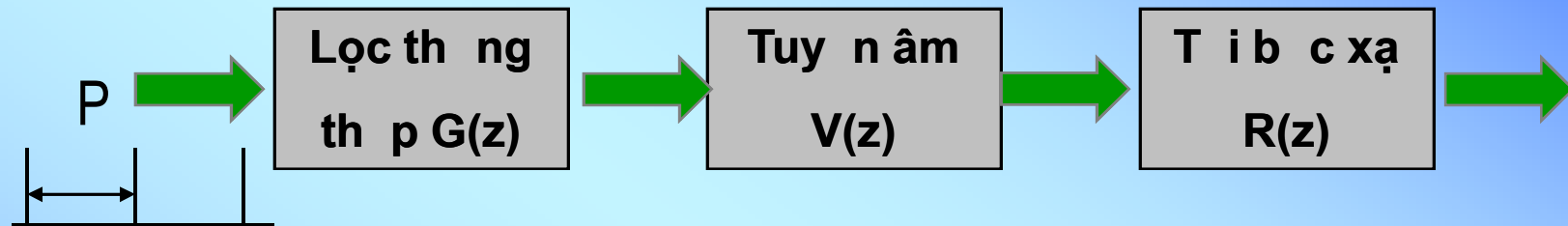
1. Một số khái niệm cơ bản



Một số đặc điểm ngữ âm tiếng Việt

- Đơn âm tiết
- Có thanh điệu (6), biến đổi thanh điệu kèm theo biến đổi nghĩa
- Không biến đổi hình thái

Mô hình tạo tiếng nói (Fant-1960)



$$G(z) = \frac{A}{(1 + \alpha z^{-1})(1 + \beta z^{-1})}$$

$$R(z) = C(1 - z^{-1})$$

$$V(z) = \frac{B}{\prod_{k=1}^K (1 + b_{1k} z^{-1} + b_{2k} z^{-2})}$$

$A(z)$: Hàm truyền đạt của bộ lọc đảo

$$T(z) = G(z)V(z)R(z) = \frac{\sigma}{A(z)}$$

Mô hình toàn điểm cực (AR)

$$T(z) = \frac{\sigma}{A(z)}$$

$$A(z) = 1 + \sum_{i=1}^{2K+1} a_i z^{-i}$$

$$x(n) + \sum_{i=1}^p a_i x(n-i) = \sigma u(n)$$



**Nếu tính đến khoang mũi \Rightarrow
xuất hiện các điểm không (ARMA)**

$$T(z) = \frac{\sigma_1}{A_1(z)} + \frac{\sigma_2}{A_2(z)} = \sigma \frac{C(z)}{A(z)}$$

$$C(z) = \sum_{i=0}^q c_i z^{-i} \quad c_0 = 1$$

$$x(n) + \sum_{i=1}^p a_i x(n-i) = \sigma \sum_{i=0}^q c_i u(n-i)$$

Bài tập

Bài 1.

Hàm truyền đạt của một bộ lọc số ở tần số formant F_k được cho bởi:

$$H_k(z) = \frac{1 - 2|z_k| \cos \theta_k + |z_k|^2}{1 - 2|z_k| \cos \theta_k z^{-1} + |z_k|^2 z^{-2}}$$

trong đó $\theta_k = 2\pi F_k T$, $|z_k| = e^{-\sigma_k T}$, T : chu kỳ lấy mẫu, $2\sigma_k$: dải thông.

1. Vẽ các điểm cực của $H_k(z)$ trong mặt phẳng Z
2. Viết phương trình sai phân mô tả quan hệ giữa tín hiệu ra $y_k(n)$ và tín hiệu vào $x_k(n)$
3. Vẽ sơ đồ khối của bộ lọc số này với 3 bộ nhân.
4. Bằng cách sắp xếp lại các số hạng của phương trình sai phân, vẽ sơ đồ khối của bộ lọc số chỉ có 2 bộ nhân

Bài tập

Bài 1.

Hàm truyền đạt của một bộ lọc số ở tần số formant F_k được cho bởi:

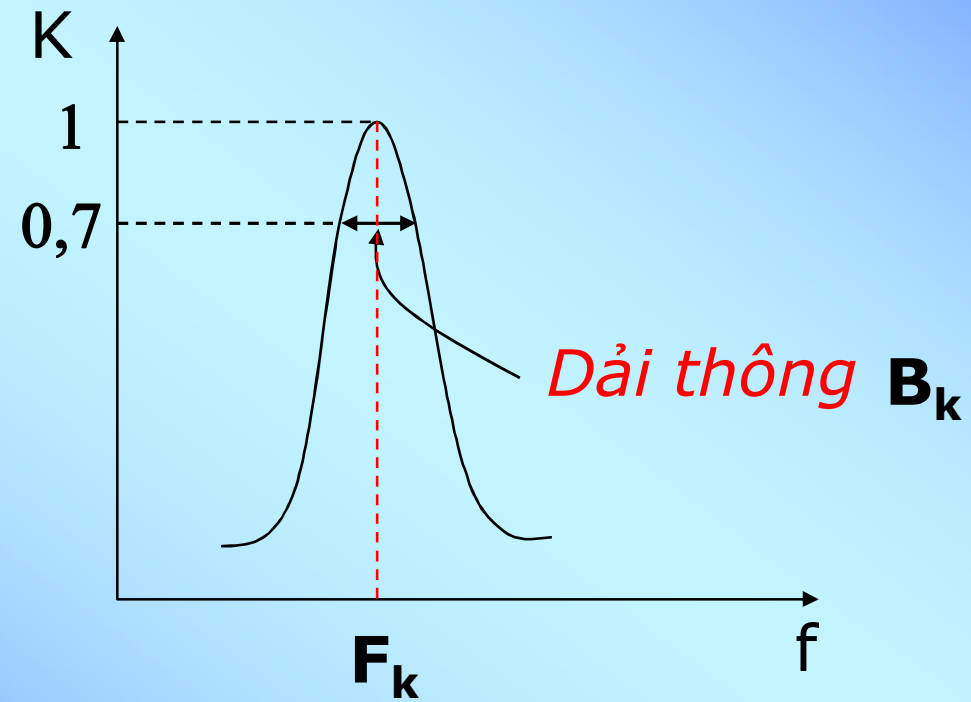
$$H_k(z) = \frac{1 - 2|z_k| \cos \theta_k + |z_k|^2}{1 - 2|z_k| \cos \theta_k z^{-1} + |z_k|^2 z^{-2}}$$

trong đó $\omega_k = 2\pi F_k T$, $|z_k| = e^{-\sigma_k T}$, T : chu kỳ lấy mẫu, $2\sigma_k$: dải thông.

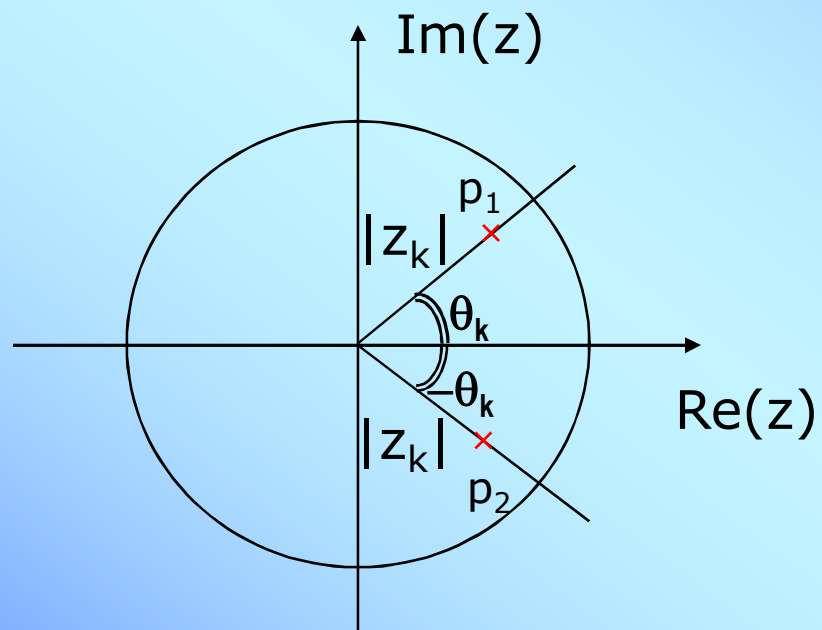
1. $H(z)$ có 2 điểm cực là nghiệm của mẫu số:

$$\begin{aligned} \Delta' &= |z_k|^2 \cos^2 \theta_k - |z_k|^2 = -|z_k|^2 \sin^2 \theta_k \\ p_{1,2} &= |z_k| \cos \theta_k \pm j |z_k| \sin \theta_k \\ &= |z_k| e^{\pm j \theta_k} \end{aligned}$$

Bài tập



Bài tập

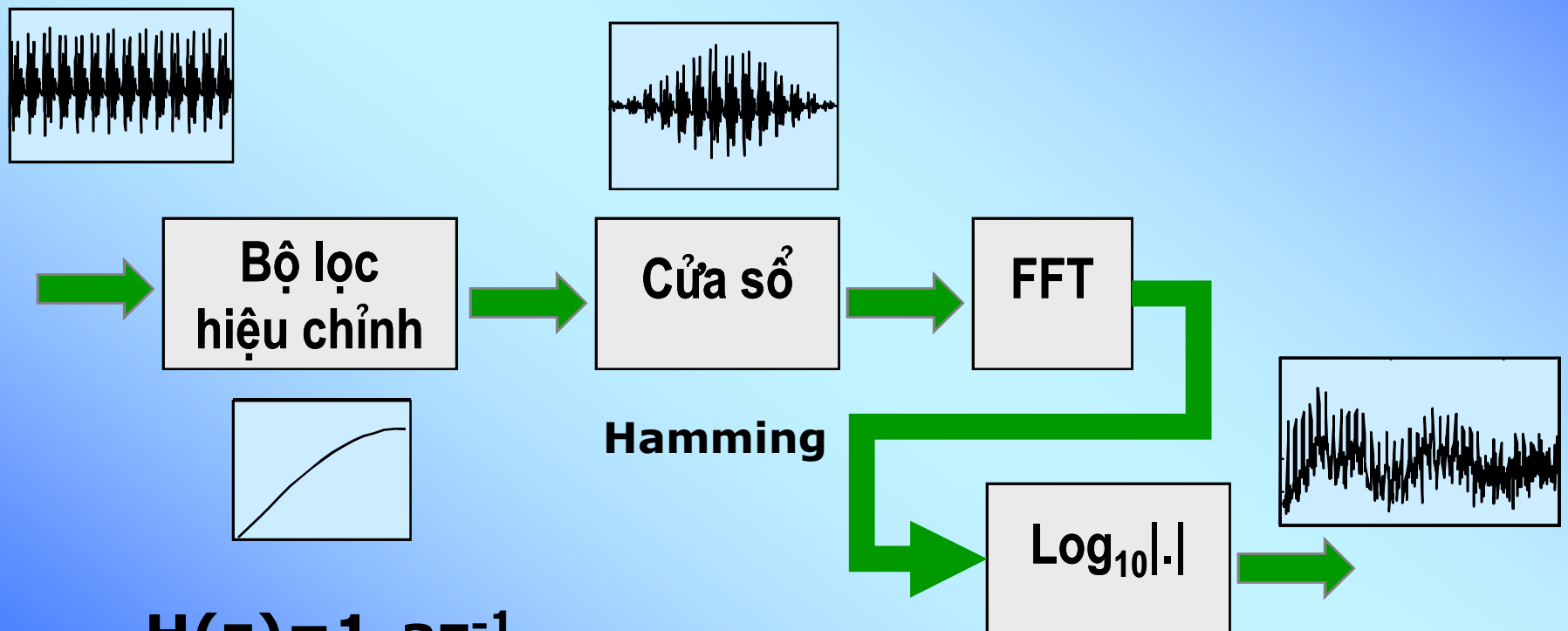


Bài tập

a) Xác định biến đổi z của $g(n)$: Tra bảng

2. Xử lý tín hiệu tiếng nói

Phân tích phổ tín hiệu tiếng nói



$$H(z) = 1 - az^{-1}$$

$$a = 0,95 \dots 0,98$$

$$\mathbf{x'(n) = x(n).w(n)}$$

$$\mathbf{X'(f) = X(f) * W(f)}$$

Xử lý đồng hình (homomorphisme)

$$s(n) = h(n) * e(n) \rightarrow S(\omega) = H(\omega).E(\omega)$$

$$\log[S(\omega)] = \log[H(\omega)] + \log[E(\omega)]$$

$$\text{FFT}^{-1} \{ \log[S(\omega)] \} =$$

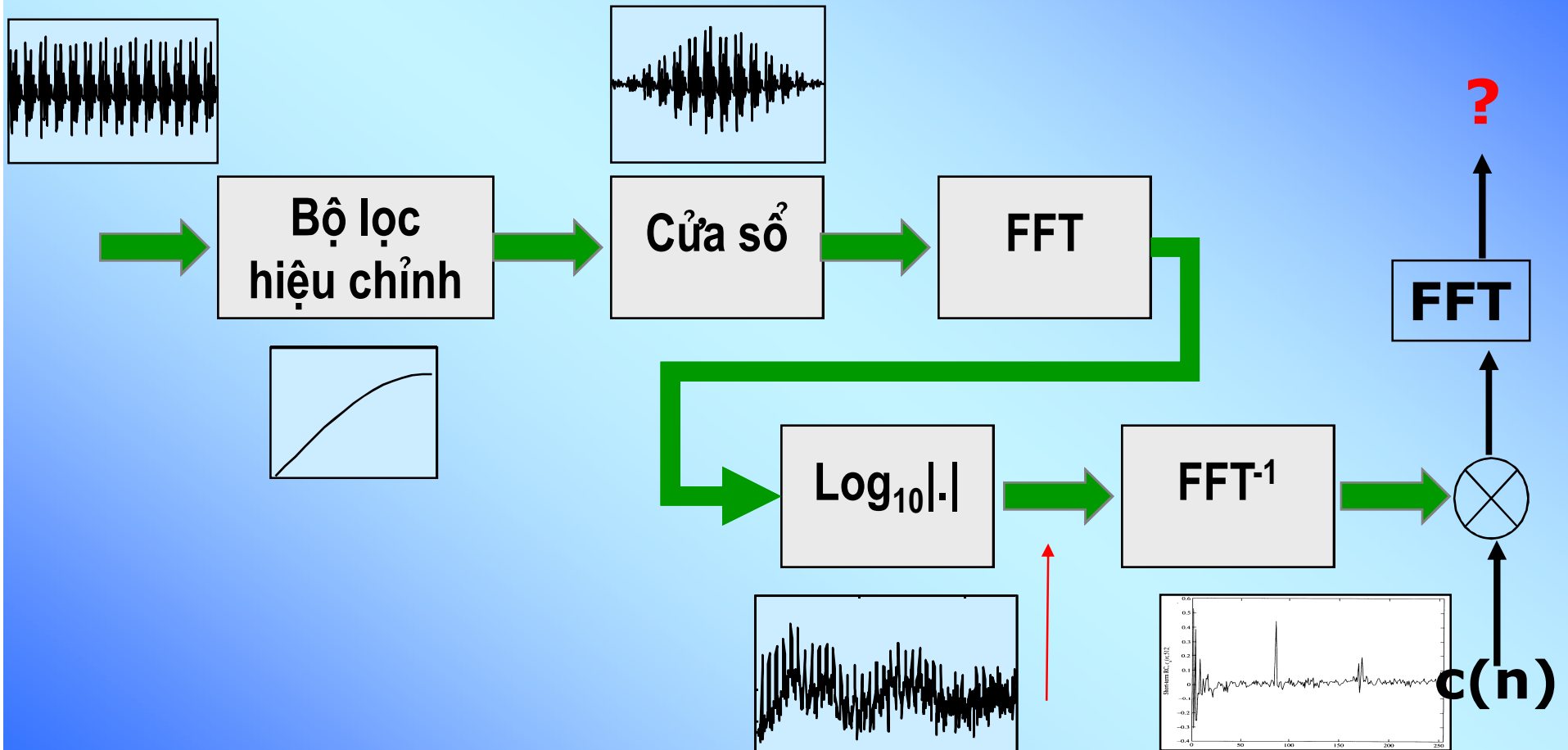
$$\text{FFT}^{-1} \{ \log[H(\omega)] \} + \text{FFT}^{-1} \{ \log[E(\omega)] \}$$

$$\text{FFT}^{-1} \{ \log[S(\omega)] \} : \text{cepstrum} : \acute{s}(n)$$

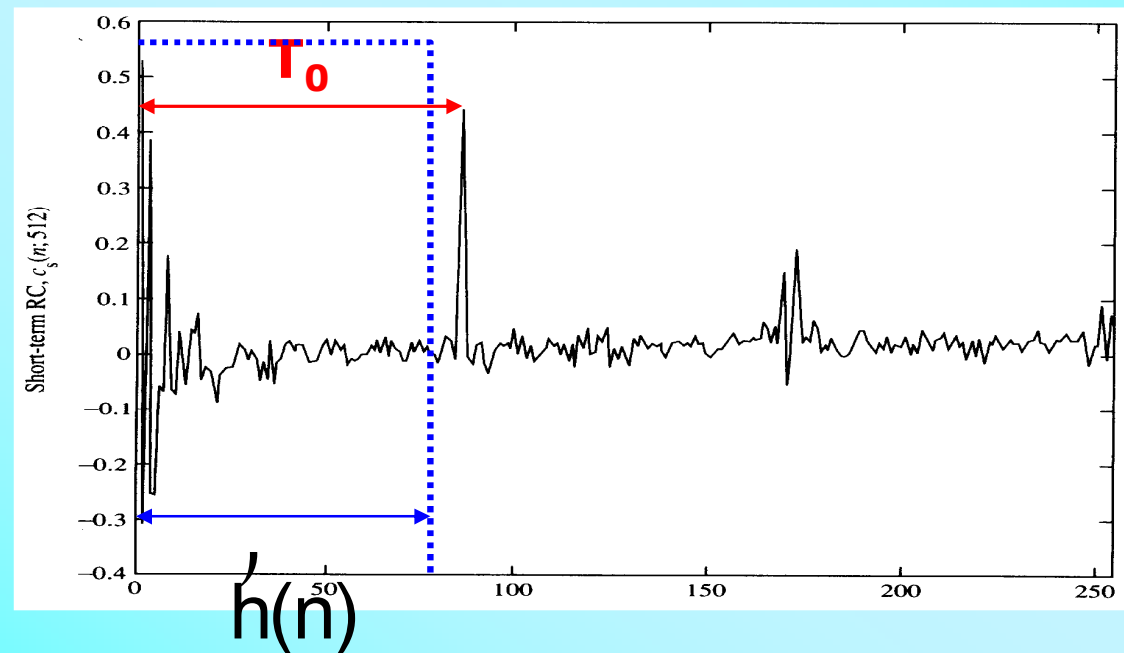
$$\text{FFT}^{-1} \{ \log[H(\omega)] \} : \text{thông tin về } h(n) : \acute{h}(n)$$

$$\text{FFT}^{-1} \{ \log[E(\omega)] \} : \text{thông tin về } \text{nguồn} : \acute{e}(n)$$

Sơ đồ khối xử lý đồng hình



$$\hat{s}(n) = \hat{h}(n) + \hat{e}(n)$$



Tiên đoán tuyến tính (Linear Prediction Coding)

Mô hình toàn điểm cực

$$x(n) + \sum_{i=1}^p a_i x(n-i) = \sigma u(n)$$

Tiên đoán

$$\hat{x}(n) = -\sum_{i=1}^p \hat{a}_i x(n-i)$$

Sai số tiên đoán

$$e(n) = x(n) - \hat{x}(n)$$

Sai số bình phương toàn phần

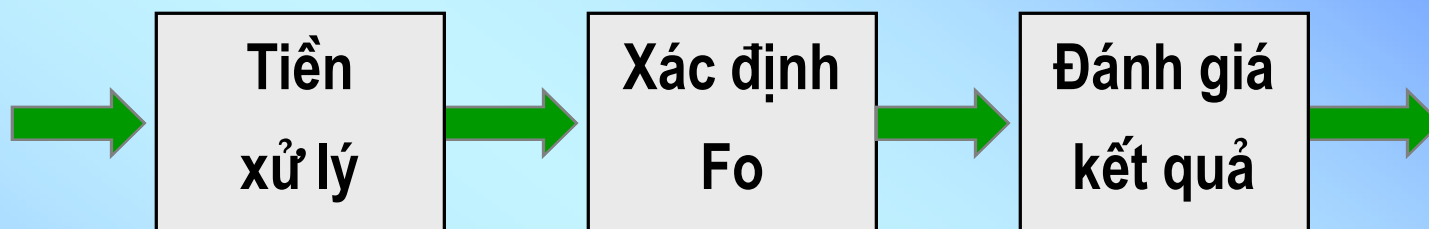
$$E = \sum_n e^2(n)$$

Tối thiểu hoá sai số

$$\frac{\partial E}{\partial \hat{a}_i} = 0, \quad i = 1, 2, \dots, p$$

Xác định tần số cơ bản F_0

Giọng nam: 80 .. 250 Hz. Giọng nữ: 150..500 Hz



Một số phương pháp xác định F_0

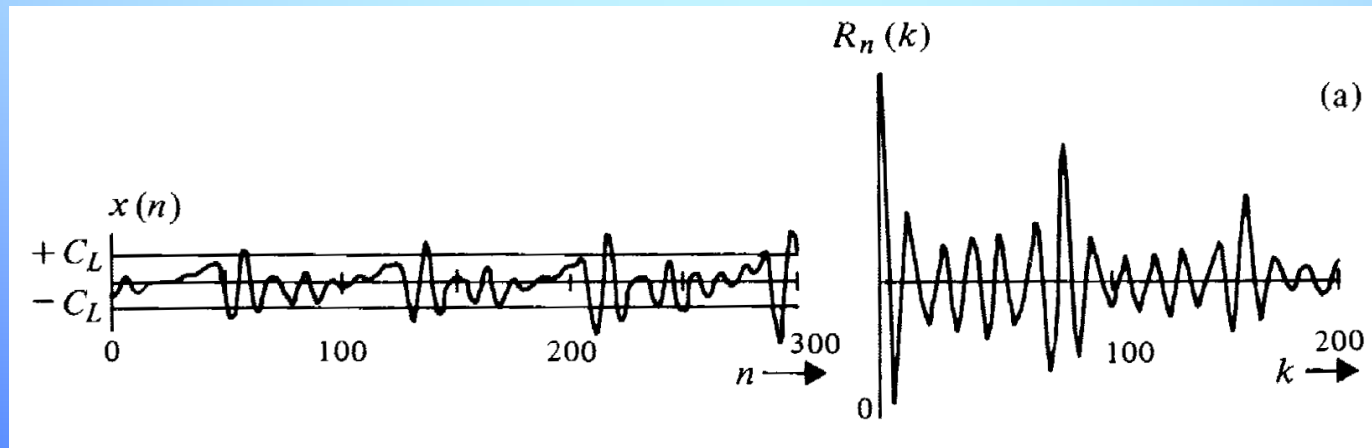
- ◆ Dựa vào hàm tự tương quan
- ◆ Dựa vào hàm vi sai biên độ trung bình
- ◆ Dùng bộ lọc đảo và hàm tự tương quan
- ◆ Xử lý đồng hình

◆ Dựa vào hàm tự tương quan

$$r(k) = \sum_{n=1}^{N-1-k} x(n)x(n+k) \quad k = 0, 1, \dots, K$$

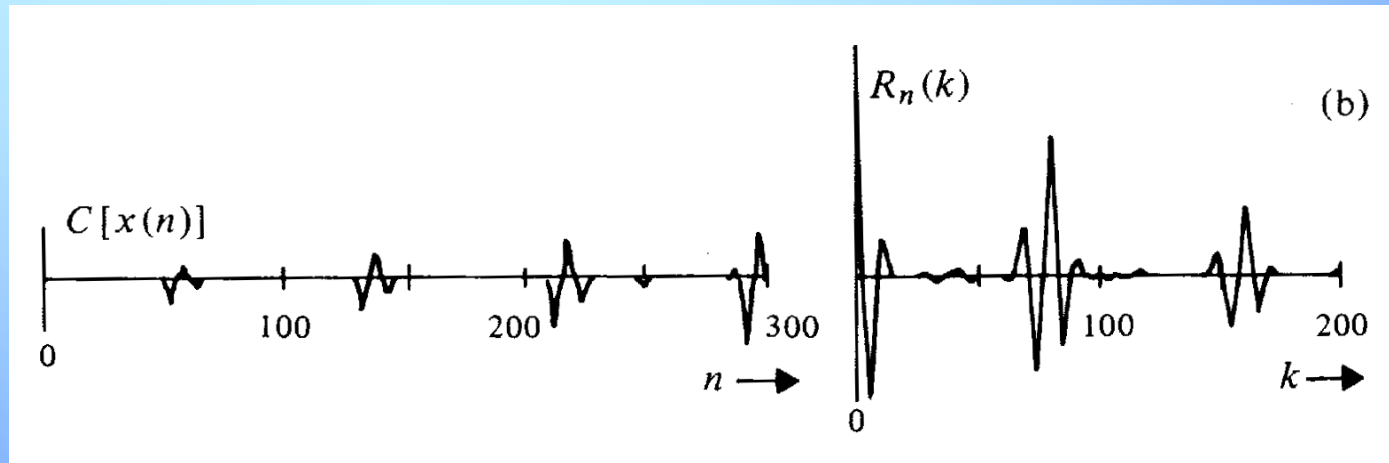
$$F_s = 10 \text{ kHz}, N = 300, K = 150$$

Tìm cực đại trong khoảng $(0, K)$



◆ Dựa vào hàm tự tương quan

Hạn chế, loại bỏ $|x| < C_L$



◆ Dựa vào hàm vi sai biên độ trung bình (Average Magnitude Difference Function) (1)

$$D(k) = \sum_{n=1}^{N-1-k} |x(n) - x(n-k)| \quad k = 0, 1, \dots, K$$

$$D(iP) = 0, \quad i = 0, 1, \dots \quad \frac{1}{N} \sum_{n=0}^{N-1} |u(n)| \leq \left[\frac{1}{N} \sum_{n=0}^{N-1} u^2(n) \right]^{1/2}$$

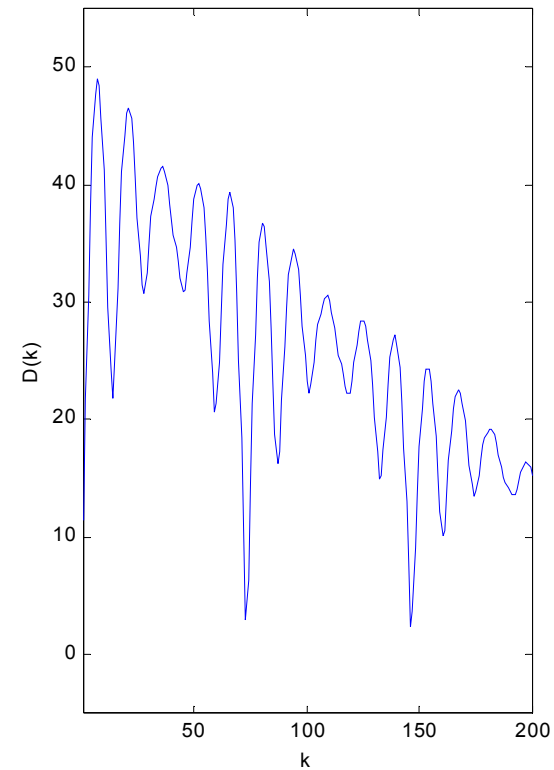
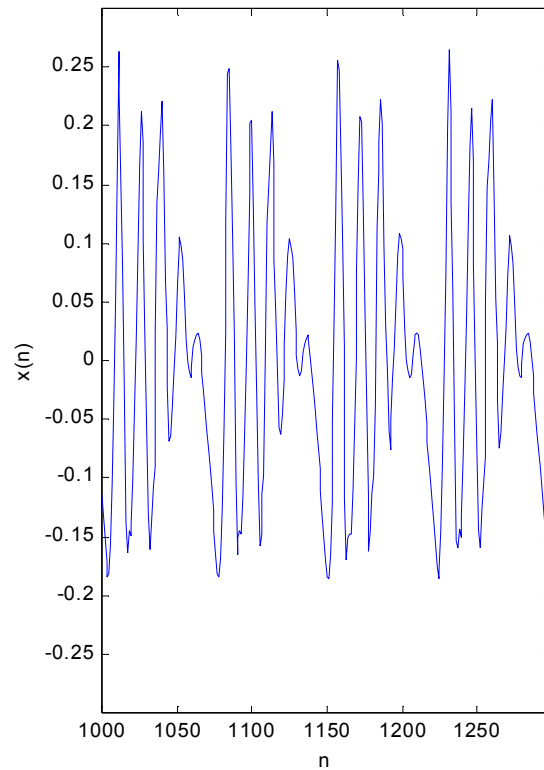
$$D(k) = \lambda \left\{ \frac{1}{N} \sum_{n=0}^{N-1} [x(n) - x(n+k)]^2 \right\}^{1/2}$$

$$= \lambda 2[r(0) - r(k)]^{1/2} \quad k = 0, 1, \dots, K$$

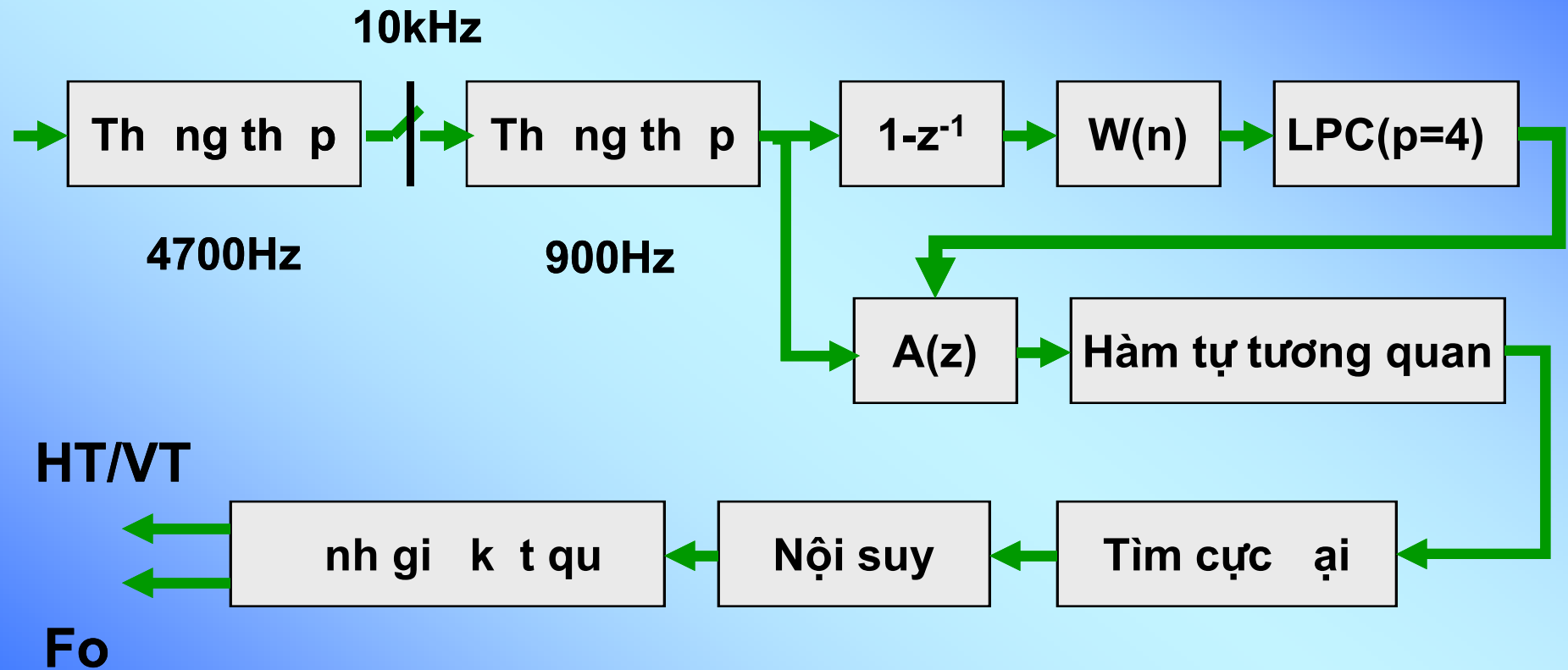
với $\lambda < 1$

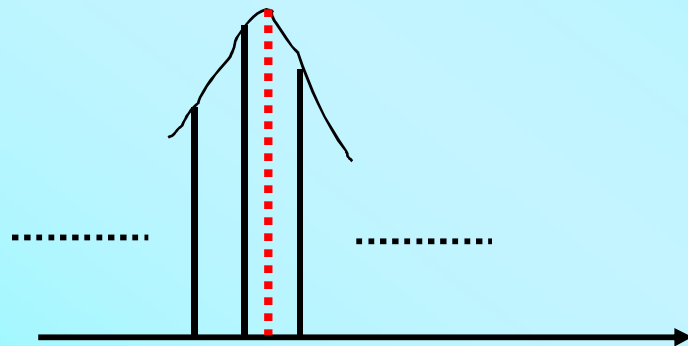


◆ Dựa vào hàm vi sai biên độ trung bình (Average Magnitude Difference Function) (2)

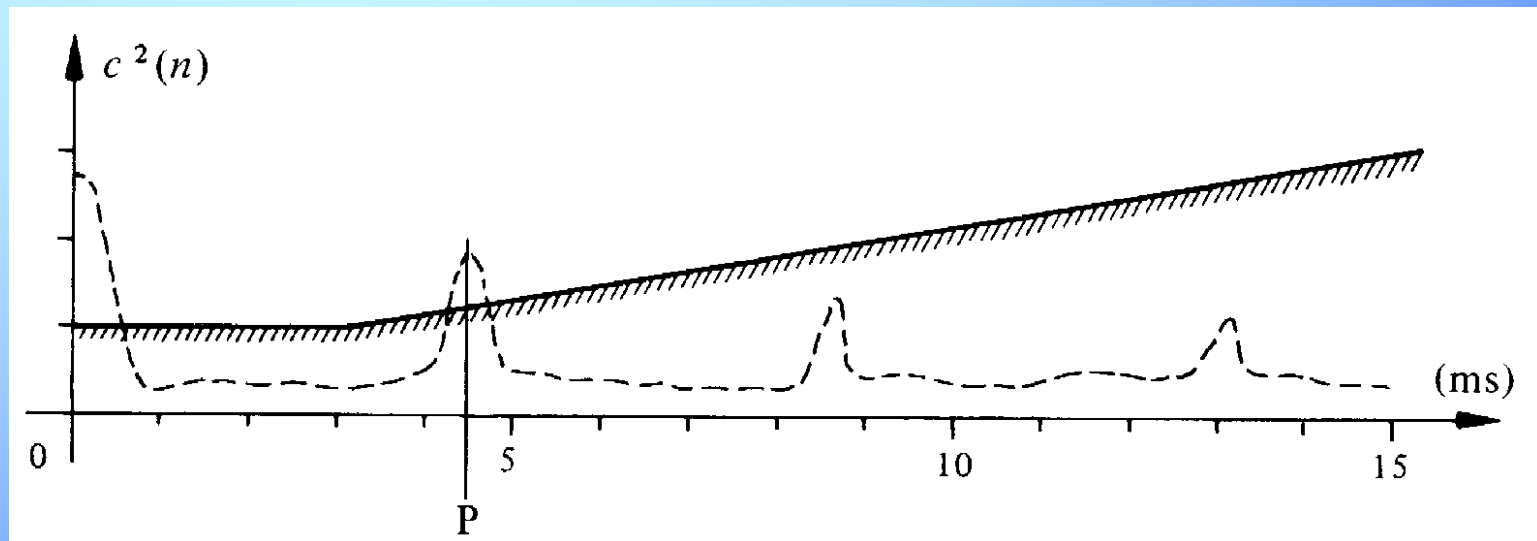


◆ Dùng bộ lọc đảo (Simplified Inverse Filter Tracking)



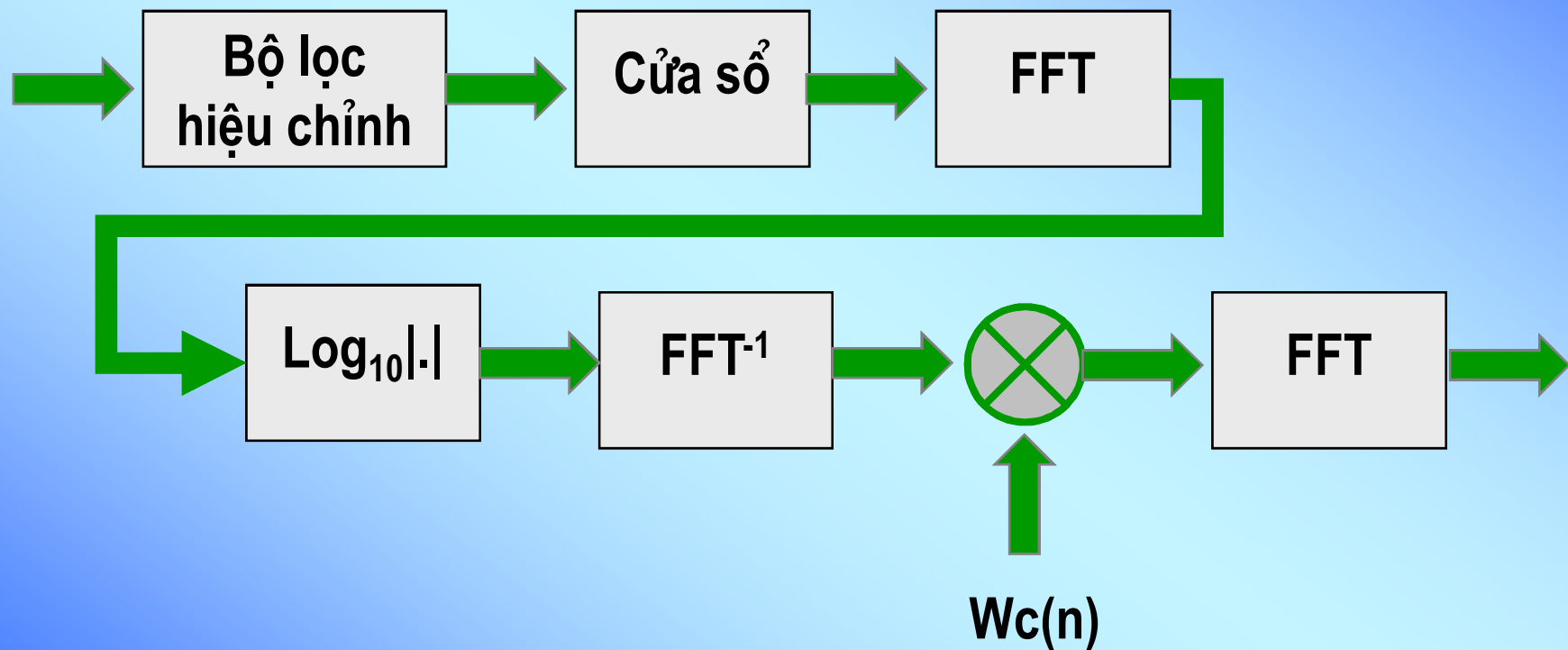


◆ Xử lý đồng hình



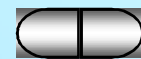
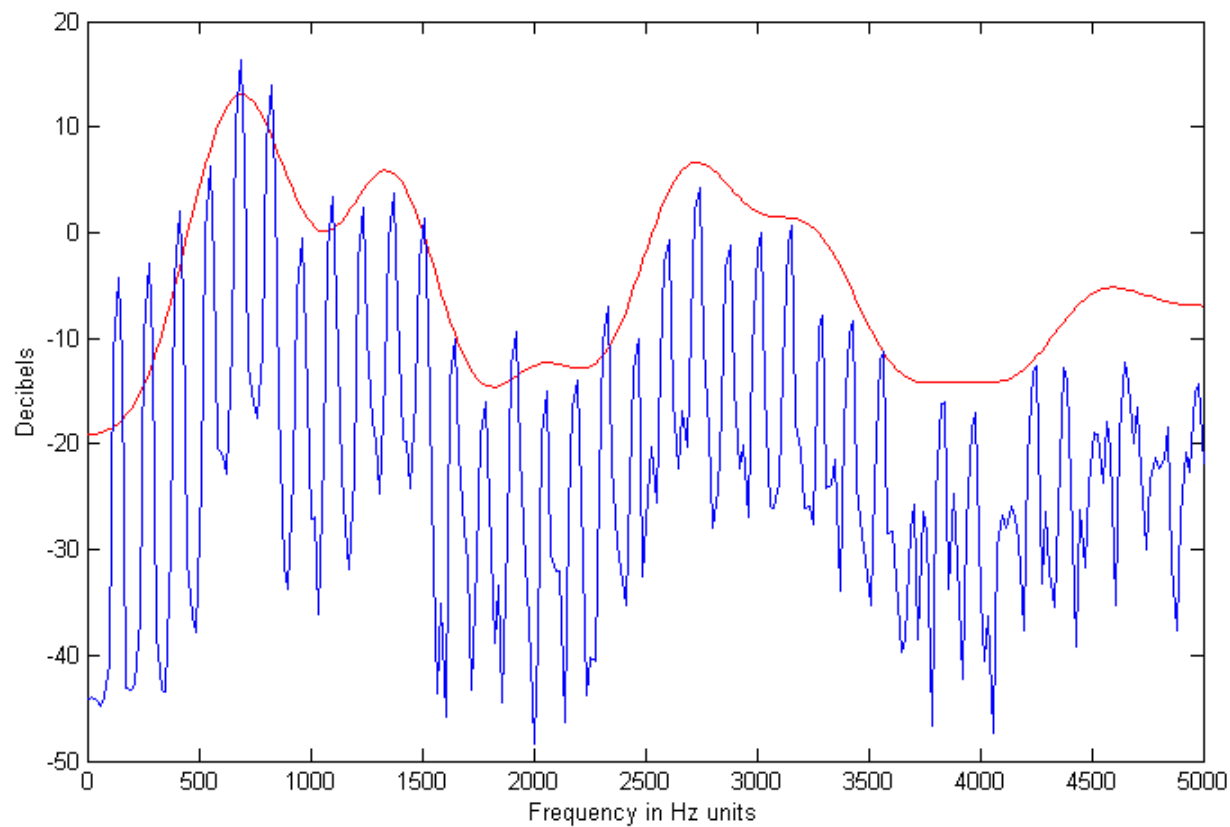
Xác định formant (1)

→ Xử lý đồng hình



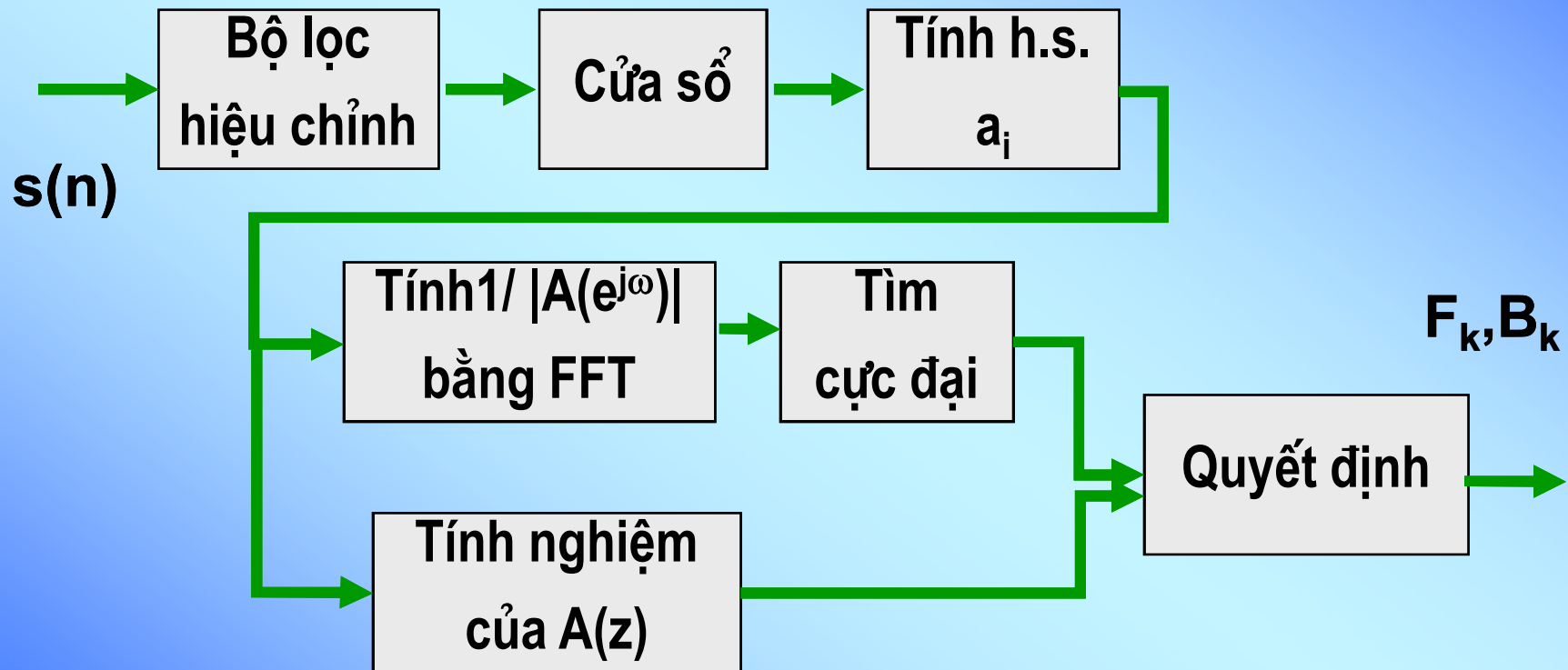
Xác định formant (1)

→ Xử lý đồng hình



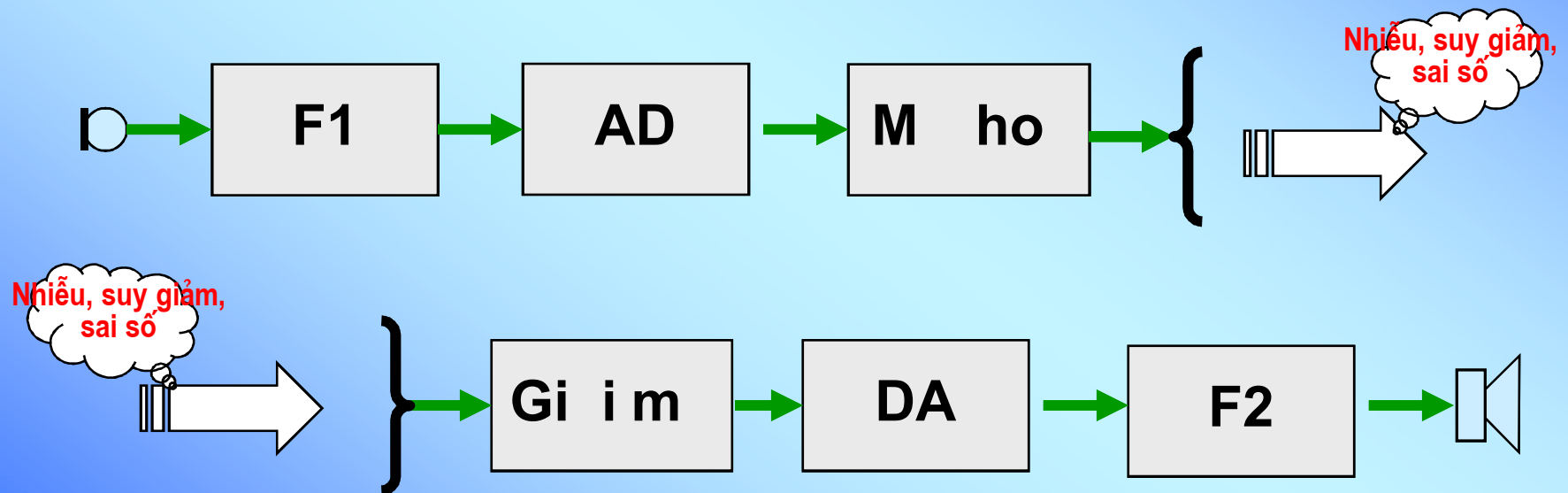
Xác định formant (2)

→ Tiên đoán tuyến tính (LPC)



3. Mã hoá tiếng nói

Dãy thao tác mã hoá và giải mã



Một số tính chất thống kê của tín hiệu tiếng nói

◆ Mật độ xác suất

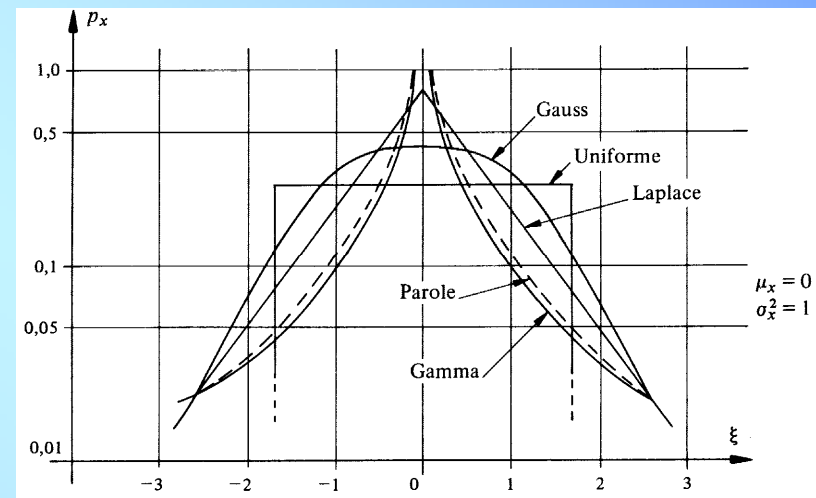
N_ξ : số lượng mẫu $x(n)$
có biên độ trong khoảng

$$[\xi - \Delta\xi/2, \xi + \Delta\xi/2]$$

$$n \in [-N, \dots, N]$$

x ergodic và dừng

$$p_x(\xi) = \lim_{\substack{N \rightarrow \infty \\ \Delta\xi \rightarrow 0}} [N_\xi / (2N + 1)]$$



◆ Giá trị trung bình và phương sai

- Giá trị trung bình của tín hiệu dừng

$$\mu_x = \int_{-\infty}^{\infty} \xi p_x(\xi) d\xi = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N x(n)$$

với tín hiệu tiếng nói, giả thiết $\mu_x = 0$

- Phương sai

$$\sigma_x^2 = \int_{-\infty}^{\infty} \xi^2 p_x(\xi) d\xi = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N x^2(n)$$





Lượng tử tức thời (không nhớ)

- Luật lượng tử $y = Q(x)$ được định nghĩa:
(L+1) mức tín hiệu $x(0), x(1), \dots, x(L)$
L mức lượng tử hoá
- Mỗi mức lượng tử hoá biểu diễn bằng từ b bit
 $L = 2^b$.
- Sai số lượng tử (tạp âm lượng tử) $e = Q(x) - x$
- Bước lượng tử : hiệu 2 mức tín hiệu kề nhau
$$\delta(i) = x(i) - x(i-1)$$
- Thông lượng $I = bF_s$ (bit/s). F_s : tần số lấy mẫu



Thông lượng (1)

- Tín hiệu lượng tử 8 bit (256 mức), $F_s = 8$ kHz  Thông lượng = 64 kbit/s
- Tín hiệu lượng tử 16 bit (65536 mức), $F_s = 16$ kHz  Thông lượng = 256 kbit/s ,
1 giờ tiếng nói ≈ 100 Mbyte
- Cần phải mã hoá tín hiệu tiếng nói (**MPEG, GSM, G723, ...**) để truyền tiếng nói trên mạng hoặc lưu trữ

Thông lượng (2)

<i>Tần số lấy mẫu (kHz)</i>	<i>Số bit cho 1 mẫu</i>	<i>Thông lượng kbit/s</i>	<i>Dung lượng / phút (kbyte)</i>	<i>Lĩnh vực</i>
48	16	768	11520	Ghi âm chuyên nghiệp
44,1	16	705,6	10584	CD Audio
32	16	512	7680	Radio FM
22	8	264	3960	Radio AM
8	8	64	960	Điện thoại

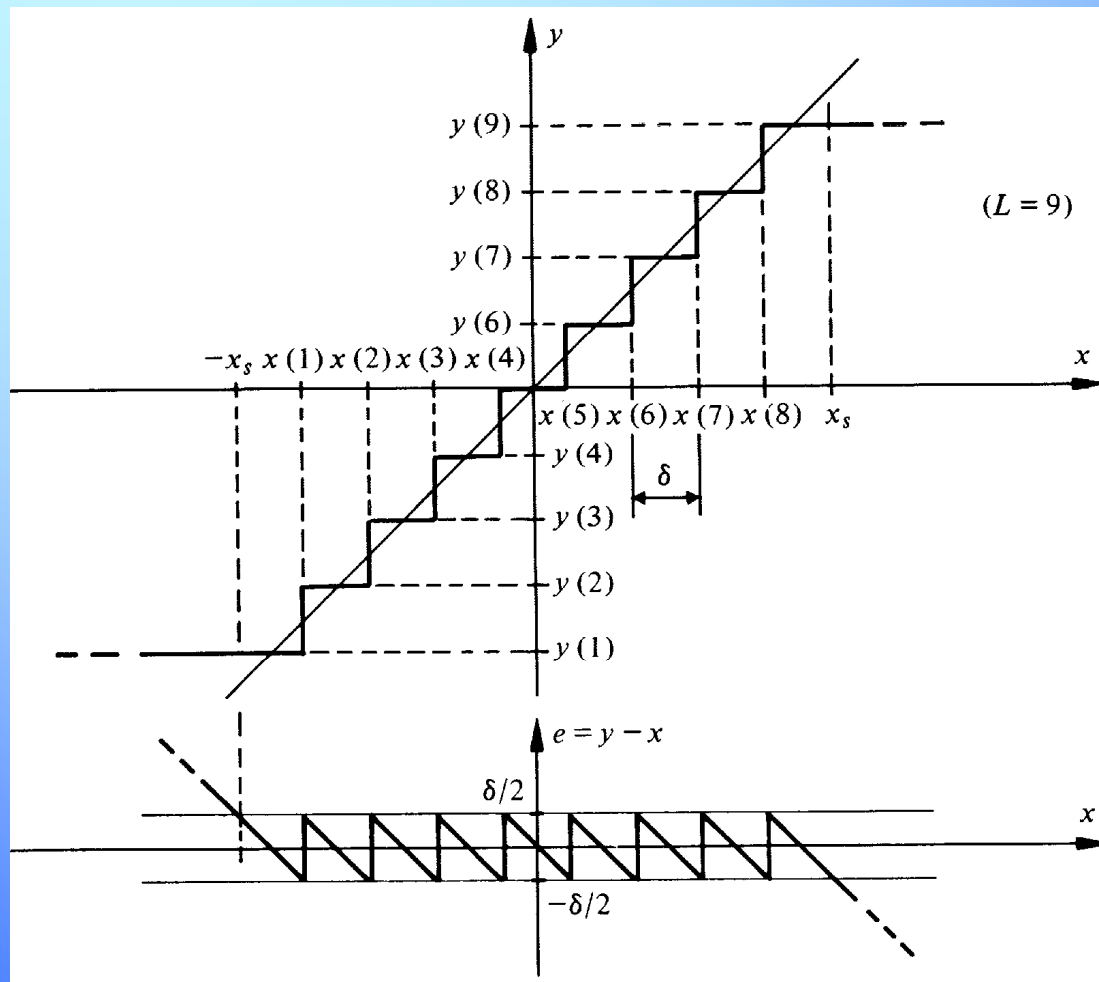
Lượng tử đều

- Tổng quát, bước lượng tử là hàm của biên độ tín hiệu x (lượng tử không đều) \rightarrow đơn giản nhất là lượng tử đều.
- Mức lượng tử được chọn giữa 2 mức tín hiệu
$$y(i) = (1/2)[x(i-1)+x(i)]$$
- Luật lượng tử đều và đối xứng đặc trưng bởi:
 - các mức bão hoà $\pm x_s$
 - mức lượng tử L hoặc $(L+1) = 2^b$.
- Bước lượng tử $\delta = 2x_s/L$



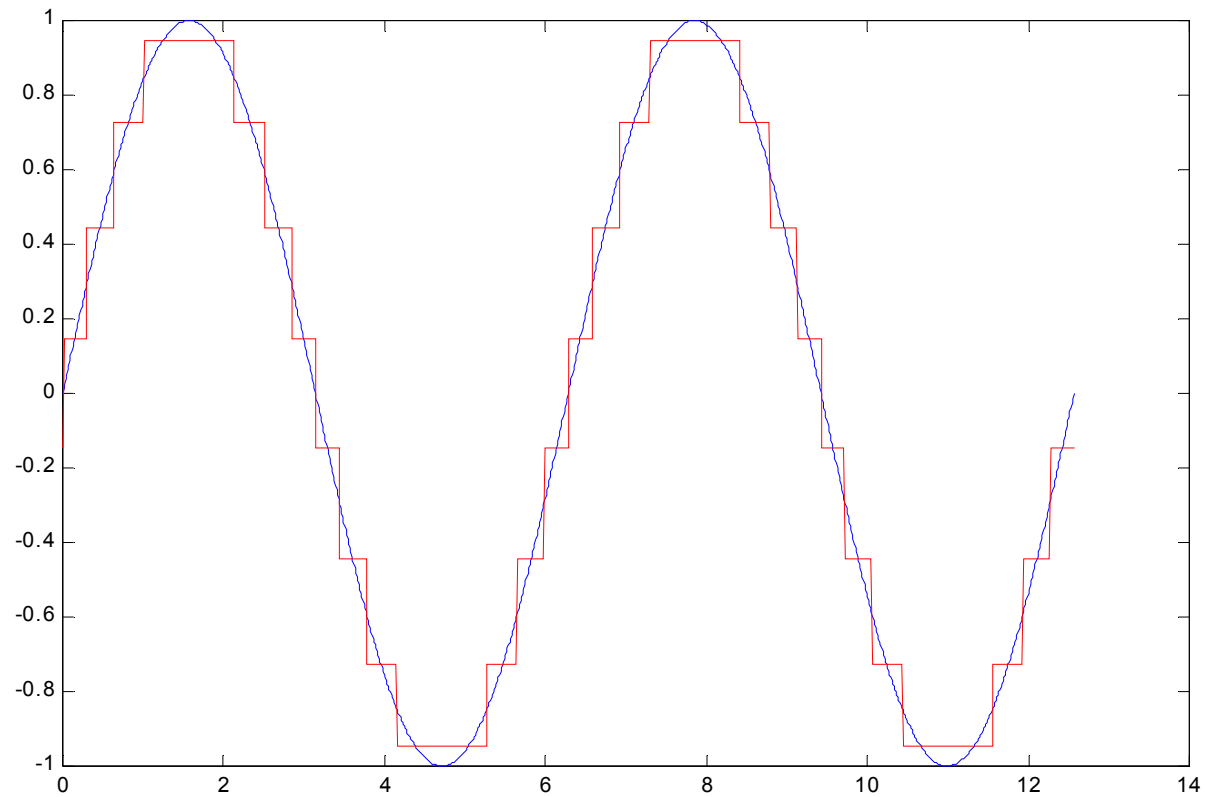
Lượng tử đều

$$L = 9$$



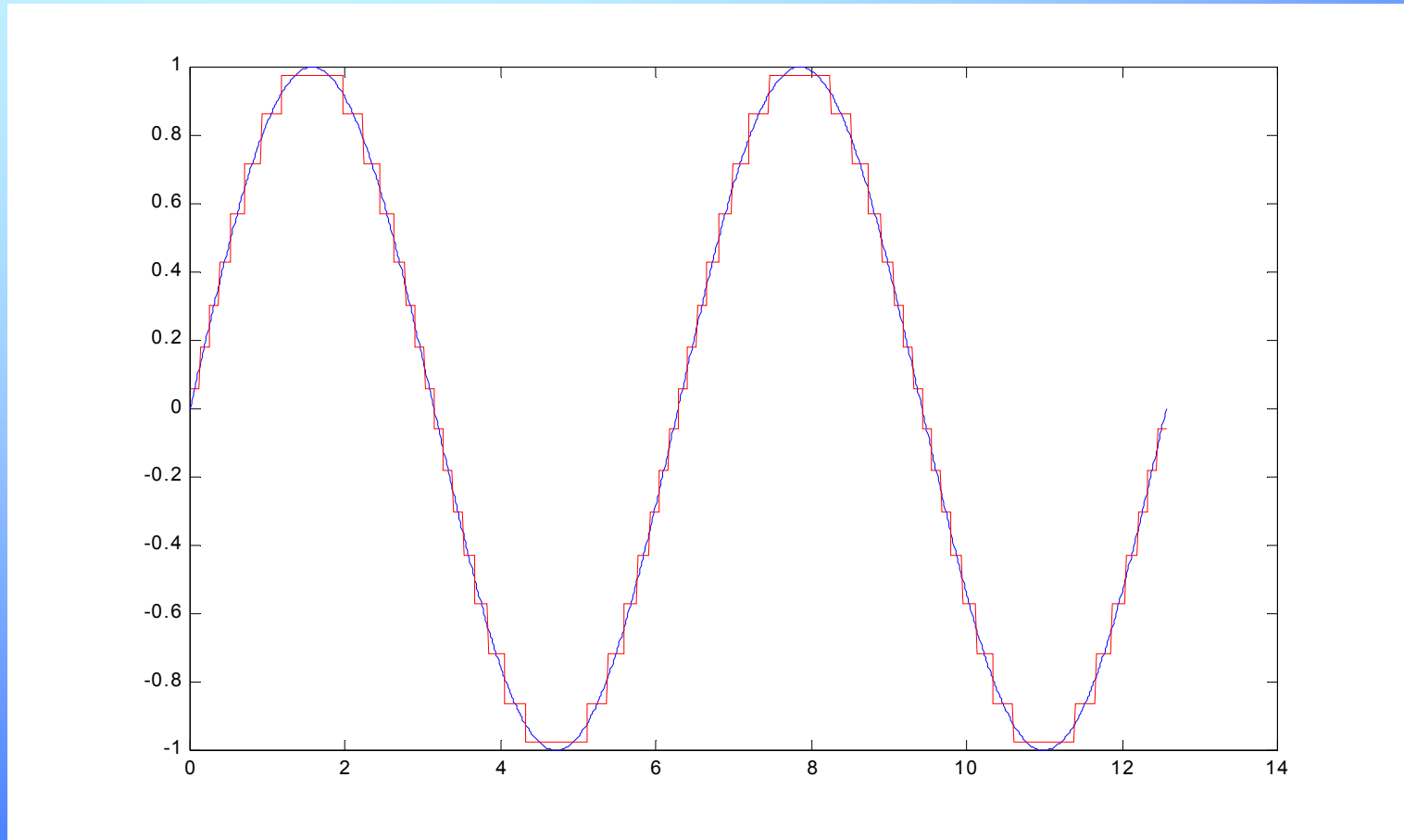
Lượng tử đều

$$L = ?$$

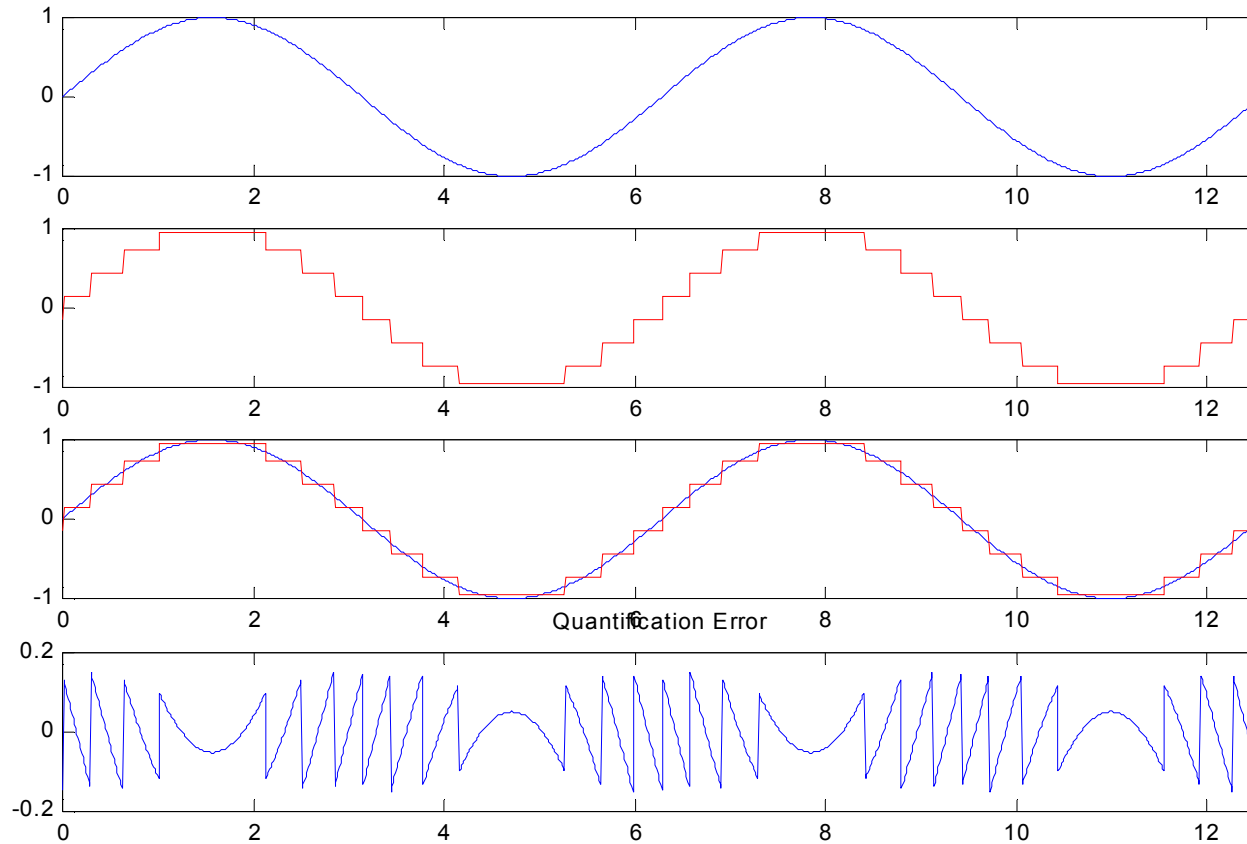


Lượng tử đều

$$L = 16$$



Lượng tử đều



Các tính chất lượng tử đều

- Mật độ xác suất sai số lượng tử

$$p_e(\xi) = \sum_{i=-1}^1 p_x(i\delta + \xi), \quad 1 = (L-1)/2$$

phân bố đều giữa $-\delta/2$ và $+\delta/2$

$$\begin{aligned} p_e(\xi) &= 1/\delta, \quad |\xi| \leq \delta/2 \\ &= 0, \quad |\xi| > \delta/2 \end{aligned}$$

- Trung bình tạp âm lượng tử = 0
- Phương sai

$$\sigma_e^2 = \int_{-\delta/2}^{\delta/2} \xi^2 / \delta \, d\xi = \delta^2 / 12$$

Các tính chất lượng tử đều

- Tỷ số tín hiệu trên nhiễu

$$SN = 10 \lg \left(\frac{\sigma_x^2}{\sigma_e^2} \right) (\text{dB}) = 6,02b + 4,77 - 20 \lg \left(\frac{x_s}{\sigma_x} \right)$$

$$\text{Nếu } x_s = 4\sigma_{\max} \rightarrow SN(\text{dB}) = 6b - 7,3$$

Với $b \geq 6$, tăng 6 dB mỗi khi tăng 1 bit lượng tử

Để có chất lượng thích hợp cần có $b \geq 11$

Tỷ số tín hiệu trên nhiễu

$$SN = \frac{\text{Năng lượng tín hiệu}}{\text{Năng lượng nhiễu}} = \frac{W_s}{W_n}$$

$$SN_{dB} = 10 \log_{10} SN$$

Hoặc

$$SN_{dB} = 20 \log_{10} \frac{\text{Biên độ tín hiệu}}{\text{Biên độ nhiễu}}$$



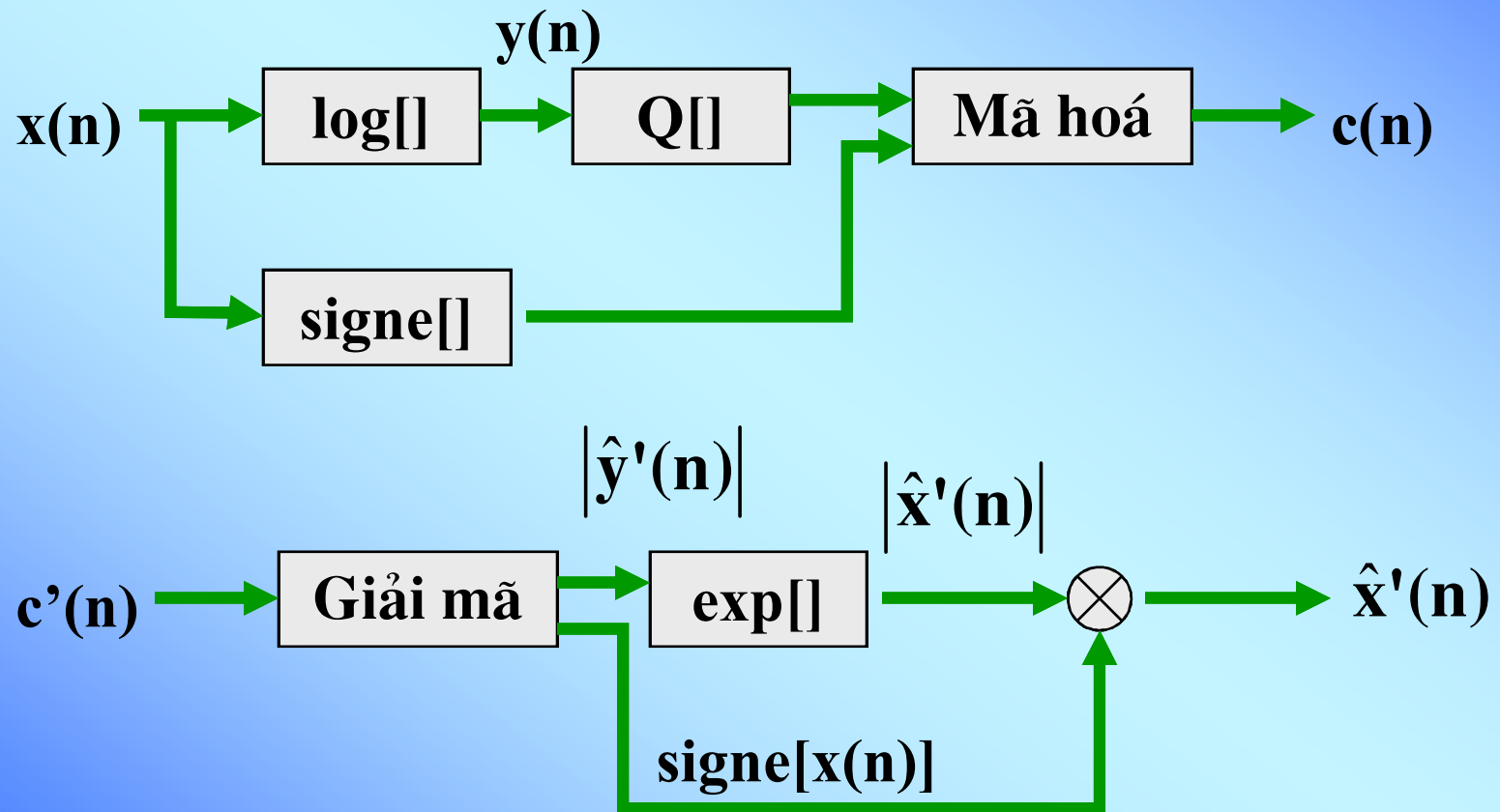
Tỷ số tín hiệu trên nhiễu

<i>Năng lượng</i>	<i>SN (dB)</i>
Tín hiệu = Nhiễu	0
Tín hiệu = 2 Nhiễu	2
Tín hiệu = 10 Nhiễu	10
Tín hiệu = 100 Nhiễu	20
Tín hiệu = 1000 Nhiễu	30
Tín hiệu = 10^N Nhiễu	$N \times 10$



Lượng tử logarit

- Sau khi lấy logarit biên độ tín hiệu sẽ mã hoá tuyến tính

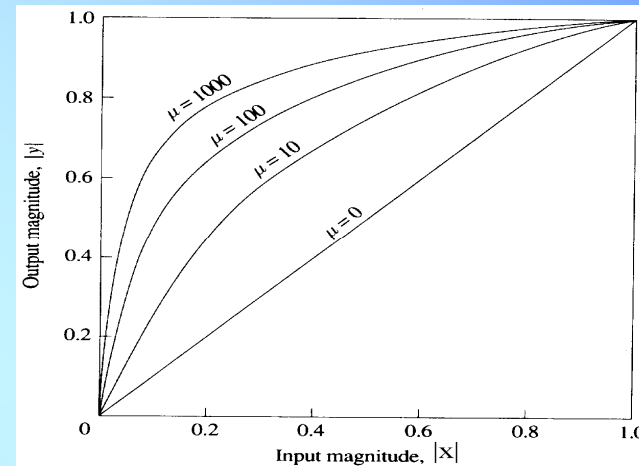
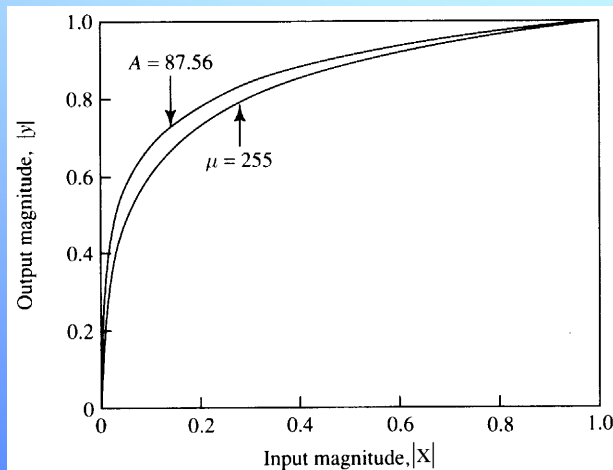


Lượng tử logarit

- Hai giải pháp dùng cho điện thoại

→ Luật μ (dùng ở Mỹ)

$$|y| = \frac{\log(1 + \mu|x|)}{\log(1 + \mu)}$$



→ Luật A (dùng ở châu Âu)

$$|y| = \frac{1 + \log A|x|}{1 + \log A}$$

$$\mu = 255 : A = 87,56$$

◆ 8 bit logarit ~ 12 bit lượng tử đều



Lượng tử thích nghi

- Bước lượng tử tuỳ thuộc vào biên độ tín hiệu

→ Thích nghi trước

