TEMA 2: LA SEÑAL DE VOZ. MODELADO Y ANÁLISIS

1.- INTRODUCCIÓN

La señal de voz es una señal natural, con bastante complejidad respecto a señales artificiales

Aspectos históricos importantes: PCM, Vocoder

Problemas básicos relativos a la señal de voz:

- codificación: representar la señal de voz como un conjunto de bits → compromiso entre calidad y velocidad
- síntesis: producción de voz a partir de texto
- reconocimiento: obtener el texto correspondiente a una señal de voz.
  Necesidad de análisis de voz + contexto (gramática)
- ayudas a discapacidades en producción/obtención de voz.

"Señal de voz" → implica siempre T_x, R_x, y la propia señal

para conocer qué aspectos de la voz son importantes para el oído del receptor

Comportamiento del oído humano ≈ analizadores de espectros

2º responde a las amplitudes de los espectros, no tanto a la fase.

Además, la escala es aproximadamente logarítmica

2º enmascaramiento: capacidad de no poder distinguir tonos cercanos a un tono de gran amplitud.

- aparato fonador: sistema realimentado (oímos lo que hablamos)

(* 2.1 *)

partes correspondientes a
- fuerza/potencia: - laringe, - pulmones, - músculos
- producción de sonido:
  - cuerdas vocales: válvula que se abre y cierra periódicamente (señal sin contenido)
  - cavidades resonantes: conforman la señal

(* 2.2 *) → transparencia "ingeniril"

(* 2.3 *) → las cuerdas se cierran solas tras el paso del aire debido a la presión

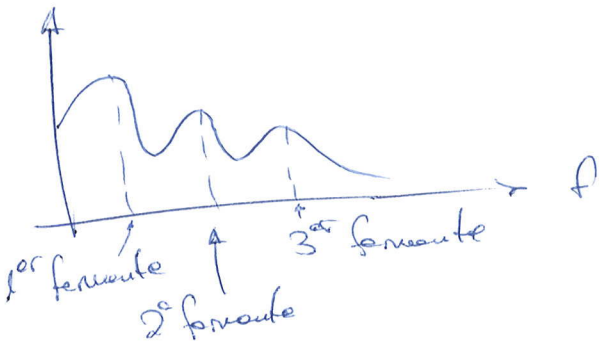
- frecuencia fundamental o "pitch": frecuencia de vibración de las cuerdas vocales. Depende de:
  - sexo: hombre con fundamental más baja (≈ 50-250 Hz) que mujer (≈ 190 - 590 Hz)
  - tensión variable sobre las cuerdas vocales
- cavidad resonante: variante, para formar la señal deseada
  ⇒ señal es estacionaria (para los procesos de interés)

=> uso de modelos lo estacionarios, pero supone dificultad para obtener los espectros

Solución: considerar segmentos de tiempo reducidos => análisis localizado de Fourier

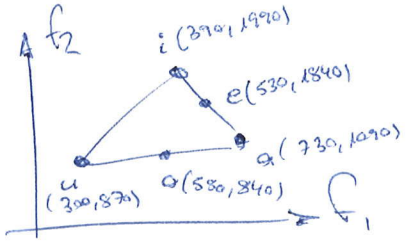
- voz constituida por fonemas, cuya concatenación da lugar a la información (voz: sistema de comunicación digital). Los fonemas se caracterizan por su espectro, y dentro de este, por sus formantes.

ejemplo: /a/



Para las vocales, basta con los 2 primeros formantes => triángulo vocálico

(frecuencias para el hombre)



Se ha intentado obtener un alfabeto universal de fonemas.

ej: el castellano tiene pocos fonemas

- el sonido se genera mediante:
  - cuerdas vocales (vibración)
  - mecanismos al deformar el tracto vocal. ej: /f/

Aparato fonador en continuo movimiento 2ª voz no estacionaria, pero fonemas sí estacionarios

- fonemas:
  - sonidos vocales (de voz) / sonoros (las cuerdas vocales vibran)
    \ sordos
  - vocales: sonidos sonoros puros (cuerdas vocales + conformación), tienen más energía (en mucha diferencia), bastante estacionarios, se pueden mantener tiempo. Se caracteriza bastante bien por sus 2 primeros formantes
  - diptongos: fonemas, no se pueden considerar 2 vocales juntas ya que la sonoridad es distinta.

fonemas:

- líquidos o semivocales: sonoros, con menos energía, se caracterizan por ej. Ej: /l/, /r/
- fricativas: totalmente sordas (/s/, /f/), o con algo de sonoros (/v/, /z/, ...). Poca energía, difíciles de modelar.
- oclusivas: cierre + apertura del tracto vocal

  sordas: /p/, /t/, ...

  sonoras: /b/, /d/, ...

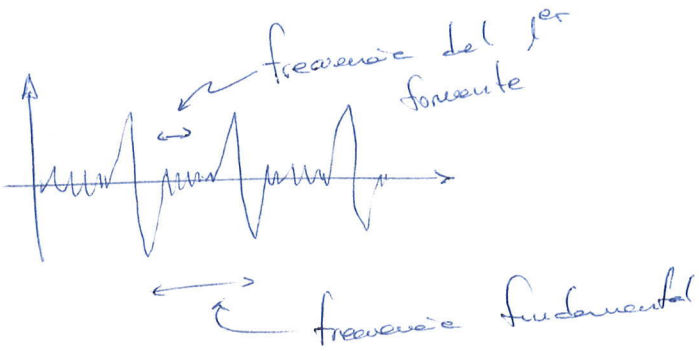
  No estacionario: parte de silencio + parte de señal
- africadas o semioclusivas: no sonoras (/ch/)
- nasales: se identifican bastante bien por el uso de la cavidad nasal (cuelo del paladar cuando cavidades bucal y nasal). Con las 2 cavidades en paralelo aparecen ceros en el espectro.

  Ej: /u/, /m/, ...

- fonemas sonoros se identifican con el pitch

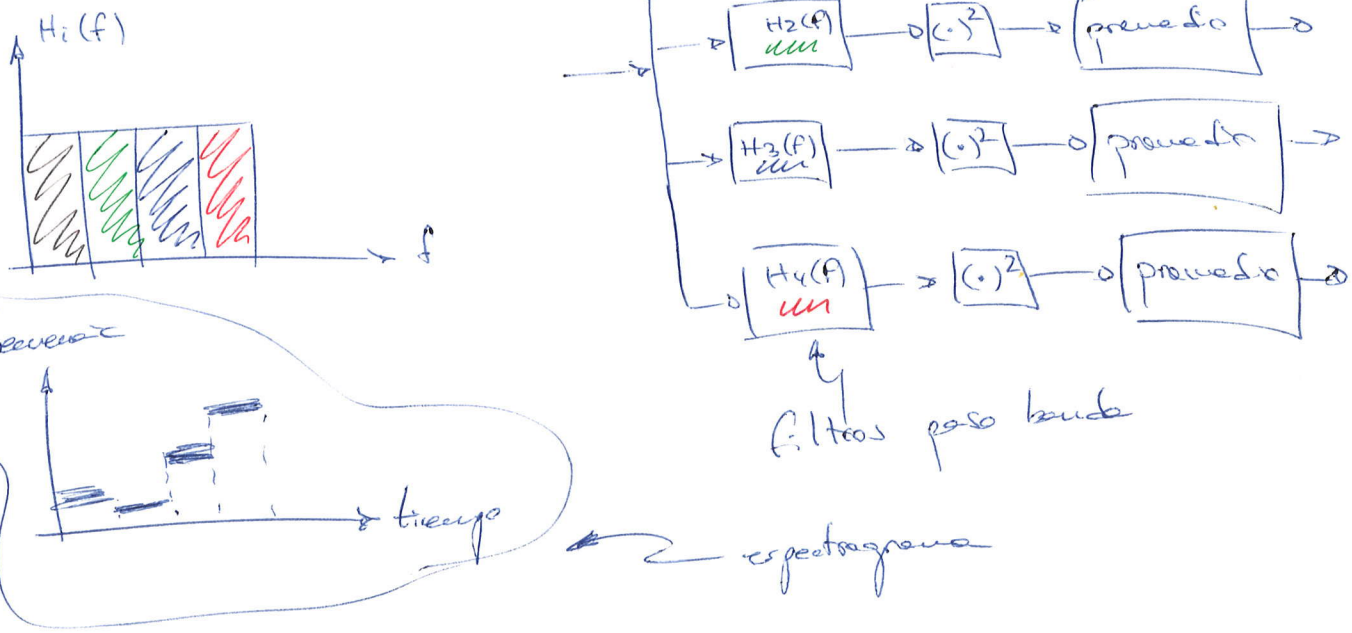
Podemos variar dicha frecuencia: entonación, realce de la voz, canto, ...

(* 2.6 *) → vocales:

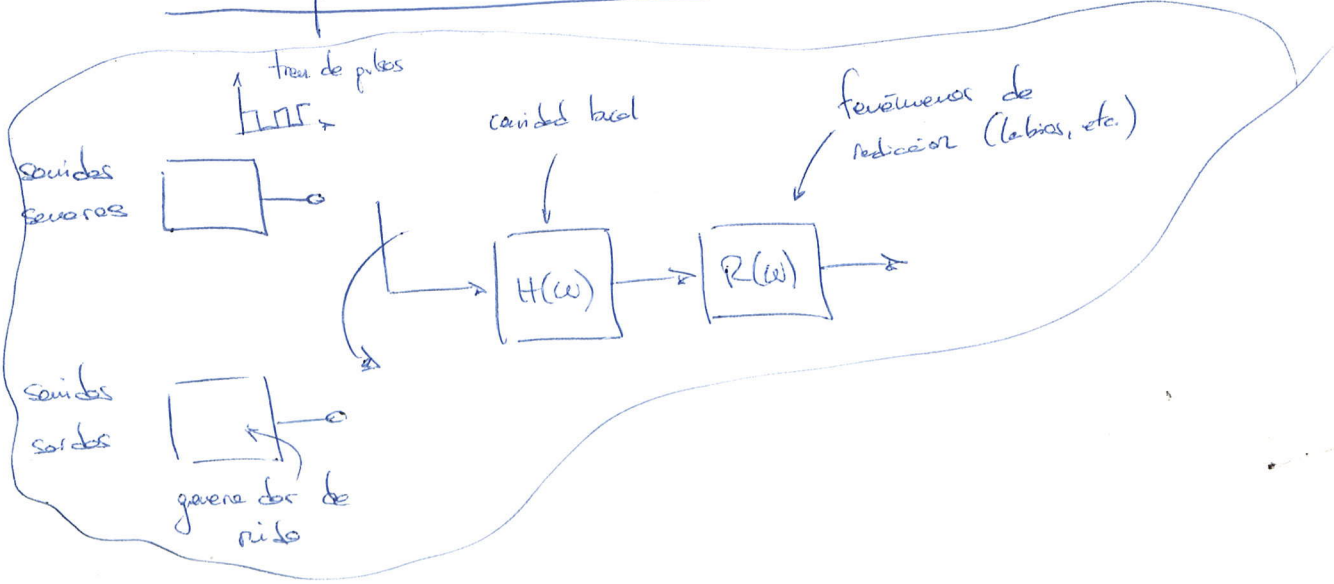


- espectrograma: intento de análisis espectral de señales no estacionarias.

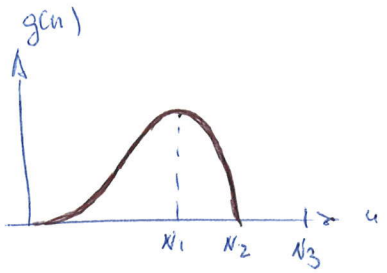
(* 2.7 *)



- modelo preliminar de la señal de voz:



Pulso usado para modelar: "pulso de Rosenberg"

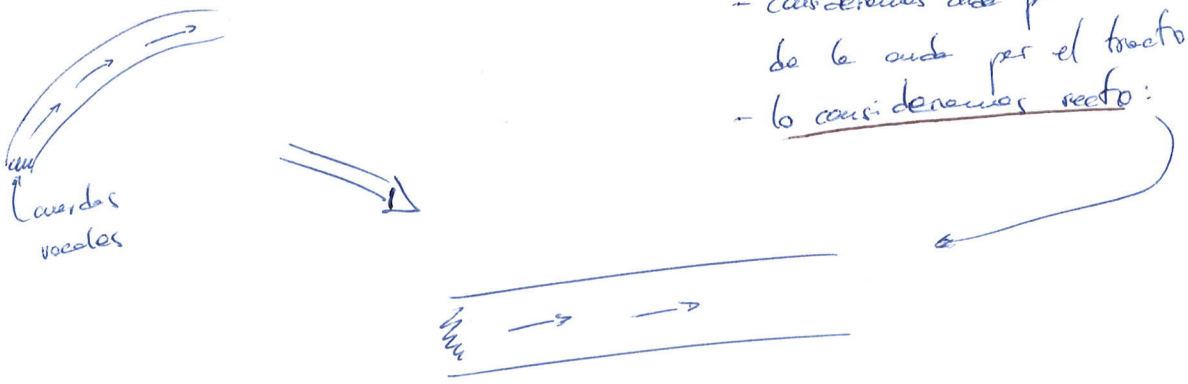


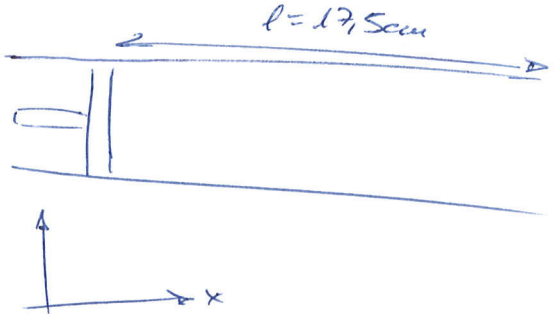
$$\begin{cases} g(u) = \frac{1}{2}\left(1 - \cos\frac{\pi u}{N_1}\right) & 0 \le u \le N_1 \\ g(u) = \cos\pi\frac{u-N_1}{2N_2} & N_1 \le u \le N_1 + N_2 \\ g(u) = 0 & N_1+N_2 < u \le N_3 \end{cases}$$

(*2.4*)

- modelo del tracto vocal: complejo, leyes de la física, etc.

Para cada fonema: hipótesis: - sección invariante con el tiempo
- paredes sin pérdidas, no absorben energía
- consideramos onda plana en la propagación de la onda por el tracto
- lo consideramos recto:





tubo acústico de sección circular constante

Magnitudes:
- velocidad: $v(x,t)$ m/s
- presión: $p(x,t)$
- densidad del gas: ρ
- área de la sección del tubo: A
- velocidad de propagación: $c = 20,1\sqrt{T}$ (T en °K)
- velocidad volumétrica: $u(x,t) = A \cdot v(x,t)$ m^3/s

ecuaciones:

$$-\frac{\partial u}{\partial x} = \frac{A}{\rho c^2}\frac{\partial p}{\partial t}$$

conservación de la masa

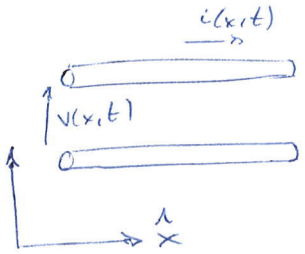
$$-\frac{\partial p}{\partial x} = \frac{\rho}{A}\frac{\partial u}{\partial t}$$

conservación del momento

(tubo uniforme sin pérdidas)

análogas a la línea de transmisión sin pérdidas:

$$\begin{cases} -\dfrac{\partial i}{\partial x} = C\dfrac{\partial v}{\partial t} \\ -\dfrac{\partial v}{\partial x} = L\dfrac{\partial i}{\partial t} \end{cases}$$



L: inductancia / ud. longitud

de esta forma:

gases	línea tx
v	i
p	v
$A/\rho c^2$	C
ρ/A	L

(en acústica se usa: $C = \frac{A}{\rho c^2}$ $L = \frac{\rho}{A}$)

C: compresibilidad

L: inductancia acústica

$$u(x,t) = u^+\left(t - \frac{x}{c}\right) - u^-\left(t + \frac{x}{c}\right)$$

$$p(x,t) = Z_0\left(u^+\left(t - \frac{x}{c}\right) + u^-\left(t + \frac{x}{c}\right)\right)$$

donde $Z_0 = \frac{\rho c}{A}$: impedancia acústica

análisis armónico: entrada/excitación ⇒ señal exponencial compleja

$$u(0,t) = u(0)e^{j2\pi ft} = e^{j2\pi ft}$$ (podemos hacer $u(0)=1$)

(consideramos también $p(l,t)=0$ ← tubo abierto)

$$u(x,t) = u(x)\,e^{j2\pi ft}$$
$$p(x,t) = P(x)\,e^{j2\pi ft}$$

condiciones de contorno:

$$\begin{cases} u(0) = 1 \\ P(l) = 0 \end{cases}$$

$$-\frac{\partial u(x)}{\partial x} = \frac{A}{\rho c^2} p(x)\, j2\pi f$$
$$-\frac{\partial p(x)}{\partial x} = \frac{\rho}{A} u(x)\, j2\pi f$$

$$\frac{\partial^2 u(x)}{\partial x^2} + \frac{4\pi^2 f^2}{c^2} u(x) = 0$$
$$\frac{\partial^2 p(x)}{\partial x^2} + \frac{4\pi^2 f^2}{c^2} p(x) = 0$$

ecuaciones de onda

Soluciones: $u(x) = a_1 e^{\gamma x} + a_2 e^{-\gamma x}$

$P(x) = a_3 e^{\gamma x} + a_4 e^{-\gamma x}$

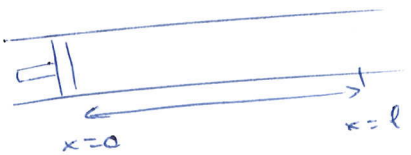
donde $\gamma = \sqrt{-\frac{4\pi^2 f^2}{c^2}} = j\frac{2\pi f}{c}$

(...) =

$$P(x) = j Z_0 \frac{\operatorname{sen} \frac{2\pi f}{c}(l-x)}{\cos \frac{2\pi f}{c} l}$$

$$u(x) = \frac{\cos \frac{2\pi f}{c}(l-x)}{\cos \frac{2\pi f}{c} l}$$

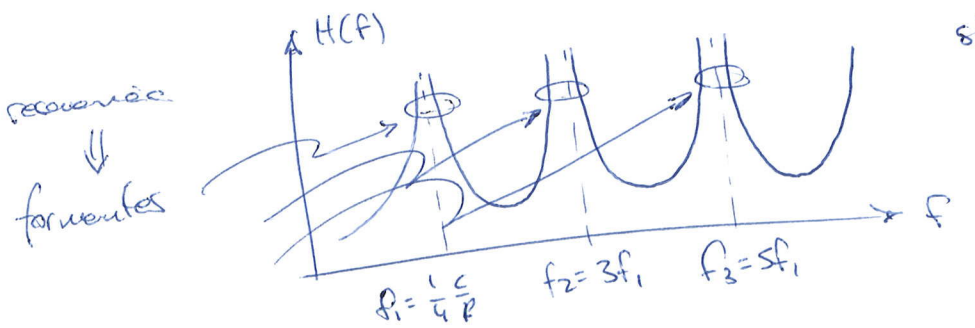
soluciones al existir con exponencial compleja



$u(0,t) = e^{j2\pi ft}$ $\quad u(l,t) = u(l)\,e^{j2\pi ft}$

⇒ respuesta:

$$H(f) = \frac{u(l)}{u(0)} = \frac{1}{\cos \frac{2\pi f}{c} l}$$



si: $\frac{2\pi f}{c} l = \frac{\pi}{2} + k\pi \Rightarrow$ polos

$\Rightarrow f_1 = \frac{1}{4}\frac{c}{l}$

→ esto es para tubo recto, abierto, sin pérdidas

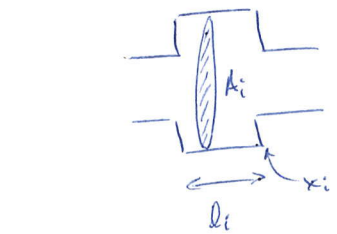
- paralelismo con línea de transmisión terminada en cortocircuito.
- en la práctica siempre hay pérdidas $\Rightarrow$ no será inestable
- señal excitadora: $S_e(f) \rightarrow S_o(f) = S_e(f)\,|H(f)|^2$

  Esta $H(f)$ es propia de sistemas distribuidos

  Como $v \cdot i$ = potencia $\Rightarrow$ **potencia acústica = u.P**

- <u>caracterización de tubos</u>: tenemos ahora una cavidad de tubos de diferentes longitudes y secciones

  sistema LTI $\Rightarrow$ respuesta en frecuencia



Para cada tubo "i", tenemos:

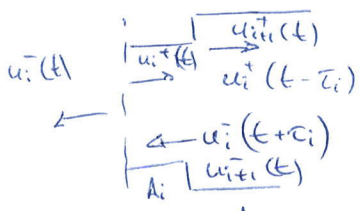
- $u_i(x,t) = u_i^+(t - x/c) - u_i^-(t + x/c)$
- $p_i(x,t) = z_i\left(u_i^+(t - x/c) + u_i^-(t + x/c)\right)$

Nos interesa especialmente lo que pase en las transiciones $\Rightarrow$ <u>continuidad</u>:

- $u_i(x_i,t) = u_{i+1}(x_i,t)$
- $p_i(x_i,t) = p_{i+1}(x_i,t)$

· tiempo de propagación de la onda en un tubo:

$$\tau_i = \frac{l_i}{c}$$



$$z_i = \frac{\rho c}{A_i}$$

Continuidad

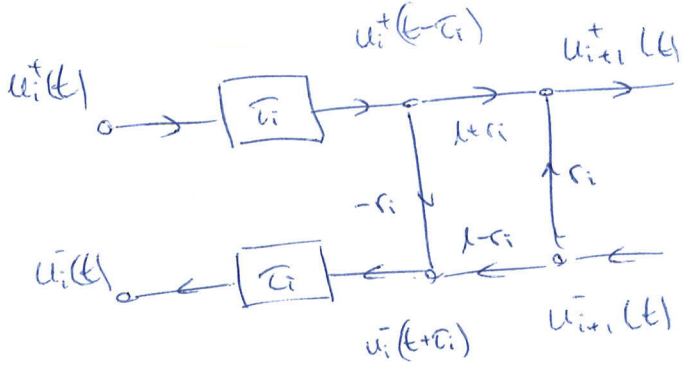
$$u_i(t, l_i) = u_i^+(t-\tau_i) - u_i^-(t+\tau_i) = u_{i+1}(t, 0) =$$

$$= u_{i+1}^+(t) - u_{i+1}^-(t)$$

$$P_i(t, l_i) = z_i\left(u_i^+(t-\tau_i) + u_i^-(t+\tau_i)\right) = P_{i+1}(t, 0) =$$

$$= z_{i+1}\left(u_{i+1}^+(t) + u_{i+1}^-(t)\right)$$

Representación de la conexión mediante grafos



$$r_i = \frac{z_i - z_{i+1}}{z_i + z_{i+1}}$$

$$|r_i| \leq 1$$

El diagrama de grafos se puede expresar matricialmente:

$$\begin{bmatrix} u_{i+1}^+(t) \\ u_i^-(t+\tau_i) \end{bmatrix} = \begin{bmatrix} 1+r_i & r_i \\ -r_i & 1-r_i \end{bmatrix} \begin{bmatrix} u_i^+(t-\tau_i) \\ u_{i+1}^-(t) \end{bmatrix}$$

(ondas salientes) (ondas entrantes)

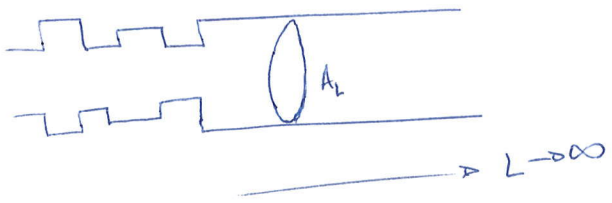
Sistema completo → cuaternario de varios tubos

Necesitamos saber además qué ocurre al principio y al final de la cuaternario completa

- Comienzo: globis
- Final: bobias

(* 2.18 *)

Bobias: Si Z_L es real, no hay onda reflejada

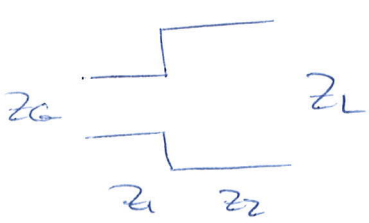


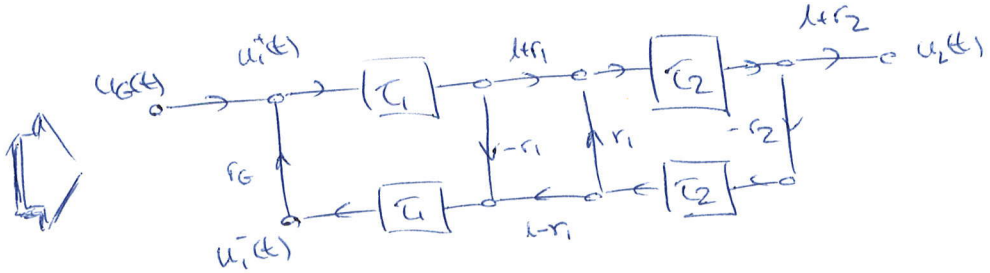
$$Z_L = \frac{\rho c}{A_L}$$

Globis:

$$\left.\begin{aligned} u_1(t) &= u_1^+(t) - u_1^-(t) \\ p_1(t) &= Z_1\left(u_1^+(t) + u_1^-(t)\right) \\ u_1(t) &= u_G(t) - \frac{p_1(t)}{Z_G} \end{aligned}\right\} \Rightarrow u_1^+(t) = \frac{1+r_G}{2} u_G(t) + r_G u_1^-(t)$$

Ejemplo:





Ya tenemos ecuaciones suficientes para calcular la respuesta en frecuencia Z_0 (* 2.19 - 2.20 *)

Lo hemos hecho suponiendo que son todos sistemas LTI, aunque la física puede no serlo (ángulo de impedancia variable), pero evidentemente complica innecesariamente el análisis.

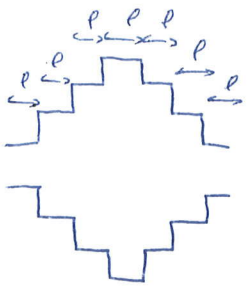
Ejercicio:

Respuesta en frecuencia de 2 tubos concatenados.

Solución: $$H(f) = \frac{\frac{1}{2}(1+r_G)(1+r_1)(1+r_2)}{1+r_1r_2\gamma_2^2 + r_1r_G\gamma_1^2 + r_Gr_2\gamma_1^2\gamma_2^2}$$

$$\gamma_i = e^{-j2\pi f\tau_i}$$

Si se particulariza para el caso de un modelo de tubos, todos con la misma longitud l, la respuesta en frecuencia que se obtiene es periódica de periodo $1/2\tau$



$$\Leftrightarrow \quad H(f) = H\left(f + \frac{k}{2\tau}\right) \qquad \tau = l/c$$

$H(f)$ periódica $\Rightarrow$ sistema discreto

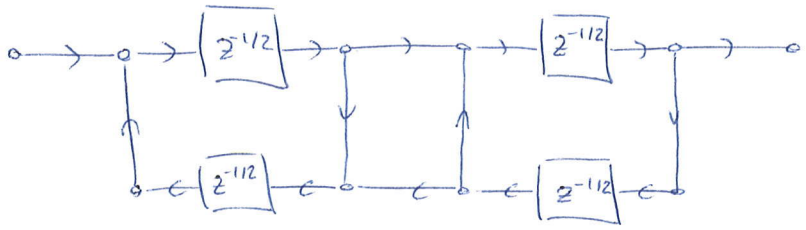
Ritmo de muestreo apropiado para este sistema:

$$\boxed{T_M = 2\tau}$$

Retardo de una muestra discreta: $T_m \rightarrow z^{-1}$

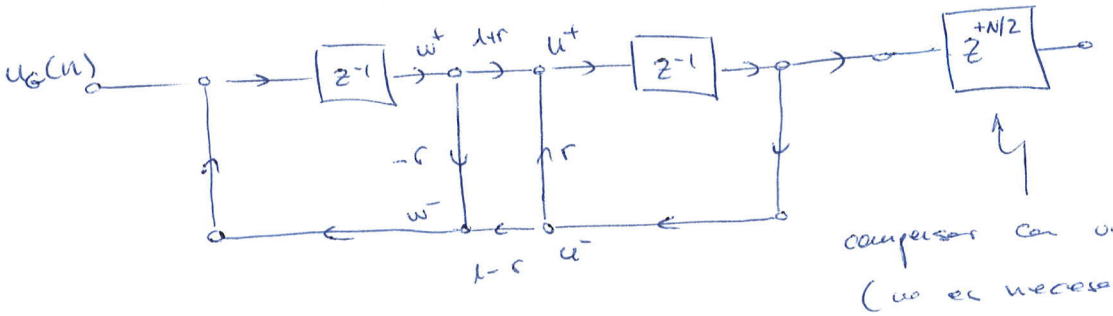
referido $\tau = \frac{T_m}{2} \rightarrow z^{-N/2}$

(grafo para la estructura digital



medio muestra
⇒ interpolador (valores)

En cada lazo cerrado tenemos z^{-1}, podemos agrupar:



Sistema discreto equivalente a 2 tubos acústicos de longitud constante.

Respuesta en frecuencia: (*2.21*)

- Comentarios sobre el modelo discreto:

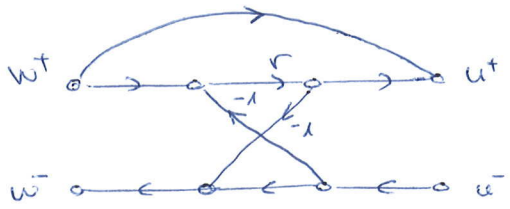
  · las operaciones que se realizan en una iteración se pueden simplificar:

$$u^+ = (1+r)w^+ + ru^- = w^+ + r(w^+ + u^-)$$

$$w^- = -rw^+ + (1-r)u^- = u^- - r(w^+ + u^-)$$

↑ reducir el número de multiplicaciones

otro grafo ↴



• relación entre τ, L, N, c si $L = 17{,}5$ cm, $c = 35 \cdot 10^3$ cm/s

$$l = \frac{L}{N} \qquad \tau = \frac{l}{c} = \frac{L}{N \cdot c} = \frac{T_m}{2}$$

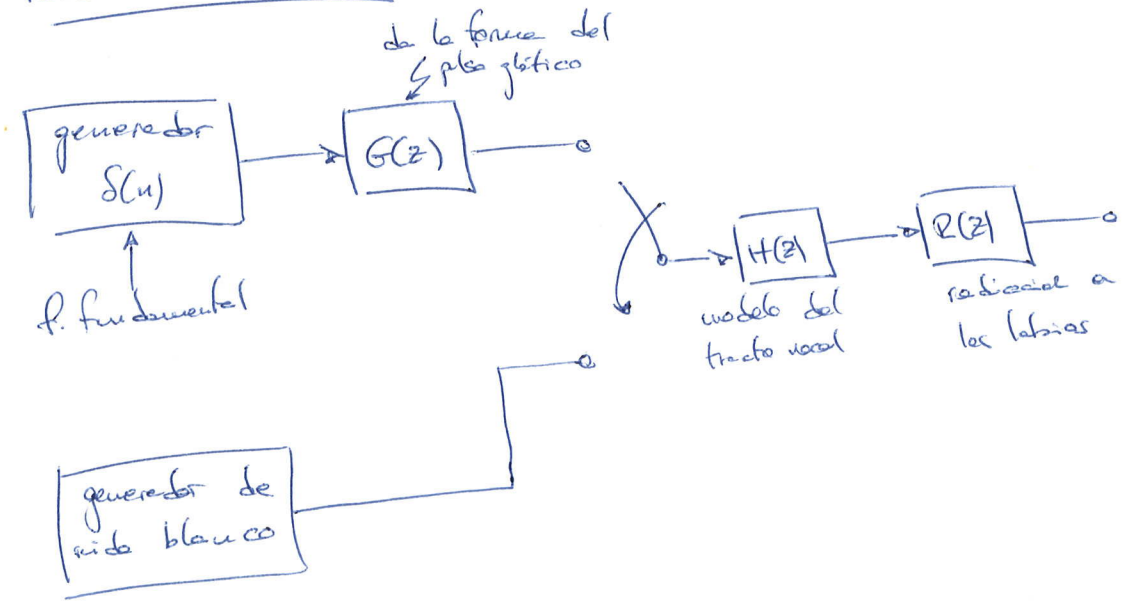
$$\frac{N}{2} = \frac{L}{c}\,\frac{1}{T_m} = \frac{1}{2}\,10^{-3} f_m = 10^{-3} B$$

$B = \frac{1}{2} f_m$ = banda que se está modelando $\boxed{\frac{N}{2} = B\,[\text{kHz}]}$

N { número de tubos
  { número de polos ⇒ nº de picos de resonancia

Cada pico está a 1kHz del otro aproximadamente

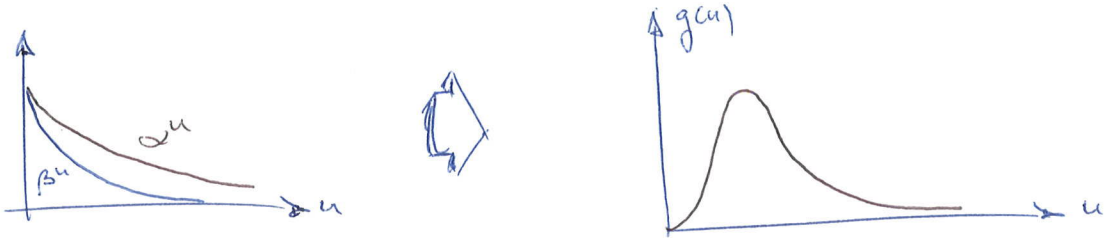
-MODELO MÁS COMPLETO:



$g(u)$ = pulso glótico:

- Rosenberg
- de duración infinita y tipo exponencial:

$$\boxed{g(u) = (\alpha^u - \beta^u)\, u(u)} \qquad 1 > \alpha > \beta > 0$$



Ventaja de este tipo de pulsos: la respuesta impulsiva del sistema que los genere sólo tiene polos

- El modelo de tracto vocal se suele hacer con sistemas sólo polos.

Los sonidos nasales tienen ceros, pero se pueden modelar, con un orden suficiente, con sistemas sólo polos.

$$H(z) = \frac{1}{1 - \sum_{k=1}^{N} a_k z^{-k}}$$

- $R(z) = 1 - z_0 z^{-1}$ $\quad z_0 \approx 1$ $\quad z_0 < 1$ (paso alto)

Normalmente se modela sólo con $H(z)$, obviando $G(z)$ y $R(z)$.

2.- MÉTODOS DE ANÁLISIS LOCALIZADO (STP)

Son imprescindibles para analizar la señal de voz, y generales en el análisis de señales no estacionarias.

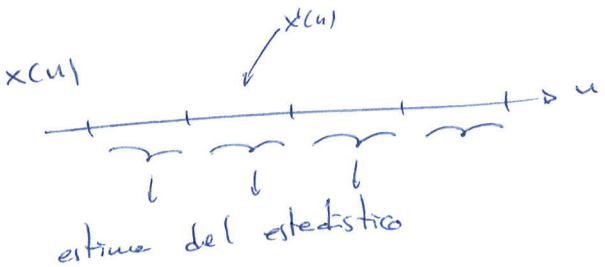
STP = Short Time Processing

Se basan en la dificultad de hacer análisis estadístico de señales no estacionarias. Se supone que la señal, en un segmento corto de duración, se comporta como una señal estacionaria.

El tamaño del segmento es determinante:

- demasiado grande: señal muy variable, sin estacionariedad
- demasiado corta: señal con pocas muestras (mejor resolución temporal), estimación del estadístico mala.

La idea es buscar segmentos en los que el estadístico no varíe demasiado, pero en el que entren varios periodos de frecuencia fundamental.



Los segmentos no tienen porqué ser contiguos, e incluso pueden solaparse

A veces puede interesarnos aplicar enventanado a los segmentos:

$$s(u) = x'(u)\, w(u) \xleftrightarrow{TF} S(f) = X'(f) * W(f)$$

y se extraen los estadísticos de $s(u)$

Usuales: rectangular y Hamming

Para el tamaño de la ventana no hay ninguna regla, y básicamente dependerá del estadístico que pretendas obtener. Requisitos:

1.- Que el estadístico no varíe mucho en el tiempo

2.- Que la ventana nos permita hacer una estima suficientemente precisa

3.- Tener un tamaño de ventana suficiente cuando queramos ver una propiedad como la frecuencia del fundamental: habrá que ver unos cuantos periodos

En voz, tamaños típicos de ventana van entre 3 y 30 ms

Duración de un fonema típico: 80 ms

La duración del segmento puede estar determinada también por el tipo de sonido:

- oclusivo → muy rápido
- vocal → bastante estacionario

Aplicaciones típicas:

- Detección del comienzo y el final de las palabras y los fonemas
- Clasificar el tipo de sonido: silencio, sordo, sonoro, ...
- Estimar la frecuencia fundamental o los formantes
- Estimar los parámetros del modelo

El análisis realizado se puede distinguir en los dominios temporal y frecuencial.

2.1.- ANÁLISIS LOCALIZADO EN EL TIEMPO

Características que veremos:

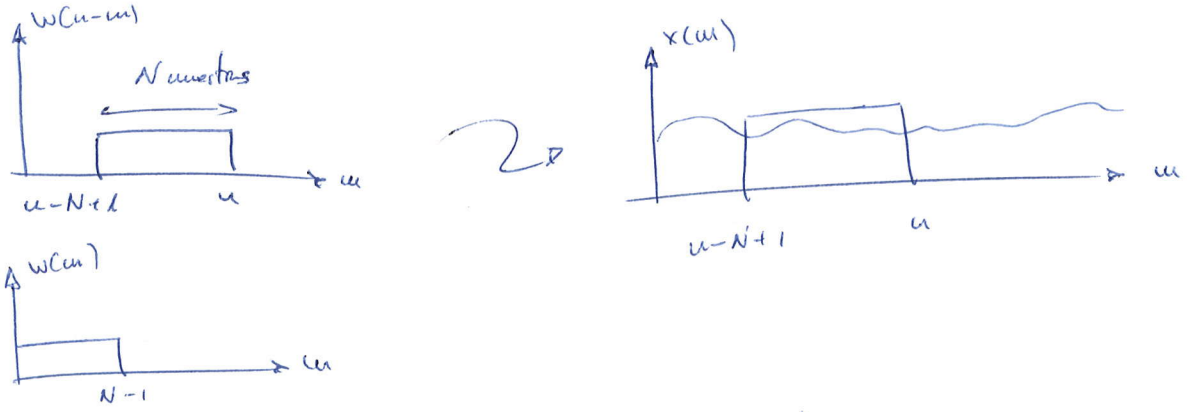
- Energía local o localizada
- Magnitud local
- Cruces por cero
- Autocorrelación localizada

} todas tienen una dependencia temporal → subíndice u

- **ENERGÍA LOCALIZADA** ("energía localizada en el instante u")

$$E_u = \sum_{m=-\infty}^{\infty} \left[x(m)\, w(u-m) \right]^2$$

Ventana de duración finita, N:



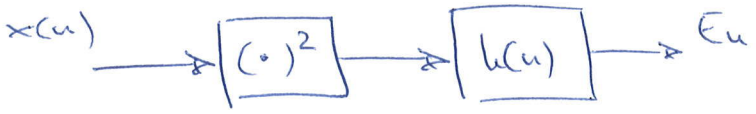
Se calcula la energía de las muestras que van de $u-N+1$ a u.

En una señal no estacionaria, E_u irá variando según el tramo.

Como la ventana es de duración finita:

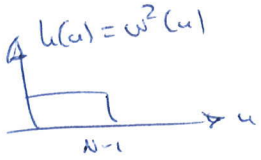
$$E_u = \sum_{m=u-N+1}^{u} \left(x(m)\, w(u-m) \right)^2 = \sum_{m=u-N+1}^{u} x^2(m)\, w^2(u-m)$$

$h(m) = w^2(m) \rightarrow$ $E_u = x^2(u) * h(u)$ $\rightarrow$ filtrado FIR

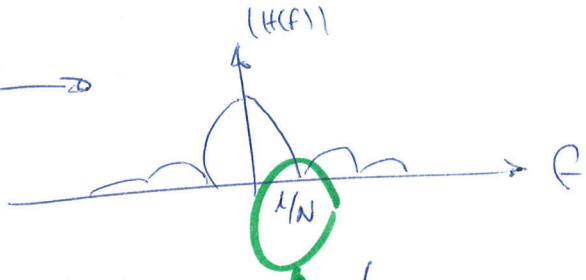


Rectangular y Hamming: retardo $\frac{N-1}{2}$ muestras con fase lineal

ventana rectangular →

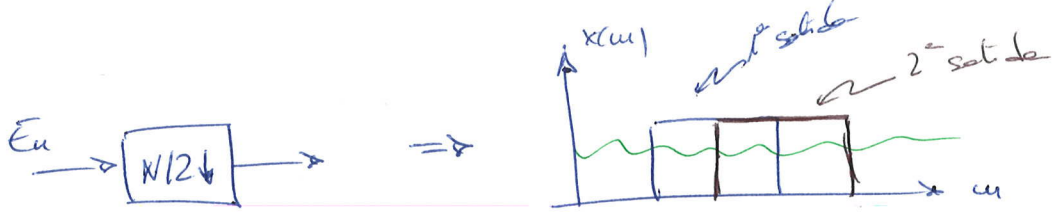


paso bajo: →



En el paso bajo, tanto más abrupto cuanto mayor sea N.

Las señales paso bajo admiten diezmado, y en este caso es usual hacer un diezmado en N/2:



se puede reconstruir la señal original tras el diezmado

⟹ no hay que computar todas las muestras en el filtro FIR

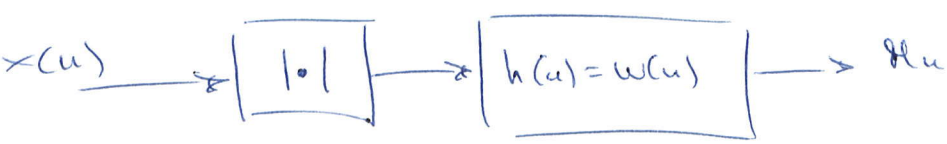
MAGNITUD LOCALIZADA:

$$M_u = \sum_{m=-\infty}^{\infty} |x(m)|\, w(u-m)$$

es interesante a veces cuando se usa aritmética en punto fijo, y que reduce el rango dinámico:

$$\frac{\max\{M_u\}}{\min\{M_u\}} = \sqrt{\frac{\max\{E_u\}}{\min\{E_u\}}}$$

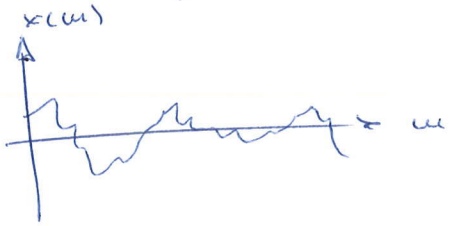
En la señal de voz, Eu puede variar varias decenas de dB



- CRUCES POR CERO:

Sirve para llevar una cuenta de las veces que una señal pasa por cero. Caracteriza muy bien ciertas señales (ej: ≈ sinusoidales)

Se utiliza porque es fácil de calcular y además es significativa del tipo de forma de voz que se está analizando.



$$Z_n = \sum_m \frac{1}{2} \left| \text{sign}(x(m)) - \text{sign}(x(m-1)) \right| w(n-m)$$

(de el número de cruces por cero en la ventana)

Para que la cuenta sea exacta, la ventana debe ser rectangular de valor $1/N$, y nos dará el número de cruces por cero por cada muestra.

Se puede asociar el número de cruces por cero de una sinusoide:

f_m → frecuencia de muestreo

f_0 → frecuencia de la sinusoide

$\frac{f_m}{f_0}$ = # de muestras por periodo de la sinusoide

En un periodo de la sinusoide hay 2 cruces por cero:

$\frac{2}{f_m/f_0} = \frac{2f_0}{f_m}$ = tasa de cruces por cero por cada muestra (para una sinusoide)

Se puede intentar estimar su frecuencia midiendo sus cruces por cero.

- Sonidos sonoros: $E_n \uparrow$, $Z_n \downarrow$
- Sonidos sordos: $E_n \downarrow$, $Z_n \uparrow$
- Ruido de fondo: $E_n \downarrow$, $Z_n \downarrow$

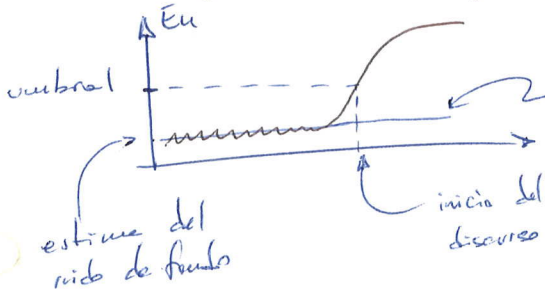
↪ importante para distinguir sonidos sordos del ruido ambiente

- APLICACIÓN: DISCRIMINACIÓN VOZ/SILENCIO EN UN DISCURSO HABLADO:

Determinar donde el locutor habla y dónde hay silencio

- Caso 1: grabación en estudio

Se puede hacer de manera que sea fácil determinar dónde hay silencio a partir de la energía localizada.



umbral por ejemplo: 45

N no puede ser muy largo, o perderemos resolución temporal.
Esto presenta problemas en ambientes ruidosos.

Ejemplos:
- energía local → (*2.22*)
- magnitud local → (*2.23*)
- cruces por cero → (*2.23*)

| ~ 10 ms de duración de la ventana

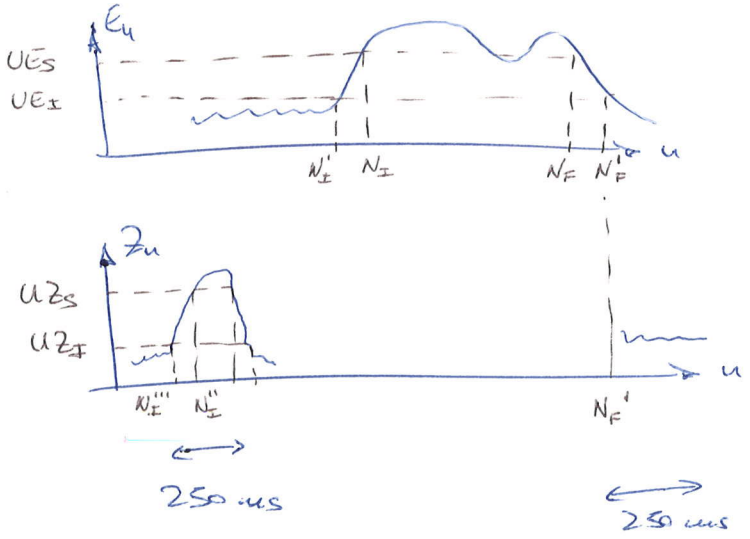
- Caso 2: ambiente ruidoso

Fricativas con poca energía pueden confundirse con el ruido ambiente, por lo que la energía localizada no es suficiente.

Se usan la magnitud o la energía local y los cruces por cero para distinguir dónde están los problemáticos.

	(Energía) E_u	(componentes espectrales) Z_u
Sonidos sonoros	Alta	Pocos
Sonidos sordos	Baja	Muchos
Ruido de fondo	Baja	~

← mayor tasa de cruces por cero que el ruido de fondo



La estimación de las magnitudes E_u y Z_u del ruido de fondo, nos permite establecer umbrales de energía UE (Superior e Inferior) y de cruces por cero UZ (Sup. e Inf.)

Superior → muy conservador, el ruido de fondo nunca lo alcanza

N_I, N_F instantes de inicio y final, muy conservadores

Ahora se puede trazar el umbral inferior, que el ruido puede superar, pero ya tenemos conocimiento de donde hay palabra con certeza → N'_I, N'_F

Aún puede darse que haya sonidos en baja energía

⇒ cruces por cero

UZS → muy conservador también

Se observan los cruces por cero desde el comienzo de la palabra hasta que transcurren 10 segundos de voz de 25ms (250ms) $\Rightarrow N_F''$

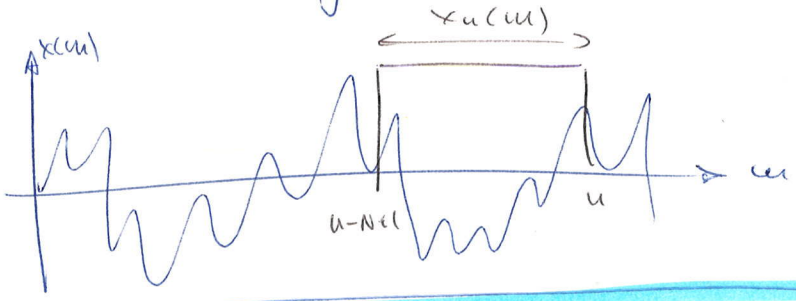
$UZ_T \Rightarrow N_F'''$

Al final, se observan 250 ms desde N_F' y si no hay cruces por cero por encima del umbral superior, se da por bueno ese instante como final de palabra (que da superar UZ_T)

Primero se mira la energía, ya que las palabras tienen valor, que son de alta energía $\Rightarrow$ energía más significativa que cruces por cero

- AUTOCORRELACIÓN LOCALIZADA:

Es el estimador sesgado de la autocorrelación del segmento, sin normalizar



$$R_u(k) = N r_{x_u}(k) = \sum_m x_u(m)\, x_u(m-k) = x_u(k) * x_u(-k)$$

Suponemos $k \geq 0$ siempre

$x_u(m) = x(m)\, w(u-m)$

$X_u(f) = TF\{x_u(m)\}$ → transformada de fourier localizada

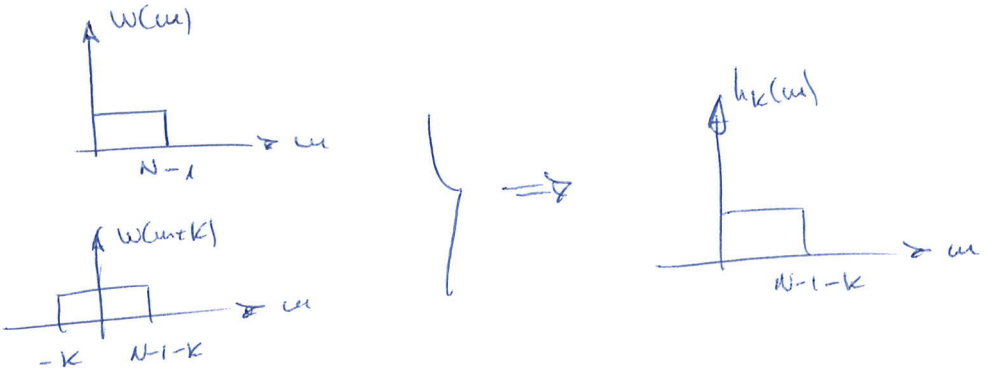
$\Rightarrow TF\{R_u(k)\} = X_u(f) \cdot X_u(-f) = |X_u(f)|^2$

señales reales: $X_u(-f) = X_u^*(f)$

$$R_u(k) = \sum_m x(m)\, w(u-m)\, x(m-k)\, w(u-m+k) = \{h_k(m) = w(m)\, w(m+k)\} =$$

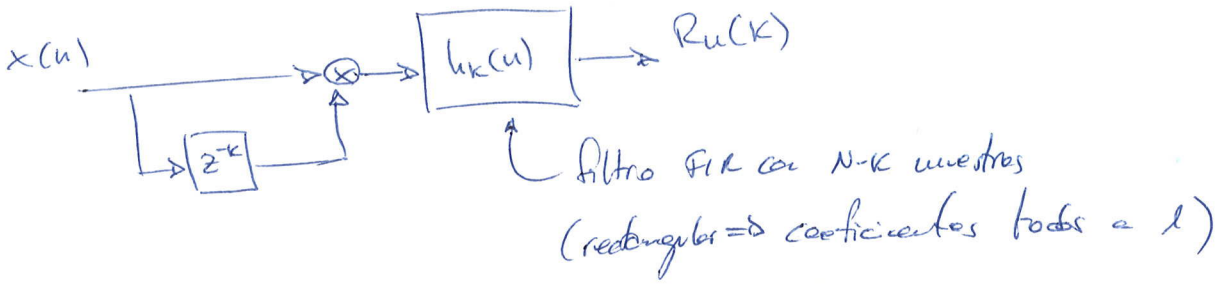
$$= \sum_m x(m)\, x(m-k)\, h_k(u-m)$$

Si tenemos la ventana rectangular:

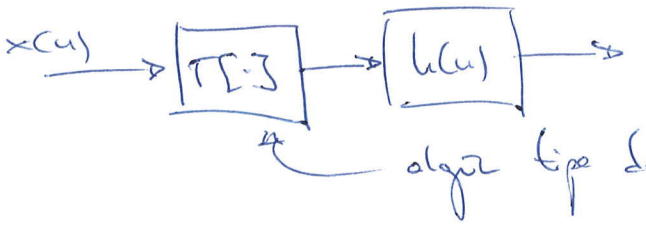


$$\Rightarrow R_x(k) = (x(n)\,x(n-k)) * h_k(n)$$

Podemos considerar que la autocorrelación localizada es:



Todas estas magnitudes responden a esta estructura:



	h(n)	T[ ]
E_n	$w^2(n)$	$(\cdot)^2$
μ_n	$w(n)$	$\lvert\cdot\rvert$
Z_n	$w(n) = 1/N$	$\frac{1}{2}(\text{sign}(x(n)) - \text{sign}(x(n-1)))$
$R_x(k)$	$w(n)\,w(n+k)$	$x(n)\,x(n-k)$

- Problemas de la autocorrelación localizada:

La duración del filtro $h_k(n)$ depende de k ($N-k$ muestras)

$$E[R_u(k)] = (N-|k|)\gamma_x(m)$$

$\Rightarrow$ la estima de la esperanza de la autocorrelación localizada decrece con k

Señal periódica $\Rightarrow$ autocorrelación periódica en k

Una de las aplicaciones más típicas de la autocorrelación localizada es la estima de la frecuencia fundamental de sonidos sonoros (casi-periódicos), en los que la autocorrelación presenta picos periódicos.

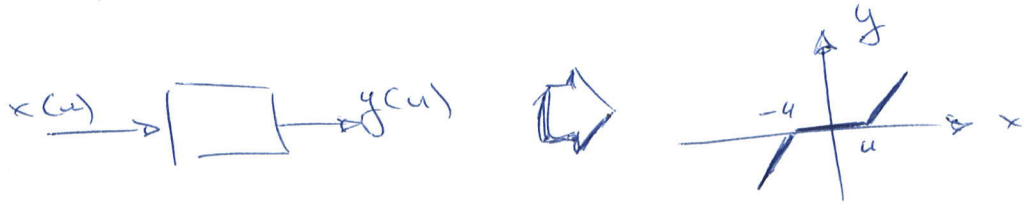
(* 2.25 *)

Los picos decaen porque:

- la señal no es totalmente periódica
- al estimar la autocorrelación hemos visto que su esperanza decrece al crecer el índice k (envolvente triangular de la autocorrelación)

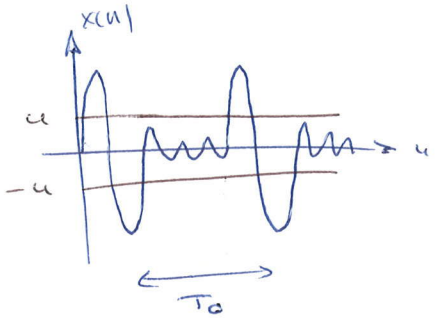
La ventana tiene que ser suficientemente grande para observar varios periodos de la señal, al menos 2.

- Central clipping (recorte central): es un dispositivo que hace una transformación no lineal sin memoria

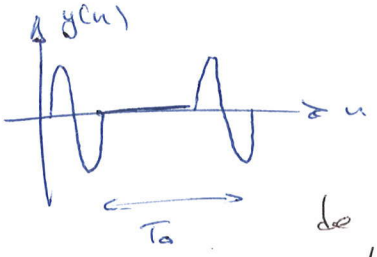


Para valores de k pequeños, hasta un umbral u, la salida es cero, y a partir de él, es proporcional a x

(* 2.24 *) Realza la frecuencia del fundamental (la periodicidad de la señal)

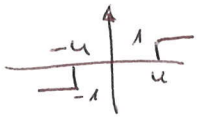


ventanas pequeñas ⇒ picos de $R_u(k)$ intermedios grandes, se pueden confundir con el fundamental ⇒ (* 2.25 *, 4.26 c)



También pesa en formantes que varían rápidamente

Recorte central de 3 niveles ⇒ simplifica el cálculo de la autocorrelación y no degrada la detección del pitch



(Rab-Schafer p. 154)

- AUTOCORRELACIÓN LOCALIZADA MODIFICADA:

$$\hat{R}_u(k) = \sum_m x_u(m)\, x_{u-k}(m-k)$$

Se mezclan 2 segmentos distintos $x_u(m)$ y $x_{u-k}(m)$

✓ no hay enventanamiento triangular de la autocorrelación ⊛

✗ no es una estima de la autocorrelación, ya que son segmentos distintos:

- ✗ pierde la simetría
- ✗ pierde las propiedades de su TF

(* 2.26 *)

⊛ ⇒ los picos no decaen

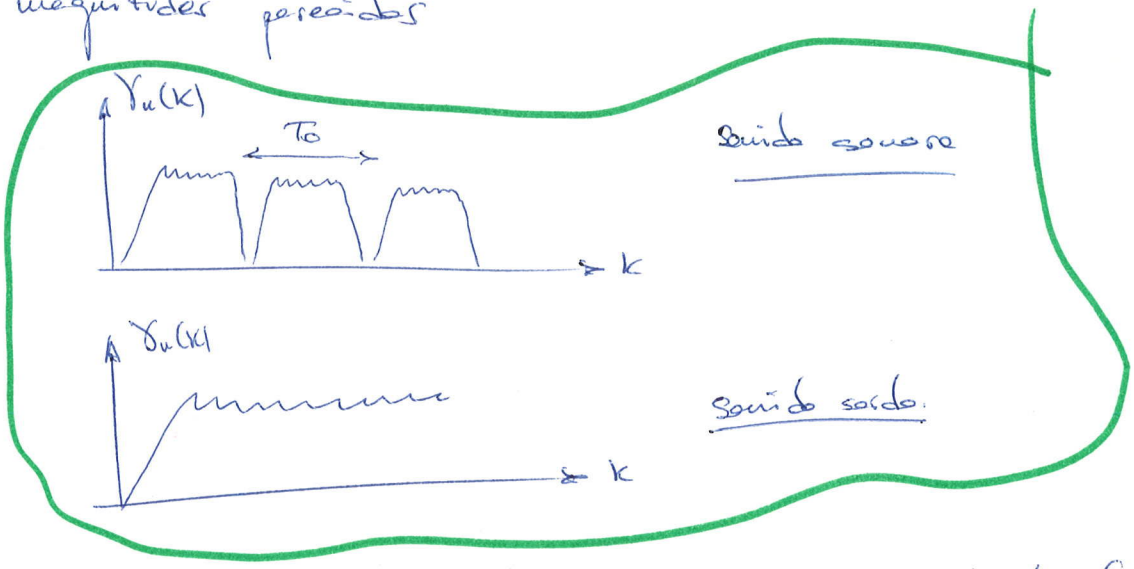
⇒ es una correlación cruzada de los dos segmentos

Autocorrelación = multiplicar = costoso (en punto fijo)

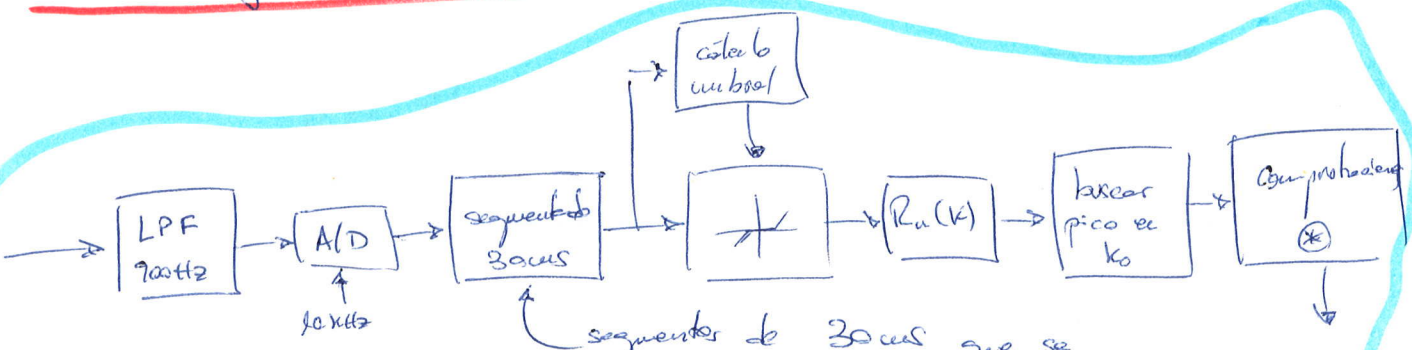
- DIFERENCIA DE MAGNITUD LOCAL:

$$\gamma_u(k) = \sum_m |x_u(m) - x_{u-k}(m-k)|$$

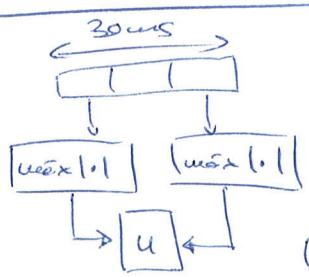
Restamos 2 segmentos separados k, por lo que si k coincide con el periodo, tenemos un mínimo ya que restamos magnitudes parecidas

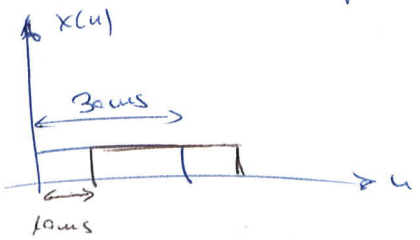


Con lo que hemos visto, hay un esquema que permite calcular la frecuencia fundamental y determinar si un sonido es sordo o sonoro:



- Cálculo del umbral:





$u = 0{,}68 \min \{ max_1, max_2 \}$

Comprobaciones:
- $R(k_0) > 0{,}3\,R(0)$
- k_0 valor adecuado (periodo fundamental posible) $\Rightarrow$ Señal sonora sólida $\equiv k_0$

Si no cumple una de ellas $\Rightarrow$ sonido sordo

2.2.- ANÁLISIS LOCALIZADO EN LA FRECUENCIA

Magnitud estable:

- TRANSFORMADA LOCALIZADA DE FOURIER (STFT = Short Time Fourier Transform):

Es la TF del segmento $x_n(m) = x(m)\,w(n-m)$

$$X_n(f) = TF\{x_n(m)\} = \sum_m x_n(m)\,e^{-j2\pi fm}$$

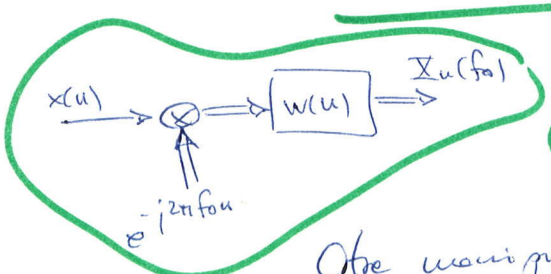
$$= \sum_m x(m)\,w(n-m)\,e^{-j2\pi fm}$$

- Propiedades:

a) Vimos antes que $TF\{R_n(k)\} = |X_n(f)|^2$

b) Si fijamos una cierta frecuencia $f = f_0$,

$$X_n(f_0) = \sum_m \left(x(m)\,e^{-j2\pi f_0 m}\right) w(n-m) = \left(x(n)\,e^{-j2\pi f_0 n}\right) * w(n)$$

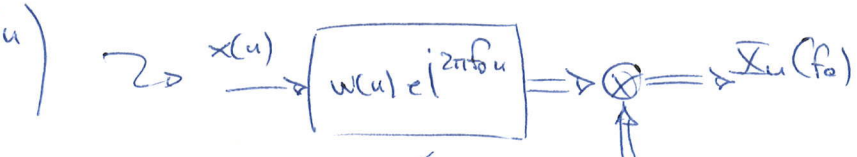


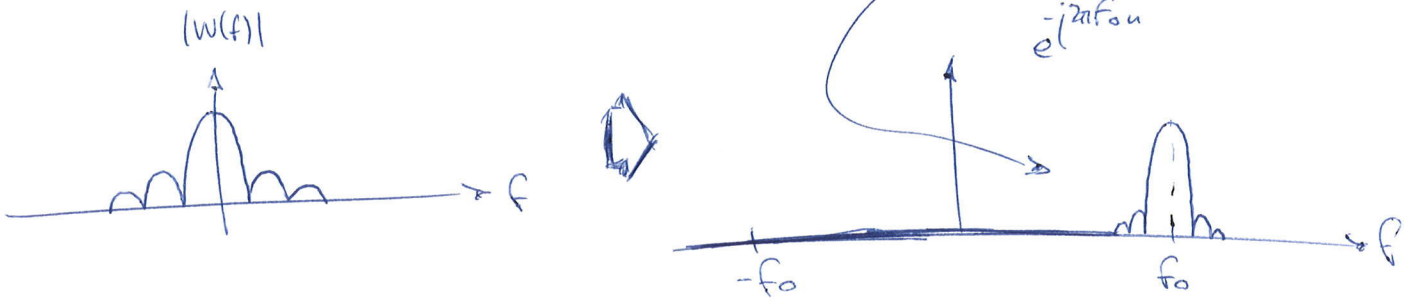
(transformada + filtro b)

Otra manipulación posible es:

$$X_n(f_0) = e^{-j2\pi f_0 n} \sum_m x(m)\,w(n-m)\,e^{j2\pi f_0 (n-m)} =$$

$= x(u) * \left(w(u) e^{j 2\pi f_0 u} \right)$ ⇝





f hace pasa bajo las frecuencias en torno a f_0 y filtra

$\mathcal{F}$ al resto

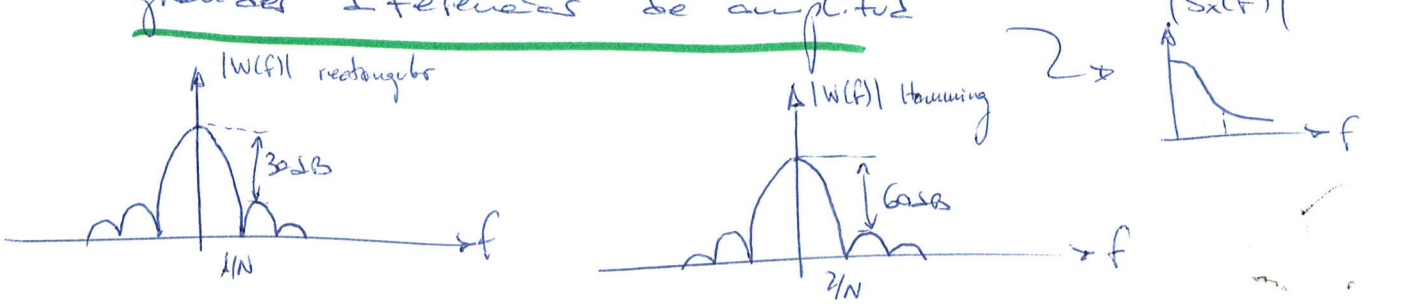
$x_u(u)$ podemos obtenerlo mediante transformada inversa:

$$x_u(u) = TF^{-1}\{X_u(f)\} = \int_{-1/2}^{1/2} X_u(f) e^{j 2\pi f u} \, df$$

$$x_u(m) = x(m) w(u-m) \longrightarrow x(u) w(0) = x_u(u)$$

$$x(u) = \frac{1}{w(0)} x_u(u) = \frac{1}{w(0)} \int_{-1/2}^{1/2} X_u(f) e^{j 2\pi f u} \, df$$

En el <u>análisis espectral</u> se suele preferir la <u>ventana de Hamming</u> para hacer el análisis localizado de Fourier, ya que el espectro de la señal de voz hay componentes en diferentes bandas con <u>grandes diferencias de amplitud</u>



Hamming tiene mayor resolución, pero mejor atenuación de los lóbulos secundarios.

El tamaño de la ventana será suficientemente grande para ver las características periódicas, como la detección del pitch (frec. fundamental)

- detección del pitch: N grande

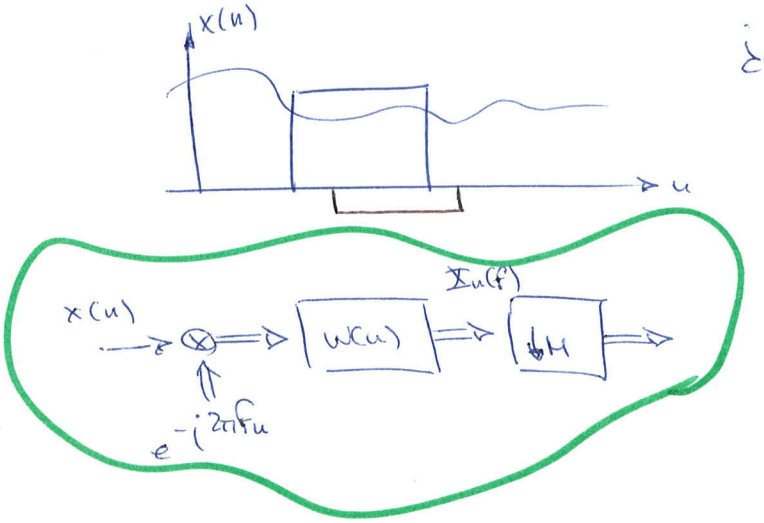
  frecuencia de muestreo: $f_m = 10\,kHz \Rightarrow N = 220$

  $\Rightarrow 22\,ms$

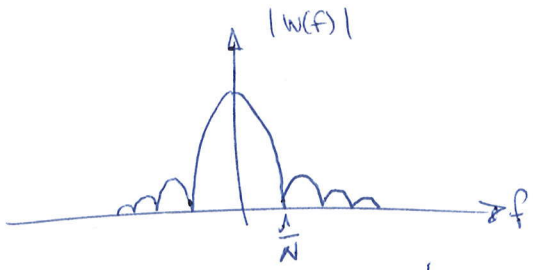
  $B_W = 45\,Hz$

  - buena resolución en frecuencia
  - pero resolución temporal

  (* 2.28 – 2.31 *)



¿rapidez de deslizamiento de la ventana?



paso bajo ⇒ admite diezmado

→ desplazar más de 1 muestra

Como el FPB no es ideal, habrá aliasing, que consideraremos despreciable si el ancho de banda es suficiente.

- rectangular: $B = 1/N \rightarrow f_m' = \frac{2}{N} f_m \Rightarrow \frac{N}{2}\downarrow \quad (M = \frac{N}{2})$
- Hamming: $B = 2/N \rightarrow f_m' = \frac{4}{N} f_m \Rightarrow \frac{N}{4}\downarrow \quad (M = \frac{N}{4})$

Cómputo: ∄ tiempo/frecuencia continuos ⇒ numéricos

$x_u(m) \xrightarrow{TF} X_u(f)$

(duración N muestras ⇒ admite DFT)

$(DFT_N \equiv st\text{-}DFT)$

$x_u(m) \xleftarrow{DFT_N} X_u(k) = X_u(f)\Big|_{f=\frac{k}{N}} \qquad k \in \mathbb{Z}$

$$X_u(k) = \sum_m x_u(m)\, e^{-i\frac{2\pi}{N}km} \qquad k = 0:N-1$$

st-DFT

- **BANCO DE FILTROS:**

Es una interpretación de la st-DFT

- iDFT: $x_u(m) \xleftarrow{DFT} X_u(k)$

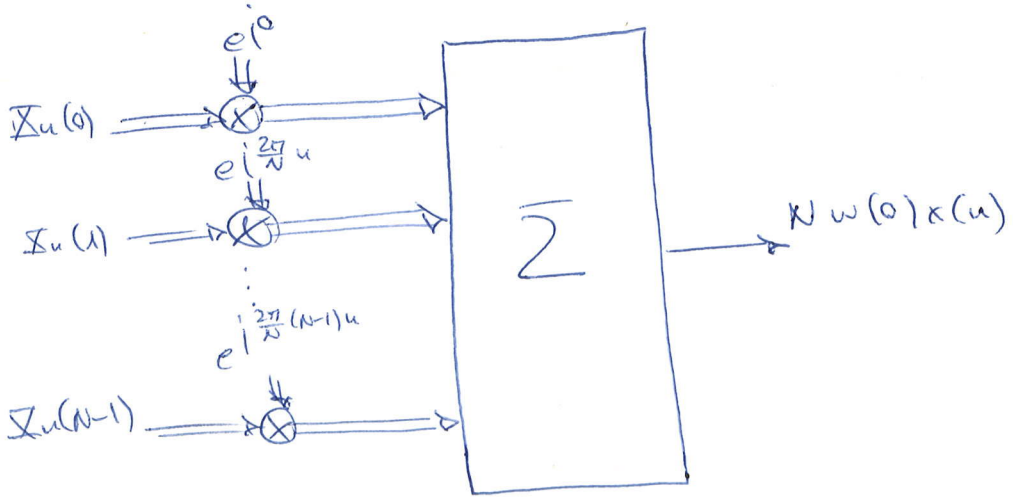
$$x_u(m) = DFT^{-1}\{X_u(k)\} = \frac{1}{N}\sum_{k=0}^{N-1} X_u(k)\, e^{i\frac{2\pi}{N}km}$$

$$x_u(m) = x(m)\, w(u-m)$$

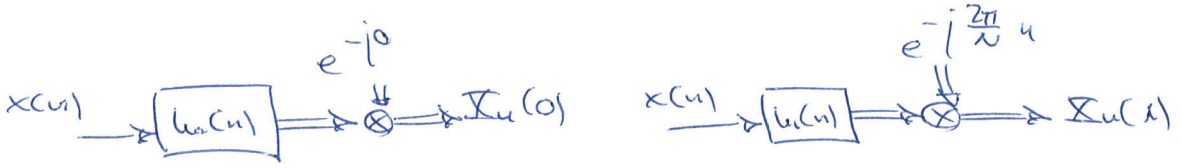
$$\hookrightarrow x_u(u) = x(u)\, w(0) = \frac{1}{N}\sum_{k=0}^{N-1} X_u(k)\, e^{i\frac{2\pi}{N}ku}$$

$x(u)\cdot N\cdot w(0)$

Se hace una modulación de $X_u(k)$:



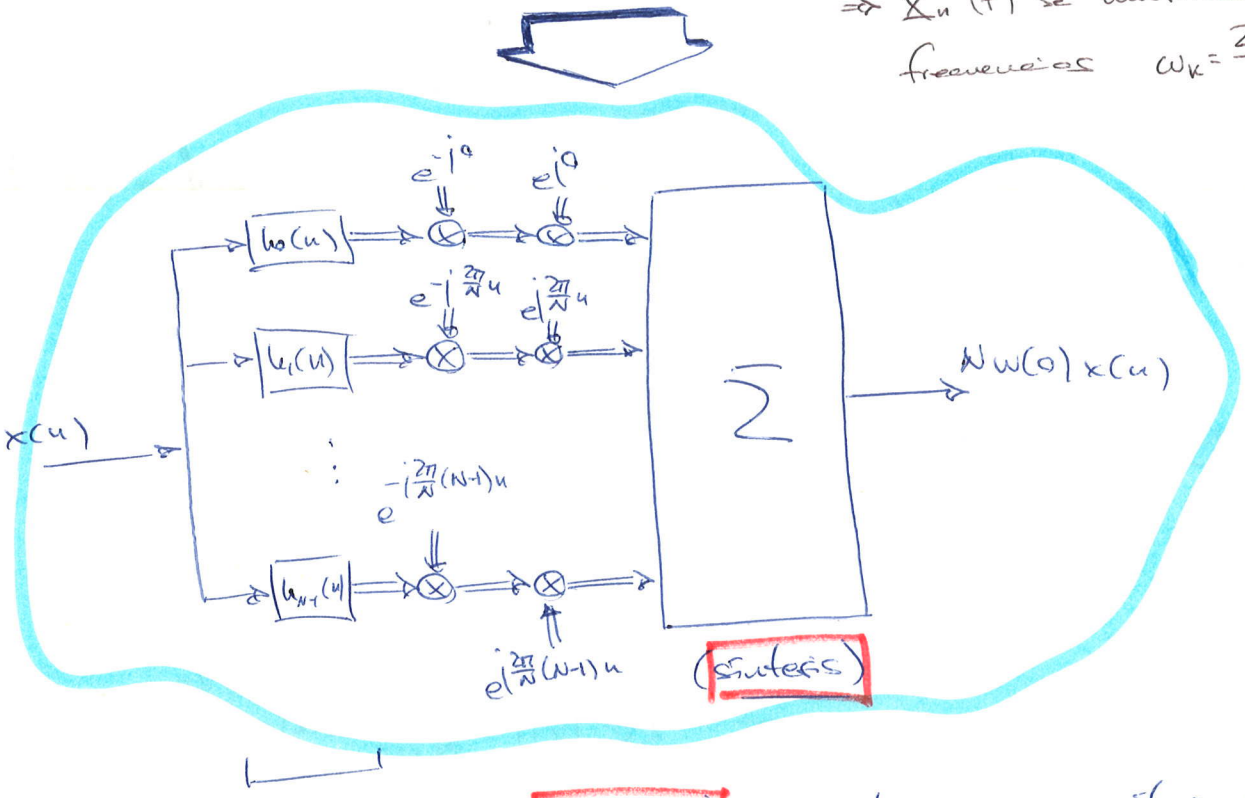
Transformada localizada a una frecuencia:



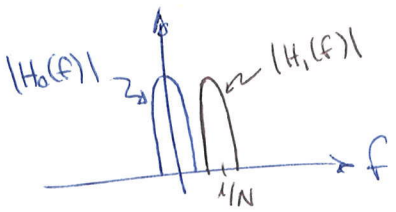
$$h_k(n) = w(n)\, e^{j\frac{2\pi}{N}kn}$$

Ventana longitud N

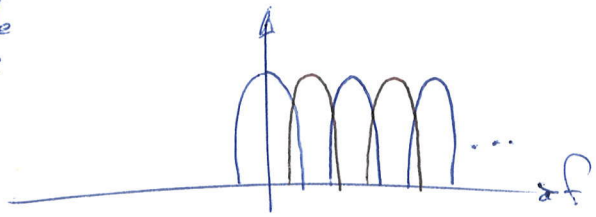
$\Rightarrow X_n(f)$ se muestrea en las frecuencias $\omega_k = \frac{2\pi k}{N}$ $k = 0 : N-$



banco de filtros (análisis) $\Rightarrow$ hace un análisis de la señal



la señal se divide en bandas



• aplicación: codificación subbanda, para quitar enmascaramientos en voz y audio

Ejemplo: Estimador de energía local

$$e(n) = \sum_{m=n-N+1}^{n} x^2(m)$$

a) Buscar una expresión recursiva para este estimador

$$e(n) = e(n-1) + a(n)$$

$$a(n) = e(n) - e(n-1) = \sum_{m=n-N+1}^{n} x^2(m) - \sum_{m=n-N}^{n-1} x^2(m)$$

$$= x^2(n) - x^2(n-N)$$

$$\Rightarrow \boxed{e(n) = e(n-1) + x^2(n) - x^2(n-N)}$$

b) Lo mismo pero usando una ventana de Hamming

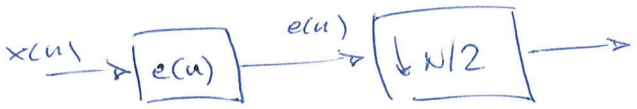
$$e(n) = \sum_{m=n-N+1}^{n} x^2(m)\, w^2(n-m)$$

$$a(n) = e(n) - e(n-1) = \sum_{m=n-N+1}^{n} x^2(m) w^2(n-m) - \sum_{m=n-N}^{n-1} x^2(m)\, w^2(n-1-m) =$$

$$= x^2(n) w^2(0) - x^2(n-1-N) w^2(N-1) + \sum_{m=n-N+1}^{n-1} x^2(m) \left(w^2(n-m) - w^2(n-1-m) \right)$$

↳ suma $N+1$ términos ⇒ es más complejo usar una recursión con una ventana no rectangular que usar la expresión directa

c) Estime del coste computacional en el caso inicial para la expresión directa [1] y la recursiva [2] si la estima de energía se calcula cada $N/2$ muestras de la entrada (diezmado por $N/2$). Calcula teniendo en cuenta sólo los productos

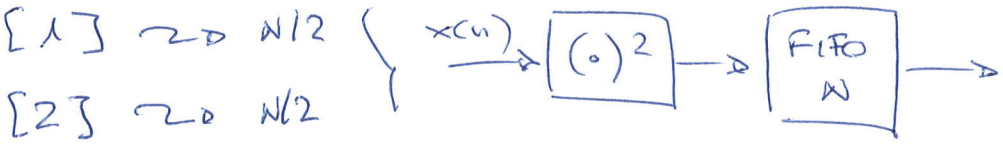


memorizando $x(n)$:

[1] → en cada cálculo requiere N productos

[2] → 2 productos por iteración, $N/2$ veces para la salida cada $N/2$ muestras ⇒ N productos

Otra opción es memorizar $x^2(n)$



d) Tenemos un estimador de la autocorrelación local definido como:

$$r(k;n) = \frac{1}{N-|k|} \sum_{m=n-N+1+|k|}^{n} x(m)\,x(m-|k|) \qquad \text{(insesgado)}$$

$$r(0;n) = \frac{1}{N} \sum_{m=n-N+1}^{n} x^2(m) = \frac{e(n)}{N}$$

e) Calcular el sesgo y la varianza de $e(n)$:

$$\begin{cases} x(n) \text{ es AWGN} \\ E[x^2(m)\,x^2(n)] = \sigma_x^4\left(\delta(m-n)+1\right) \\ \sigma_x^2 = \mathrm{Var}(x(n)) = P_x \end{cases}$$

$$E[e(n)] = \sum_{m=n-N+1}^{n} E[x^2(m)] = N\sigma_x^2 \qquad \text{es insesgado}$$

$$\mathrm{Var}(e(n)) = E[e^2(n)] - E^2[e(n)] = E[e^2(n)] - N^2\sigma_x^4$$

$$E[e^2(n)] = E\left[\sum_{m=n-N+1}^{n}\sum_{\ell=n-N+1}^{n} x^2(m)\,x^2(\ell)\right] =$$

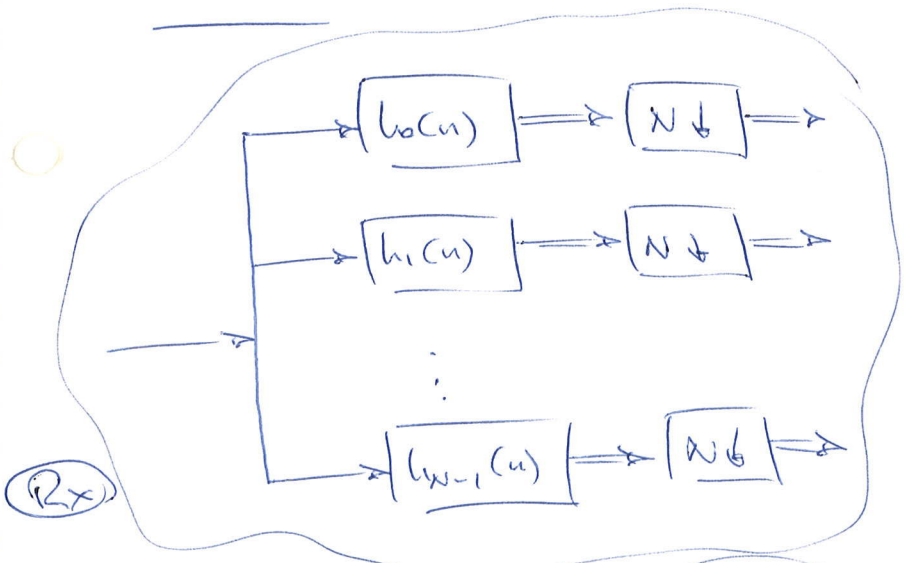
$$= \sum_{m}\sum_{\ell} E\left[x^2(m)\,x^2(\ell)\right] =$$

$$= \sum_{m}\sum_{\ell} \sigma_x^4\left(\delta(m-\ell)+1\right) =$$

$$= \sum_{m}\sigma_x^4 + \sum_{m}\sum_{\ell}\sigma_x^4 = \left(N+N^2\right)\sigma_x^4$$

$$\Rightarrow \boxed{Var\left(e(n)\right) = N\,\sigma_x^4}$$

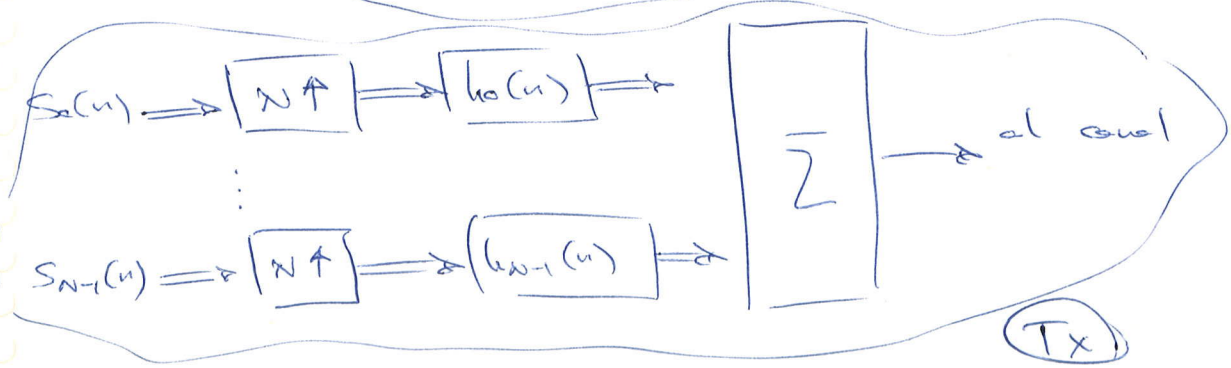
Curiosidad:



modelo OFDM o multiportadora

señal de entrada: modelos PAM

ej: ADSL, radiodifusión digital terrestre



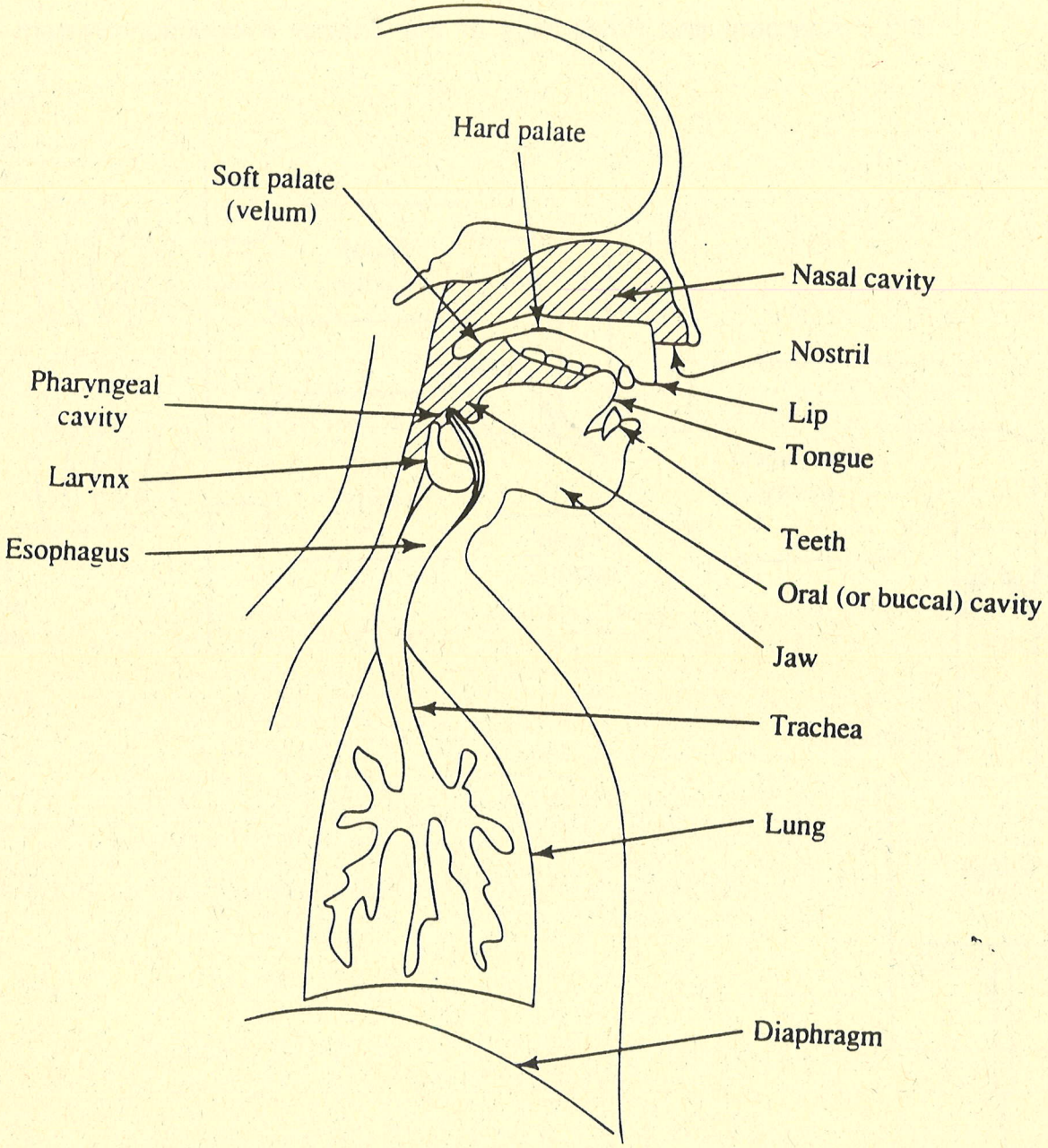


FIGURE 2.1. A schematic diagram of the human speech production mechanism.

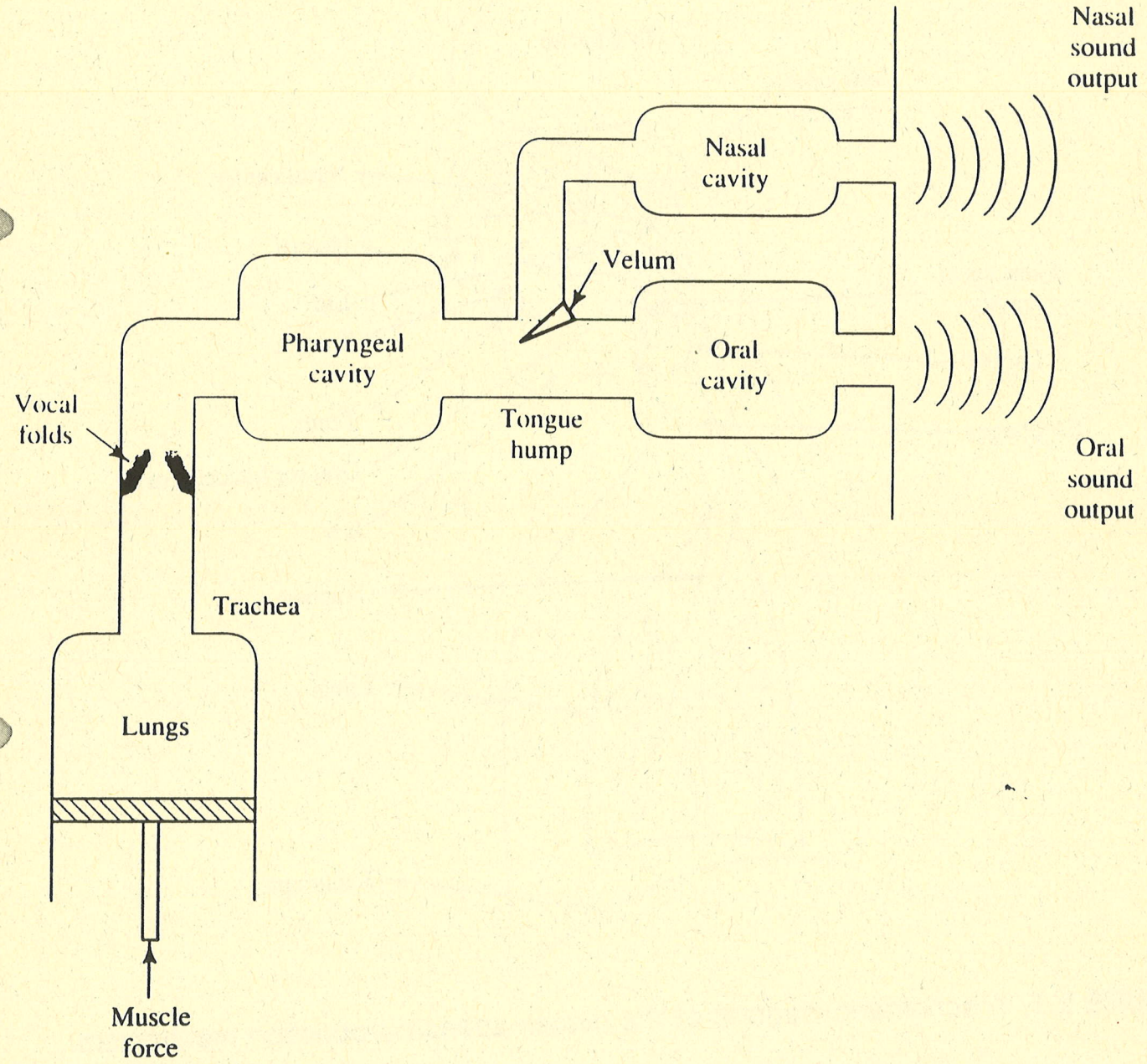


FIGURE 2.2. A block diagram of human speech production.

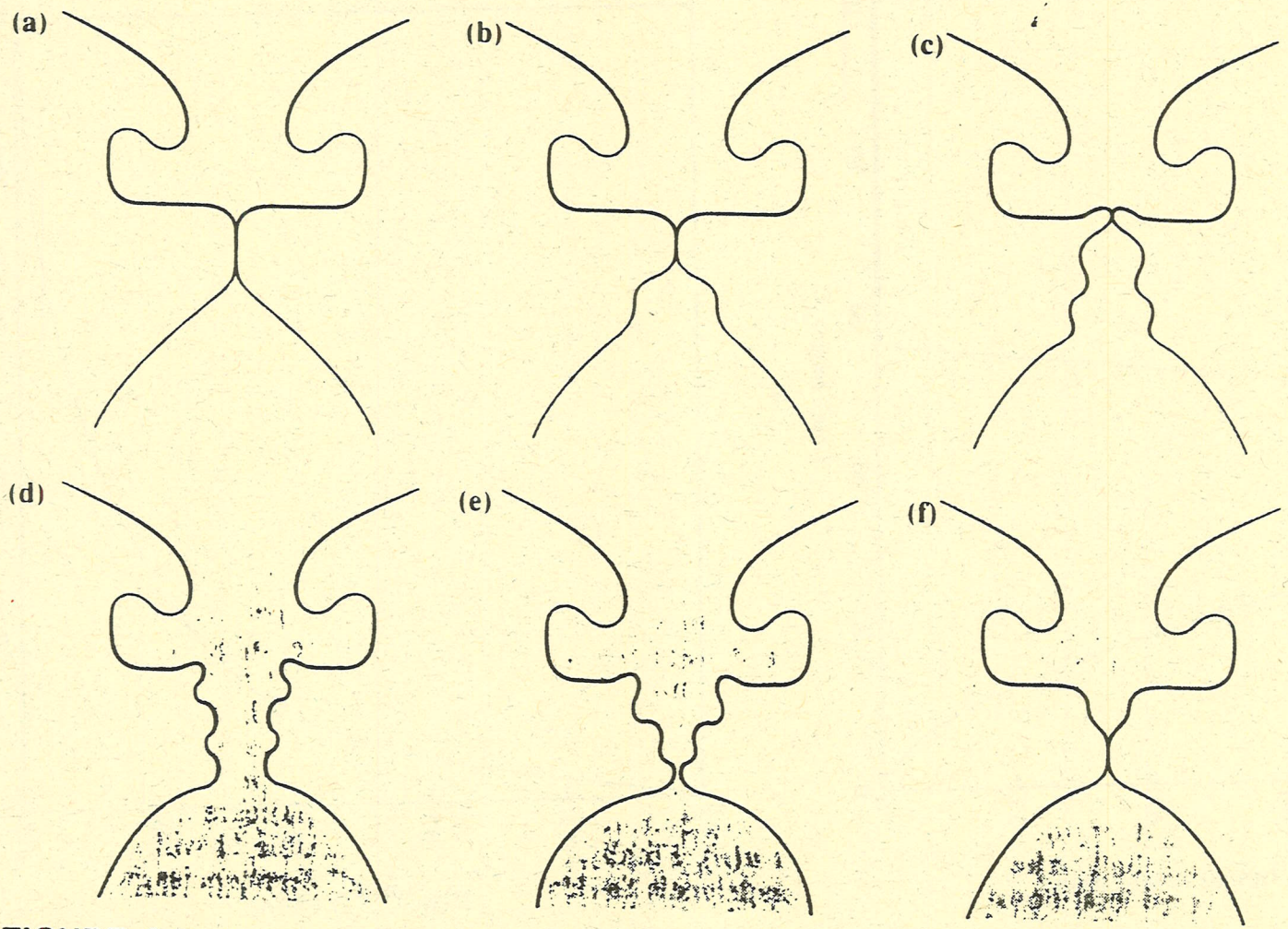
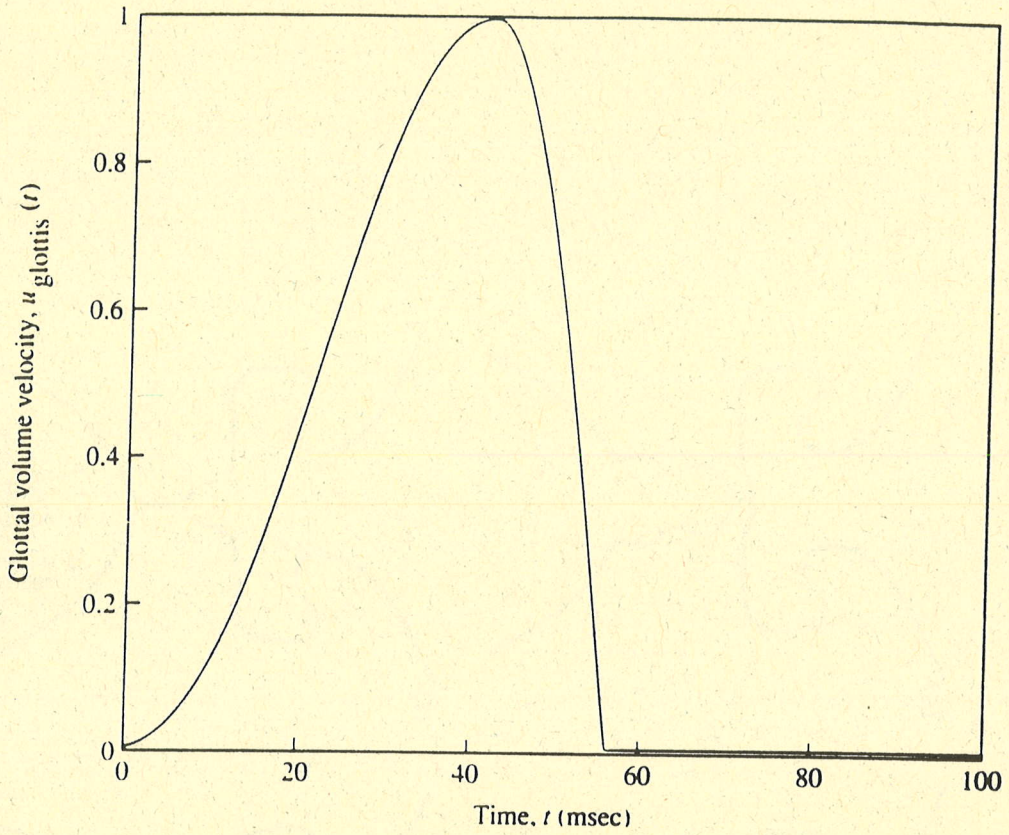
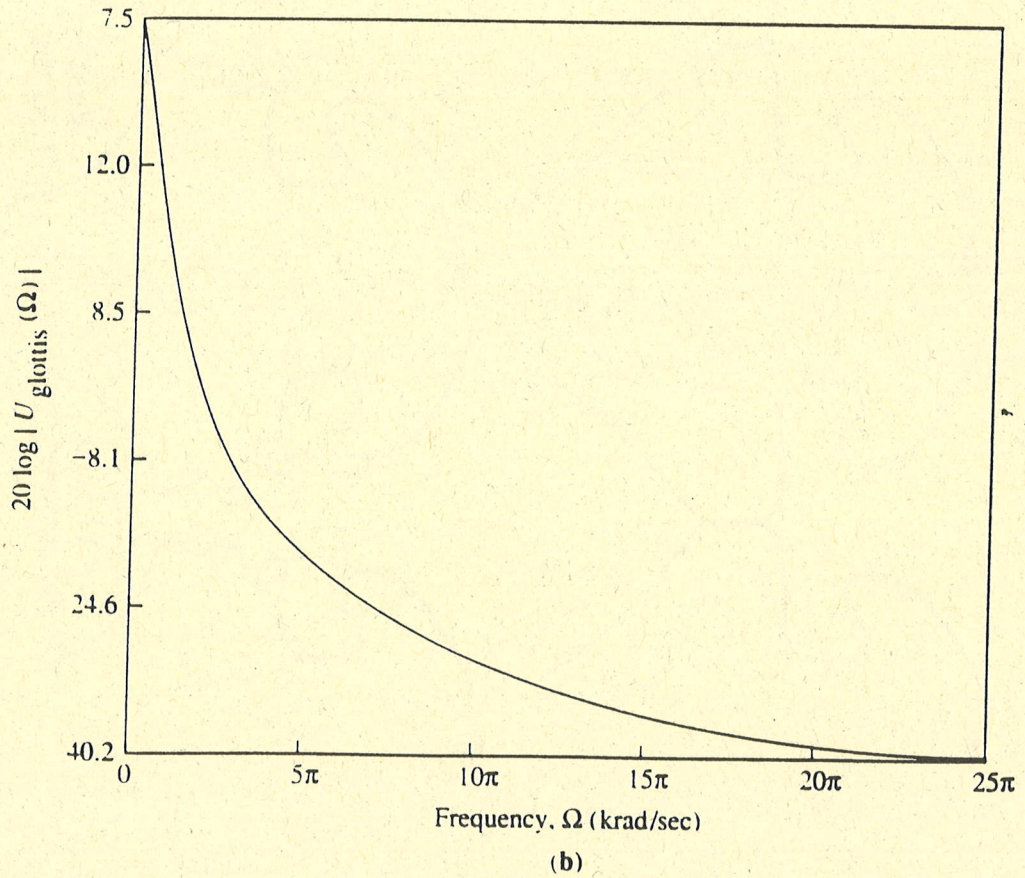


FIGURE 2.7. A sequence of cross sections of the larynx illustrating a complete phonation cycle. After Vennard (1967).

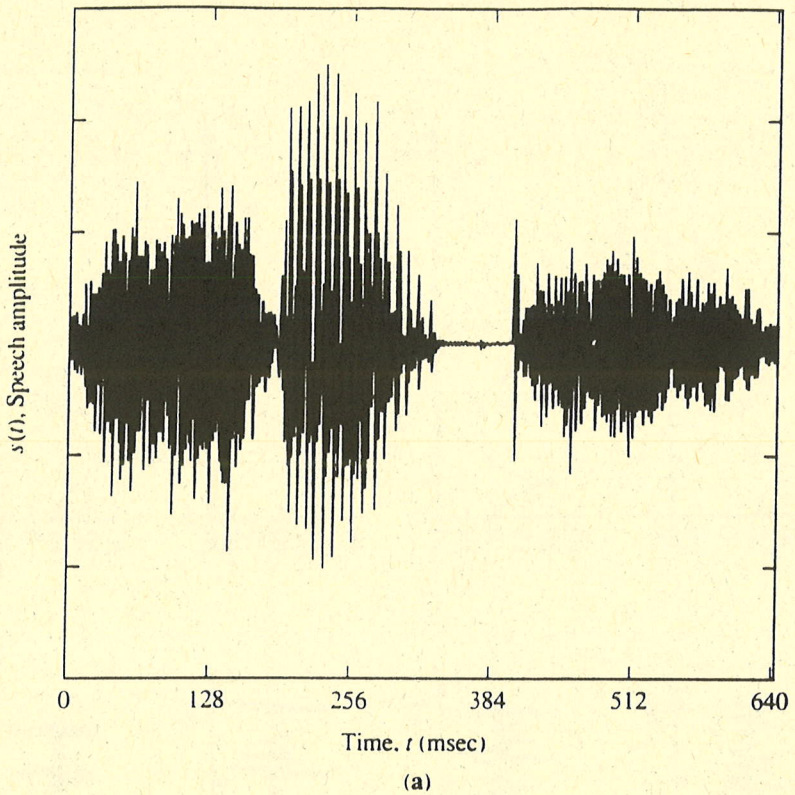


(a)



(b)

FIGURE 2.8. (a) Time waveform of volume velocity of the glottal source excitation. (b) Magnitude spectrum of one pulse of the volume velocity at the glottis.



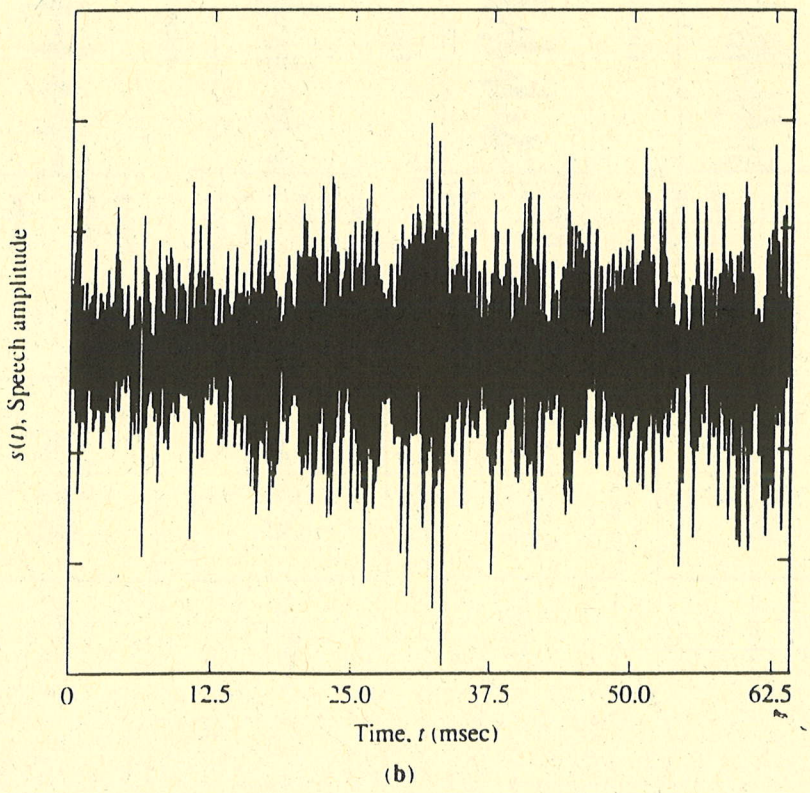
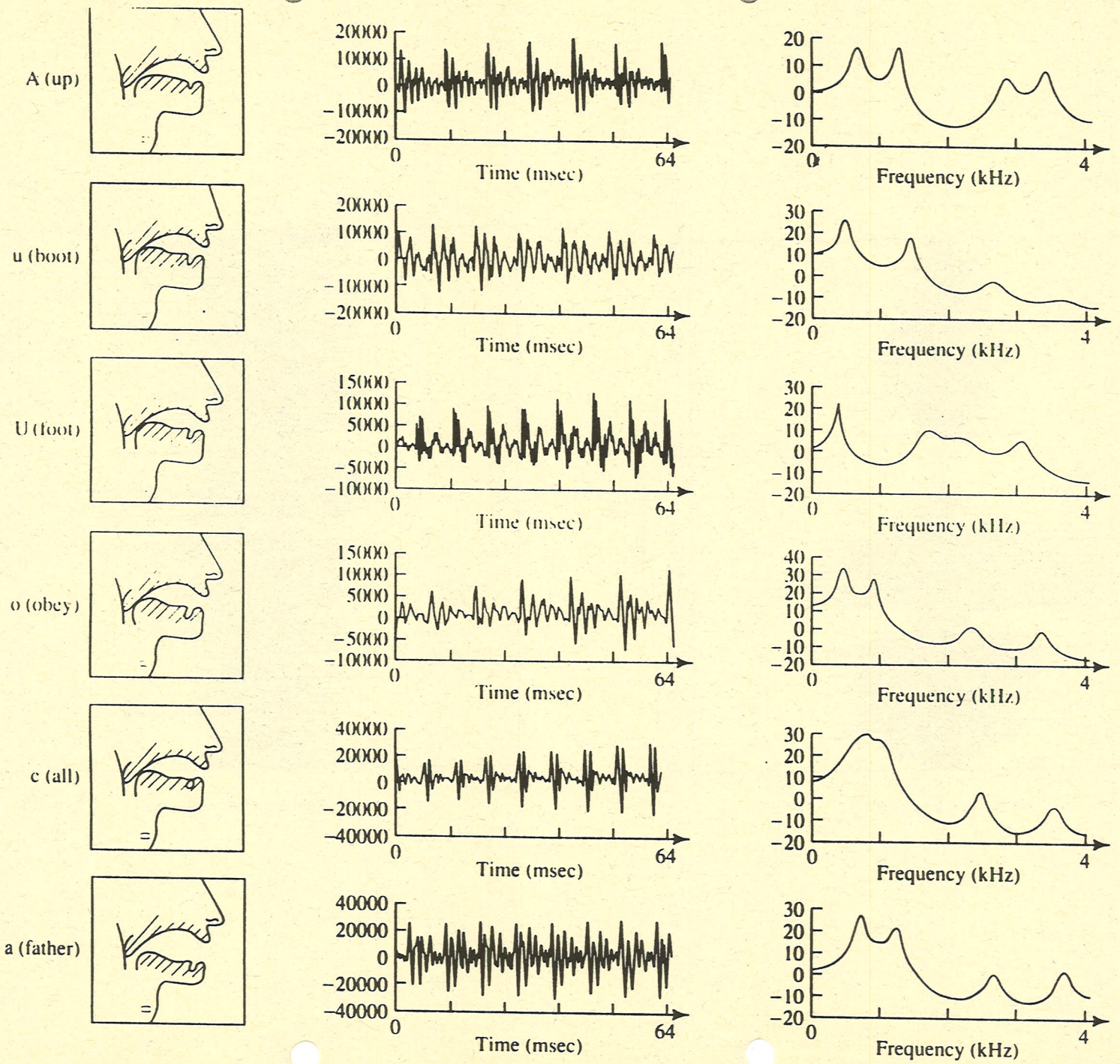


FIGURE 2.4. (a) A speech signal of the word "six" spoken by a male speaker; (b) blowup of a frame of the steady-state region of the initial /s/ sound; (c) blowup of the vowel /I/.


A (up)
u (boot)
U (foot)
o (obey)
c (all)
a (father)
Time (msec)
Frequency (kHz)

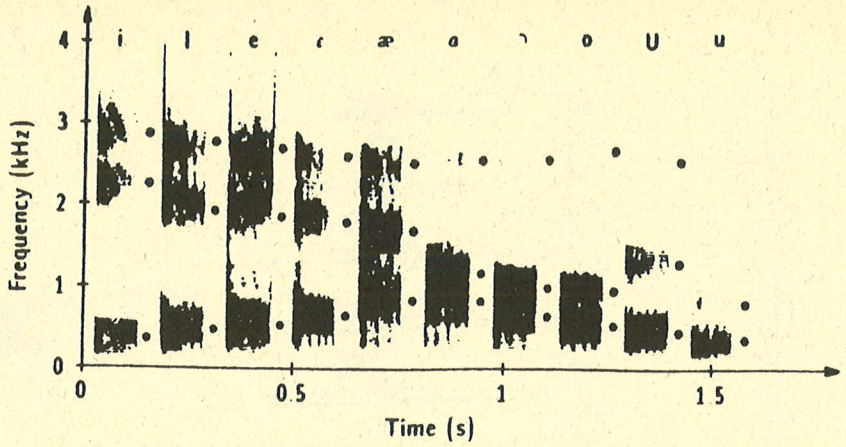


Fig. 3.10 Spectrogram of short sections of English vowels from a male speaker. Formants for each vowel are noted by dots.

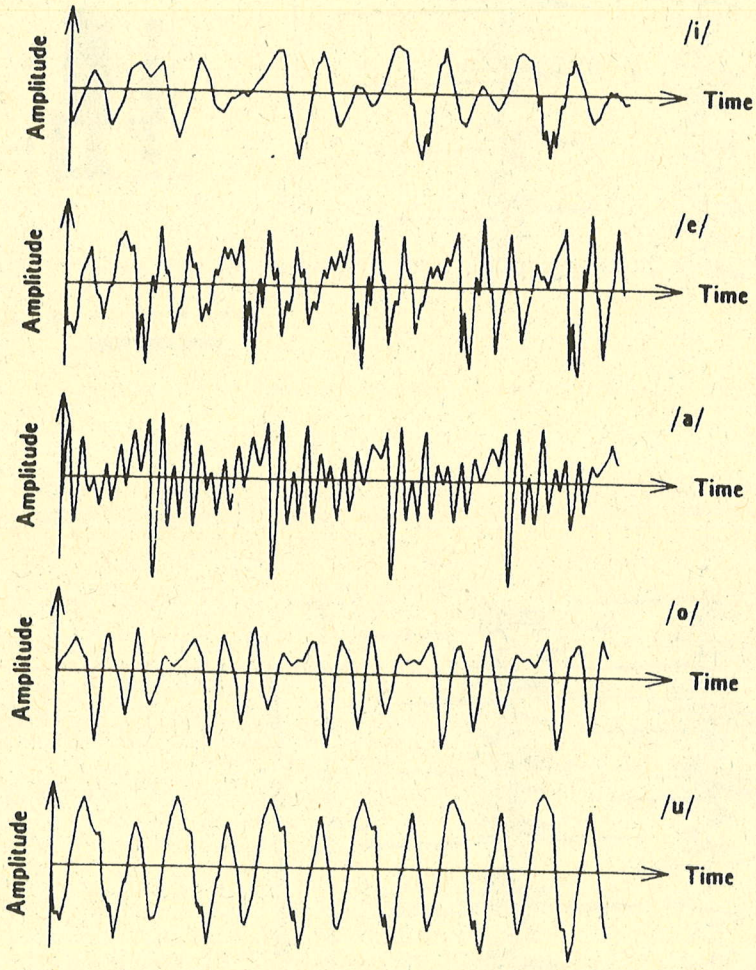


Fig. 3.12 Typical acoustic waveforms for five English vowels. Each plot shows 40 ms of a different vowel, which comprises about 5–6 pitch periods for this speaker. Note the quasi-periodic nature of such voiced speech as well as the varying spectral content for different vowels.

		/i/	/ɪ/	/ɛ/	/æ/	/ɑ/	/ɔ/	/ʊ/	/u/	/ʌ/	/ɝ/
F1	male	270	390	530	660	730	570	440	300	640	490
	female	310	430	610	860	850	590	470	370	760	500
F2	male	2290	1990	1840	1720	1090	840	1020	870	1190	1350
	female	2790	2480	2330	2050	1220	920	1160	950	1400	1640
F3	male	3010	2550	2480	2410	2440	2410	2240	2240	2390	1690
	female	3310	3070	2990	2850	2810	2710	2680	2670	2780	1960

Table 3.2 Average formant frequencies (in Hz) for English vowels by adult male and female speakers. (After Peterson and Barney [17].)

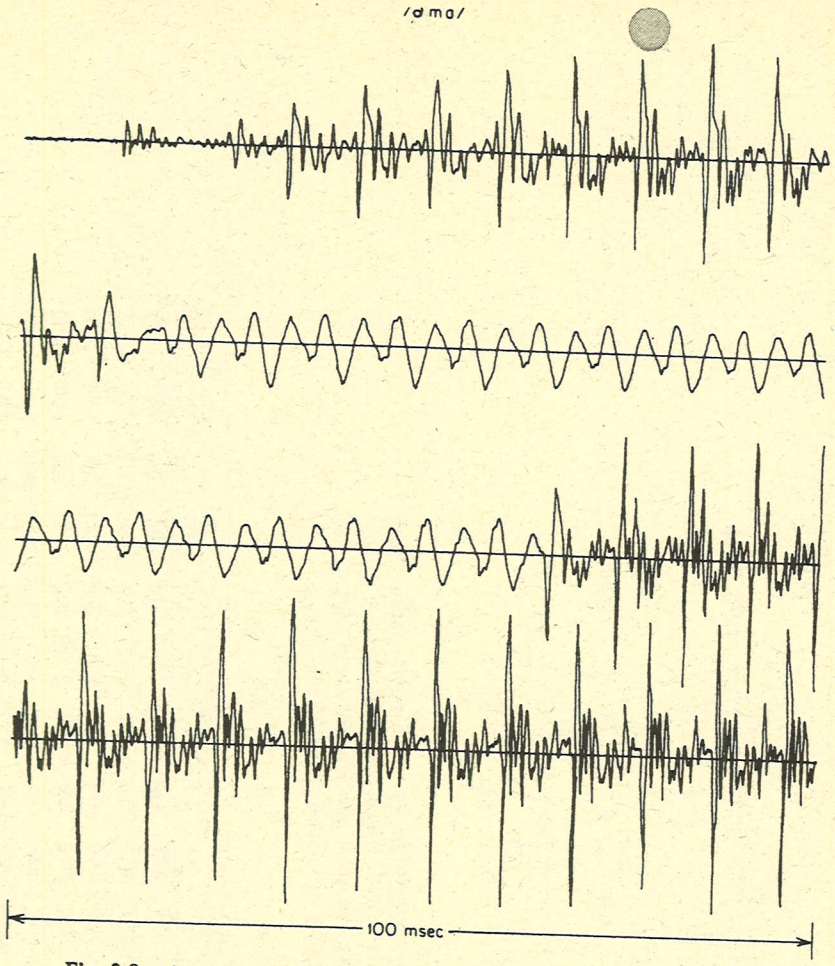


Fig. 3.8 Acoustic waveforms and spectrograms for utterances */UH–M–A/* and */UH–N–A/*.

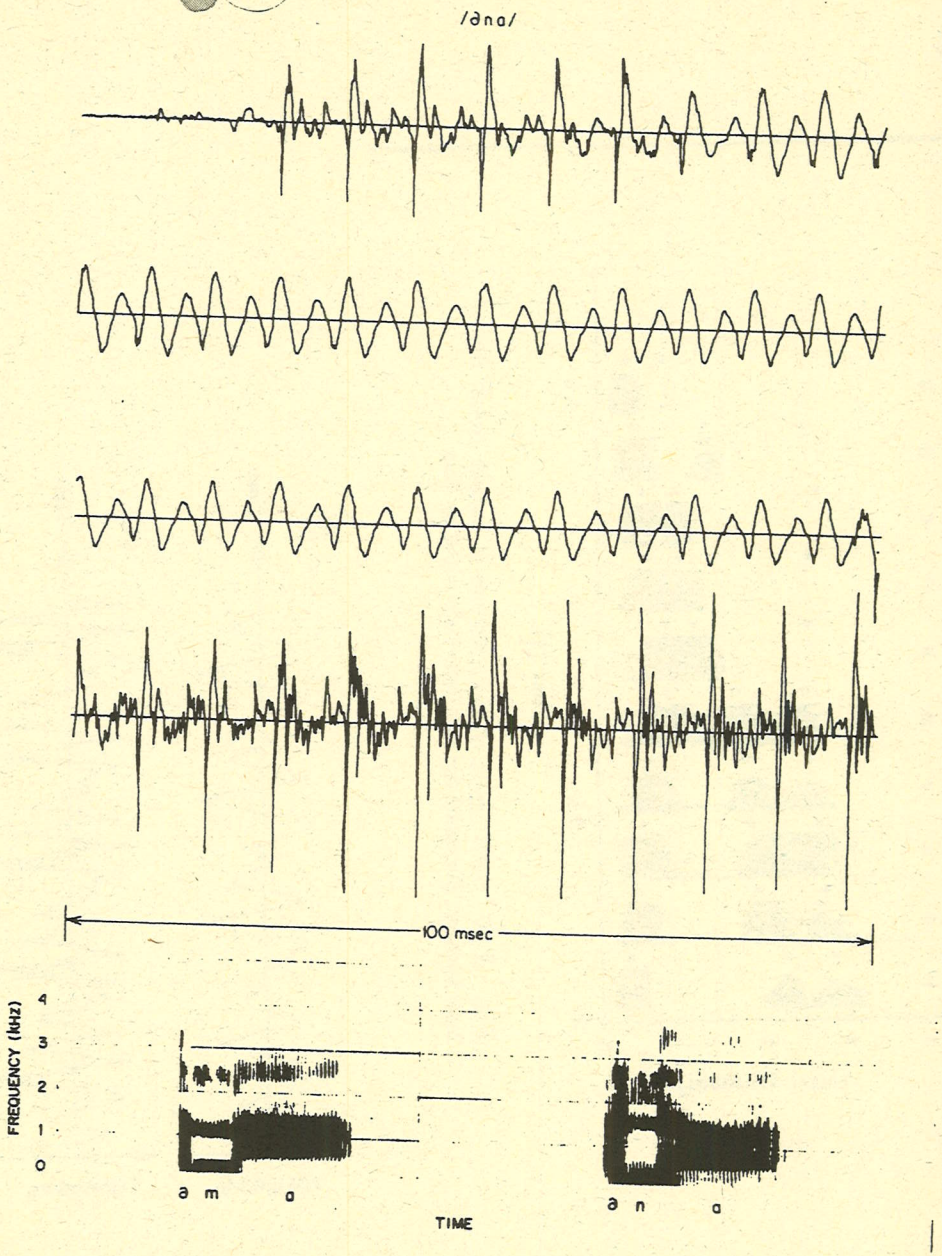


Fig. 3.8 (Continued)

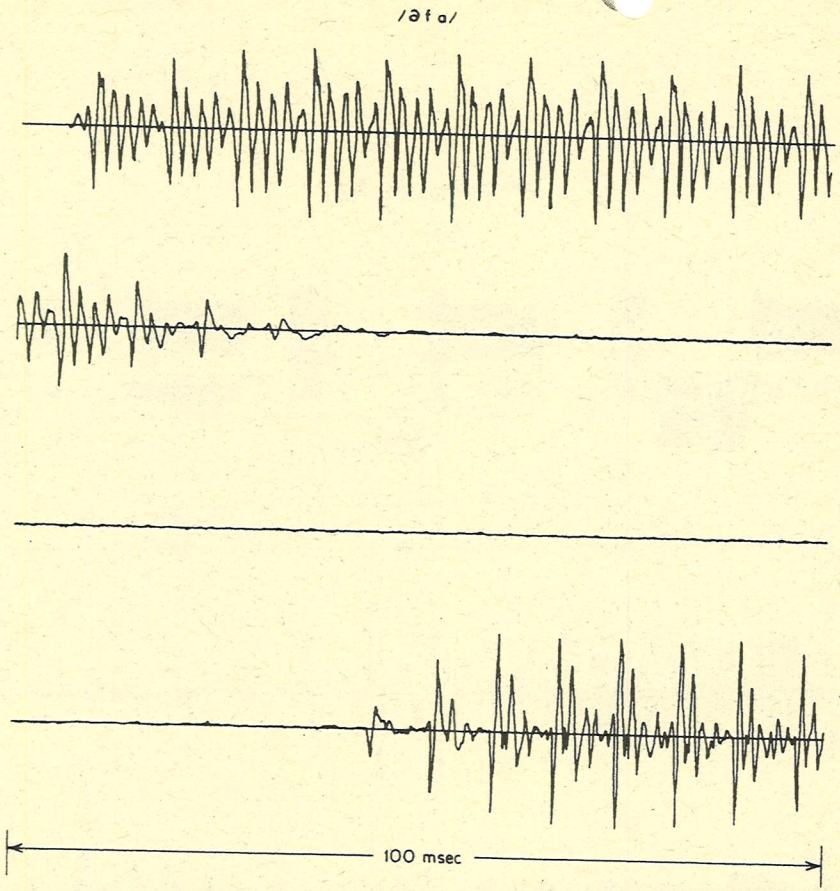


Fig. 3.9 Acoustic waveforms and spectrograms for */UH–F–A/*, */UH–S–A/*, and */UH–SH–A/*.

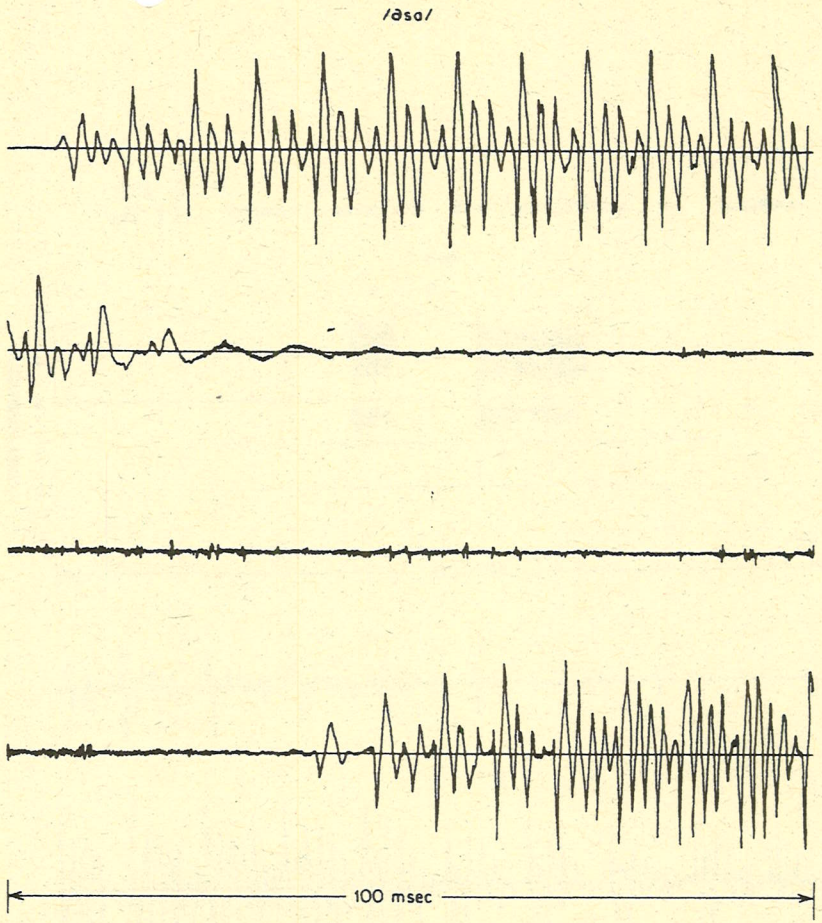


Fig. 3.9 (Continued)

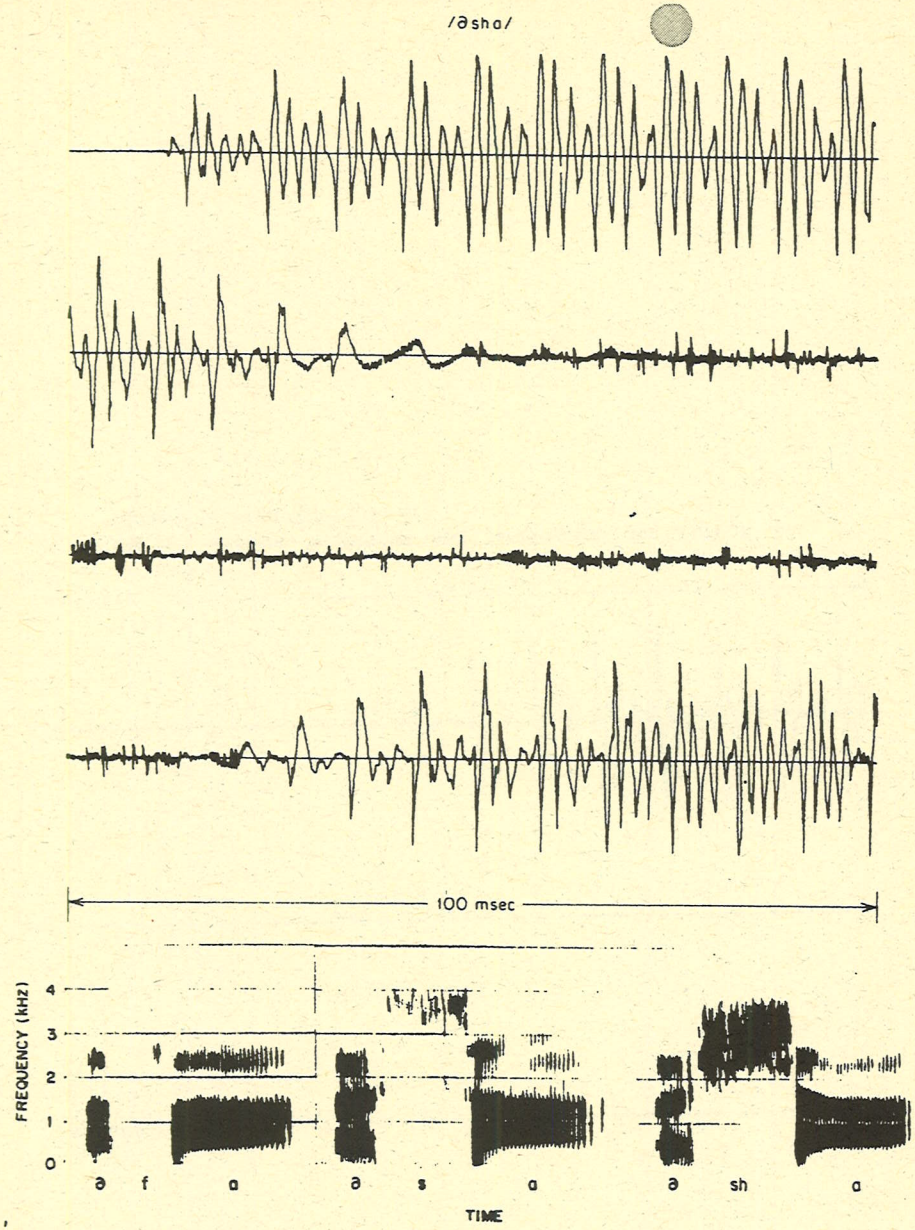


Fig. 3.9 (Continued)

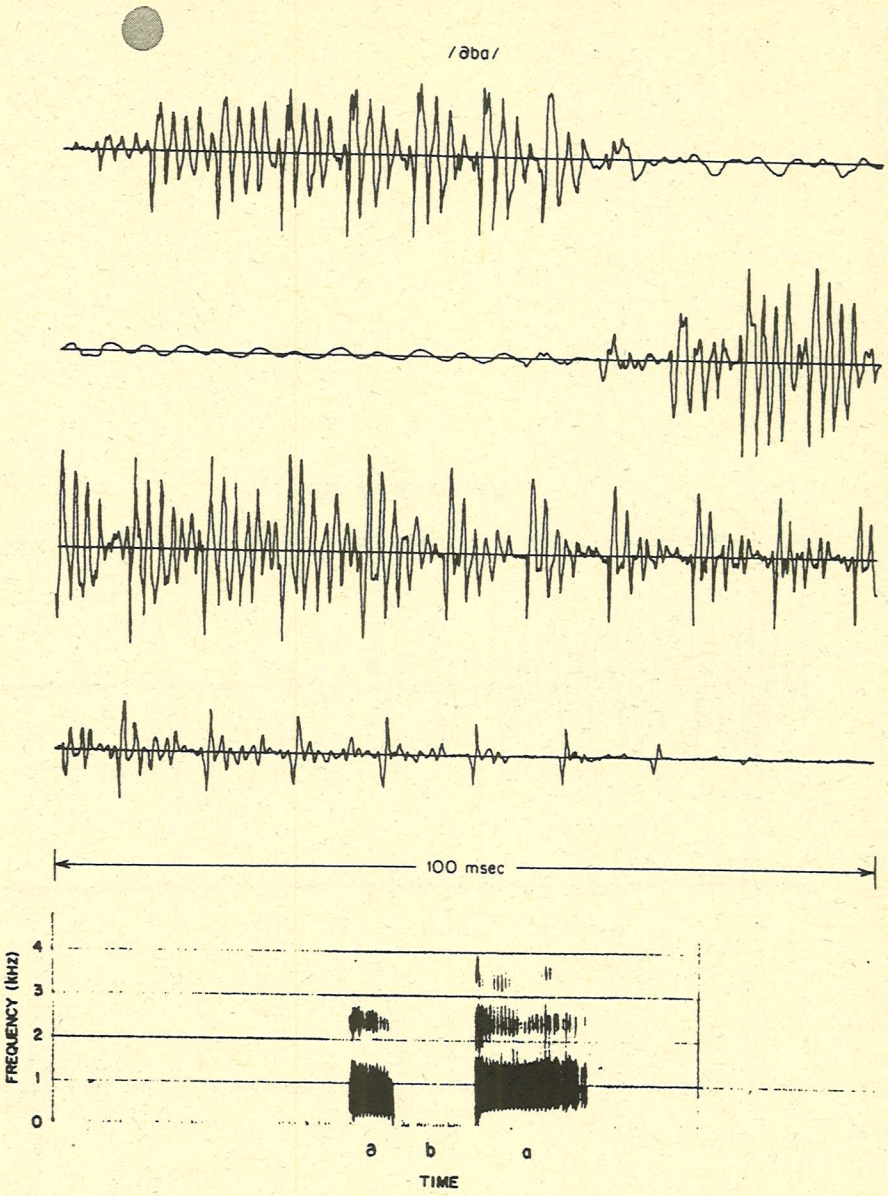


Fig. 3.11 Acoustic waveform and spectrogram for utterance */UH–B–A/*.

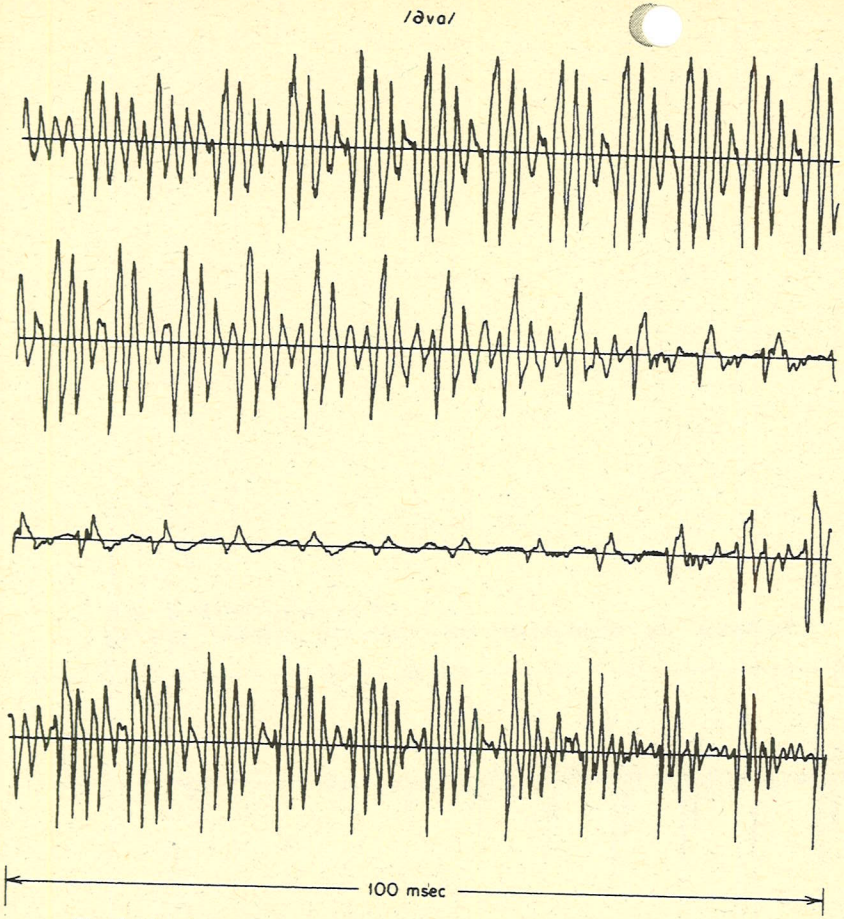


Fig. 3.10 Acoustic waveforms and spectrograms for utterances */UH−V−A/* and */UH−ZH−A/*.

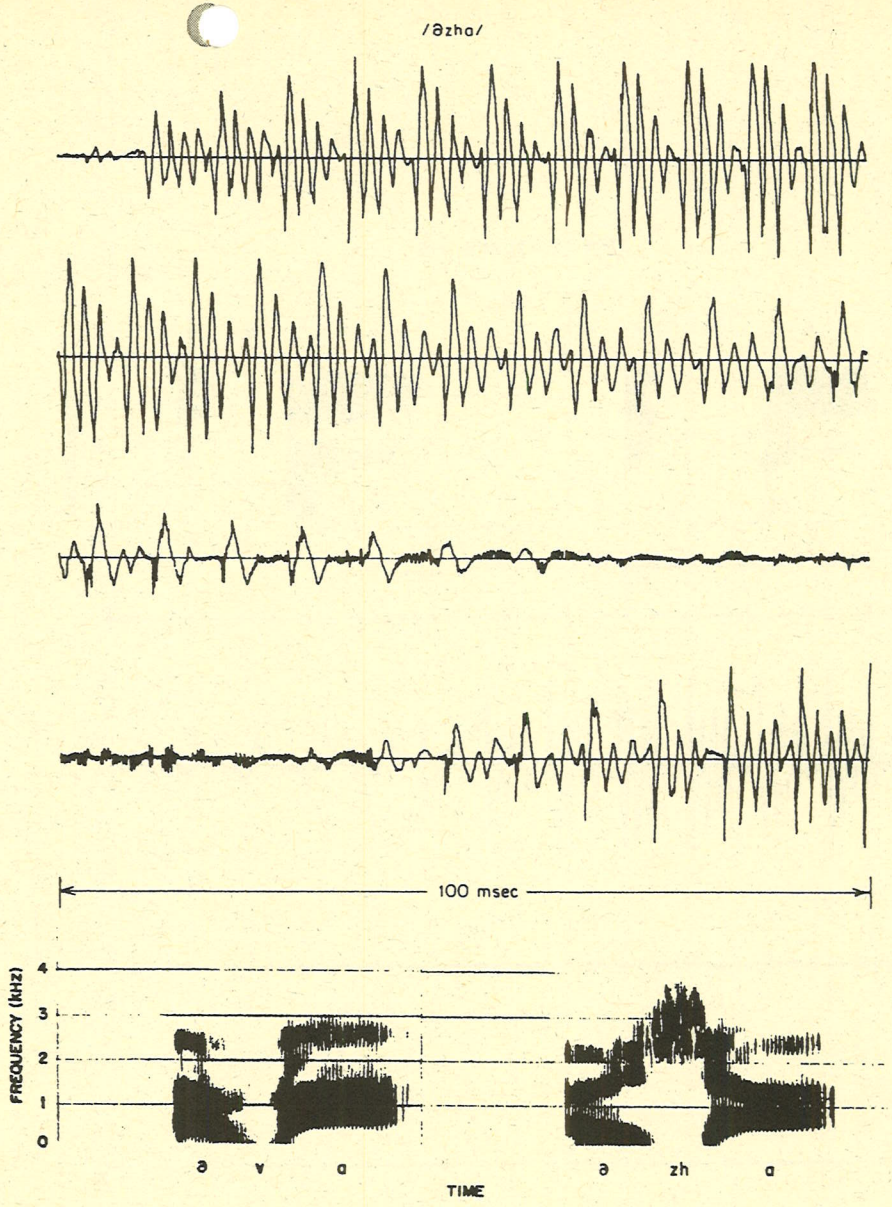


Fig. 3.10 (Continued)

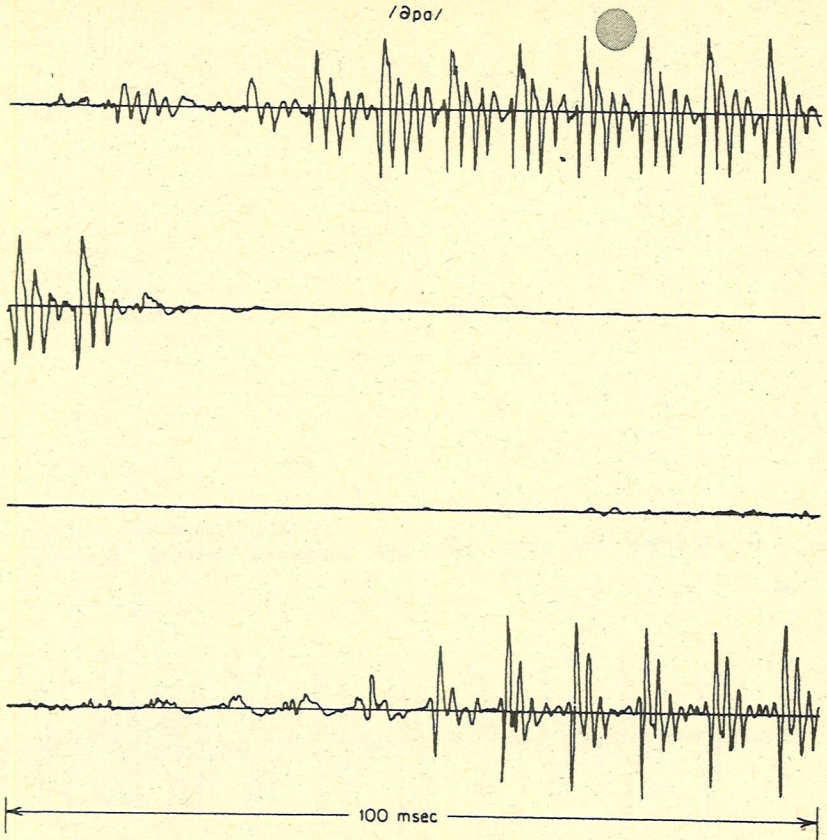


Fig. 3.12 Acoustic waveforms and spectrograms for utterances */UH–P–A/* and */UH–T–A/*.

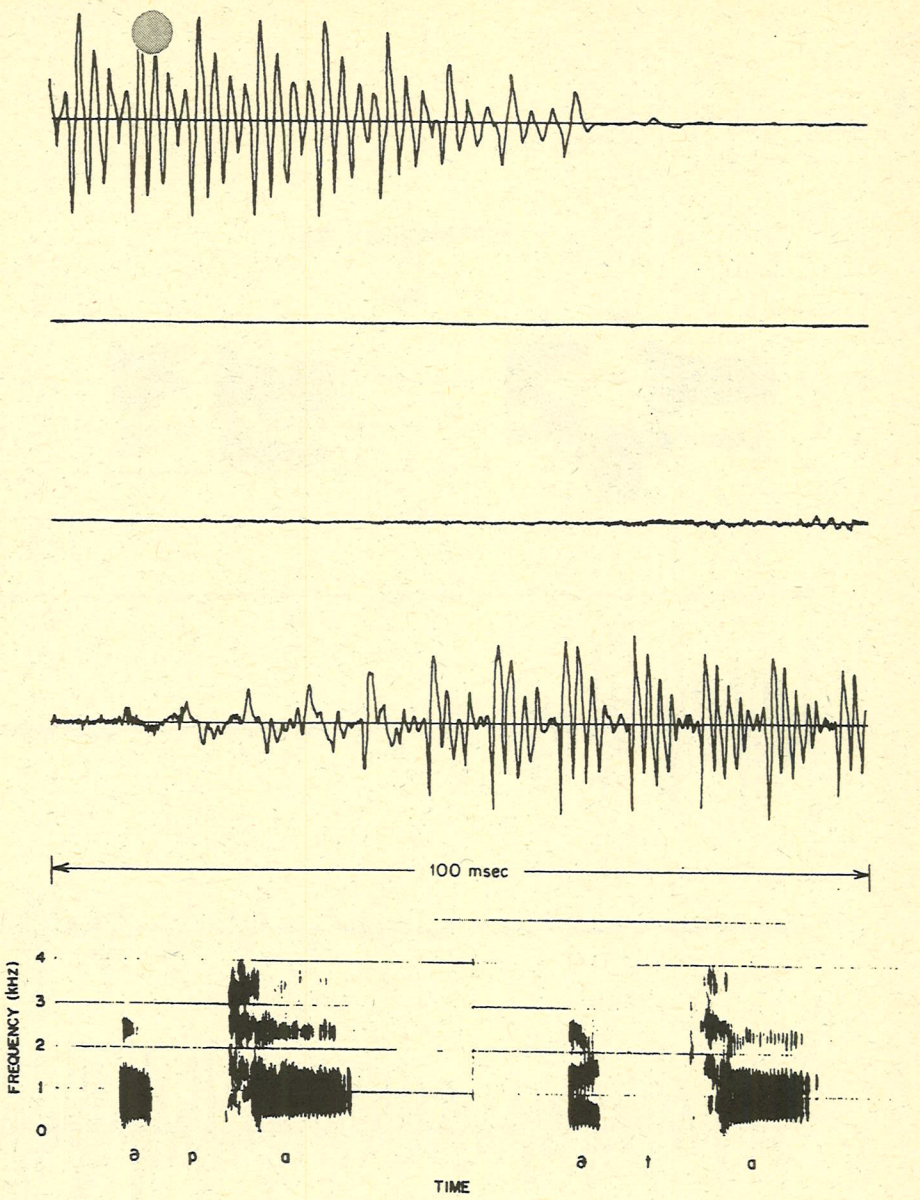


Fig. 3.12 (Continued)

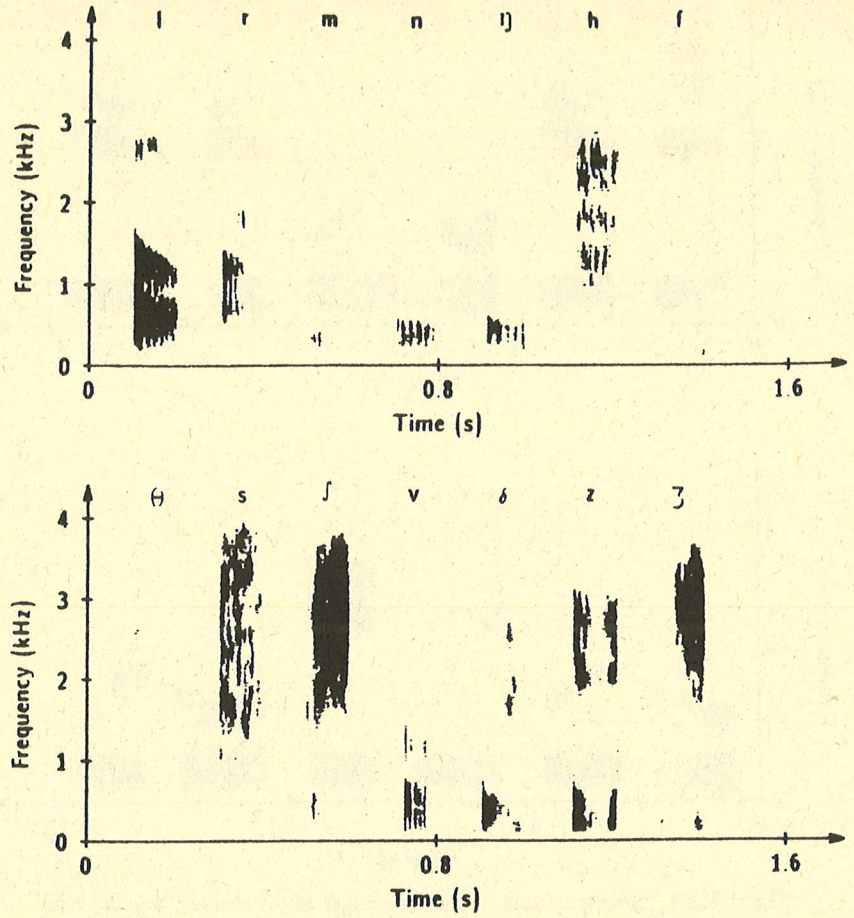


Fig. 3.18 Spectrograms of 14 English steady-state consonants /l.r,m.n.ŋ.h.f. θ,s.ʃ,v,ð,z.ʒ/.

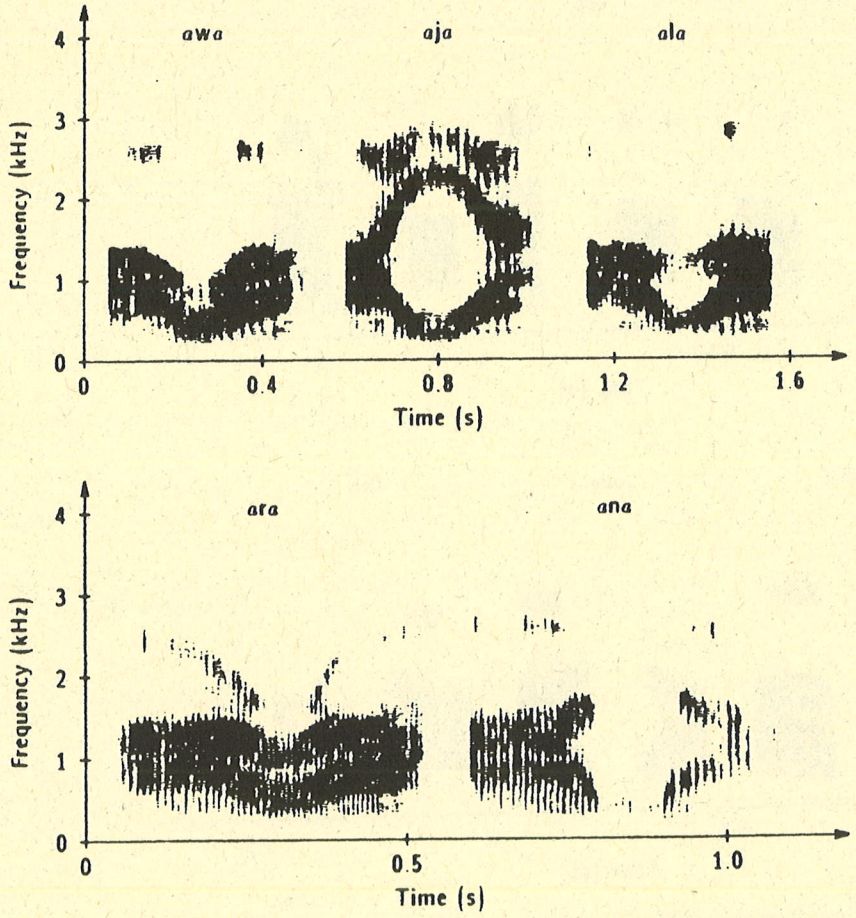


Fig. 3.19 Spectrograms of English sonorants in vocalic context: /awa.aja.ala.ara.ana/.

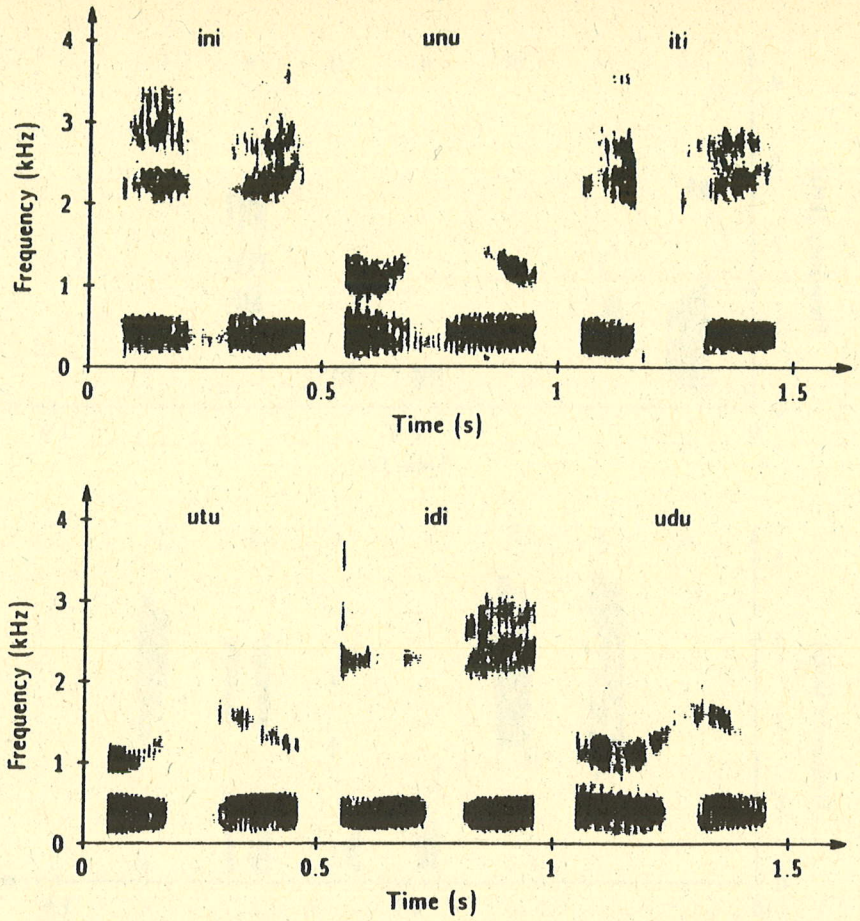


Fig. 3.20 Spectrograms of English stops and nasals in vocalic context: /ini,unu,iti,utu,idi,udu/.

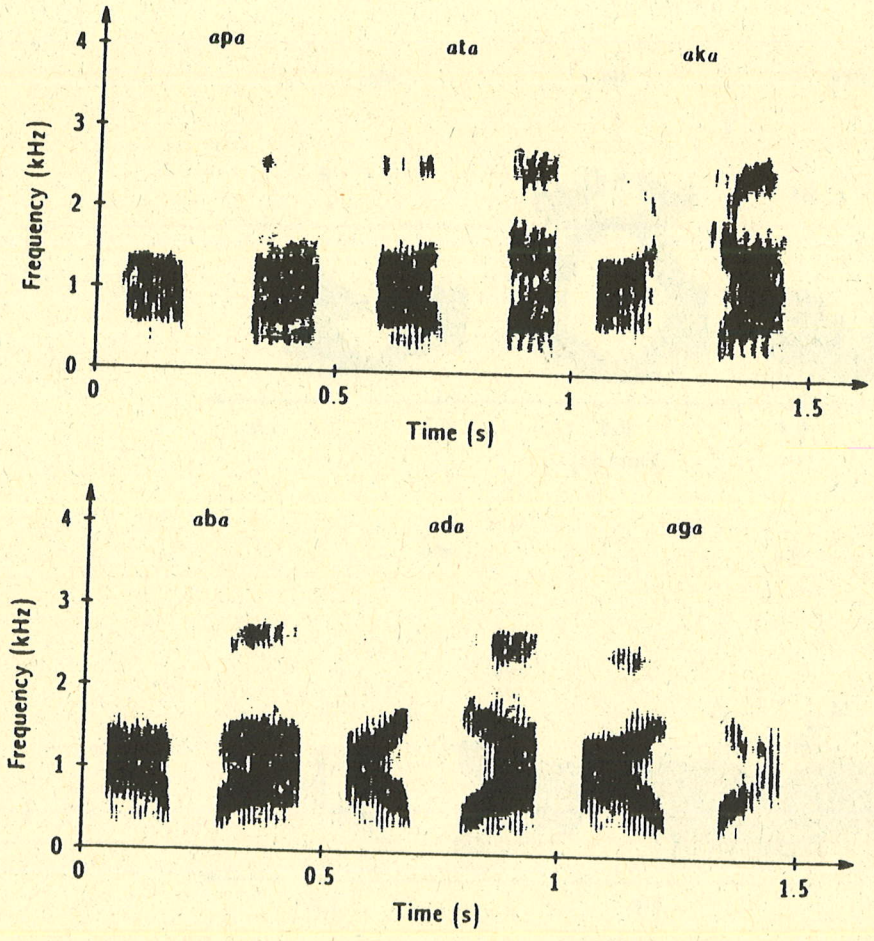


Fig. 3.21 Spectrograms of the six English stops in vocalic context: /apa,ata,aka,aba,ada,aga/.

Fig. 6.36 Wideband and narrowband spectrograms of a sentence

Tubo acústico de sección constante. Línea de transmisión.

Ecuaciones

$$-\frac{\partial u}{\partial x}=\frac{A}{\rho.c^2}\frac{\partial p}{\partial t} \quad \text{(Conservación de la masa)} \quad -\frac{\partial i}{\partial x}=C\frac{\partial v}{\partial t}$$

$$-\frac{\partial p}{\partial x}=\frac{\rho}{A}\frac{\partial u}{\partial t} \quad \text{(Conservación del momento)} \quad -\frac{\partial v}{\partial x}=L\frac{\partial i}{\partial t}$$

Analogía

$u \leftrightarrow i$	$\frac{A}{\rho c^2} \leftrightarrow C$	Compresibilidad
$p \leftrightarrow v$	$\frac{\rho}{A} \leftrightarrow L$	Inductancia acústica

Solución (Superposición onda directa y reflejada)

$$u(x,t)=u^{+}\left(t-\frac{x}{c}\right)-u^{-}\left(t+\frac{x}{c}\right)$$

$$p(x,t)=Z_0\left[u^{+}\left(t-\frac{x}{c}\right)+u^{-}\left(t+\frac{x}{c}\right)\right]$$

donde Z0 es la impedancia acústica característica (del medio):

$$Z_0=\frac{\rho.c}{A}$$

3.16

Análisis armónico

$$u(0,t) = \exp\left(j2\pi ft\right) \qquad \text{(Excitación)}$$

$$u(x,t) = U(x).\exp\left(j2\pi ft\right)$$
$$p(x,t) = P(x).\exp\left(j2\pi ft\right)$$

$$p(l,t) = 0 \quad \Rightarrow \quad P(l) = 0 \qquad \text{(Final del tubo abierto)}$$

Ecuaciones

$$\left.\begin{aligned} -\frac{\partial U(x)}{\partial x} &= \frac{A}{\rho.c^2} j.2\pi f.P(x) \\ -\frac{\partial P(x)}{\partial x} &= \frac{\rho}{A} j.2\pi f.U(x) \end{aligned}\right\} \Rightarrow \begin{aligned} \frac{\partial^2 U}{\partial x^2} + \frac{4\pi^2 f^2}{c^2} U &= 0 \\ \frac{\partial^2 P}{\partial x^2} + \frac{4\pi^2 f^2}{c^2} P &= 0 \end{aligned}$$

Soluciones

$$U(x) = a_1 \exp[\gamma.x] + a_2 \exp[-\gamma.x]$$
$$P(x) = a_3 \exp[\gamma.x] + a_4 \exp[-\gamma.x]$$

$$\gamma = j.\frac{2\pi f}{c}$$

Condiciones de contorno: $U(0) = 1; \qquad P(l) = 0$

$$U(x) = \frac{\cos\left(\frac{2\pi f}{c}(l-x)\right)}{\cos\left(\frac{2\pi f}{c} l\right)} \qquad P(x) = j.Z_0 \frac{sen\left(\frac{2\pi f}{c}(l-x)\right)}{\cos\left(\frac{2\pi f}{c} l\right)}$$

Modelo de tubos concatenados.

Ecuaciones en la sección iésima

$$u_i(x,t) = u_i^+\left(t-\frac{x}{c}\right) - u_i^-\left(t+\frac{x}{c}\right)$$

$$p_i(x,t) = Z_i\left[u_i^+\left(t-\frac{x}{c}\right) + u_i^-\left(t+\frac{x}{c}\right)\right] \quad ; \qquad Z_i = \frac{\rho c}{A_i}$$

Condiciones de continuidad en la transición i / (i+1)

$$u_i(l_i,t) = u_{i+1}(0,t) \;\Rightarrow\; u_i^+(t-\tau_i) - u_i^-(t+\tau_i) = u_{i+1}^+(t) - u_{i+1}^-(t)$$

$$p_i(l_i,t) = p_{i+1}(0,t) \;\Rightarrow\; Z_i\left[u_i^+(t-\tau_i) + u_i^-(t+\tau_i)\right] = Z_{i+1}\left[u_{i+1}^+(t) + u_{i+1}^-(t)\right]$$

donde: $\tau_i = \dfrac{l_i}{c}$

Y expresando las ondas salientes en función de las entrantes:

$$u_{i+1}^+(t) = (1+r_i).u_i^+(t-\tau_i) + r_i.u_{i+1}^-(t)$$

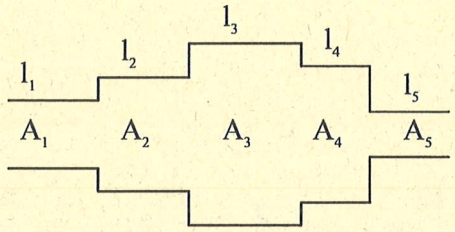
$$u_i^-(t+\tau_i) = -r_i.u_i^+(t-\tau_i) + (1-r_i).u_{i+1}^-(t)$$

O en forma vectorial:

$$\begin{bmatrix} u_{i+1}^+(t) \\ u_i^-(t+\tau_i) \end{bmatrix} = \begin{bmatrix} 1+r_i & r_i \\ -r_i & 1-r \end{bmatrix} \begin{bmatrix} u_i^+(t-\tau_i) \\ u_{i+1}^-(t) \end{bmatrix}$$

donde: $r_i = \dfrac{Z_i - Z_{i+1}}{Z_i + Z_{i+1}} = \dfrac{A_{i+1} - A_i}{A_{i+1} + A_i}$

Modelo de tubos concatenados.



$$Z_i = \frac{\rho.c}{A_i}$$

Ondas directas y reflejadas en la transición de los tubos i / i+1

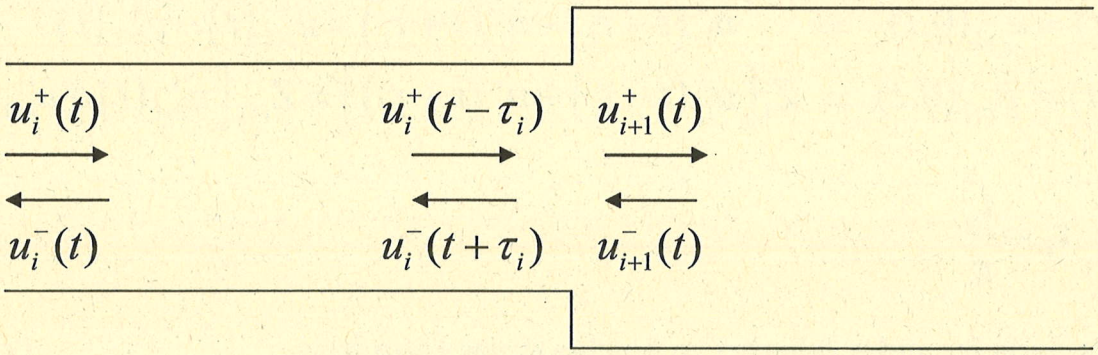
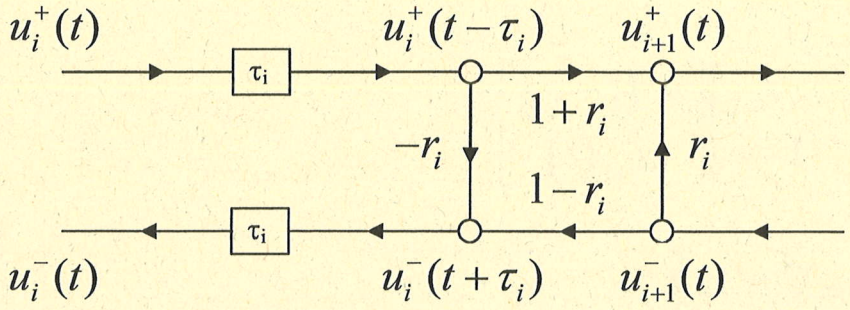


Diagrama de flujo en la transición de los tubos i / i+1

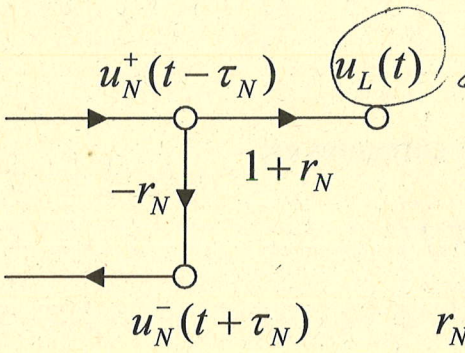


$$r_i = \frac{Z_i - Z_{i+1}}{Z_i + Z_{i+1}}$$

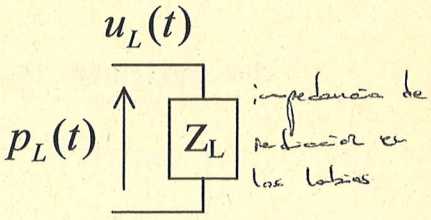
Condiciones de contorno: Radiación en los Labios.

- **Se asume que el medio exterior no refleja nada.**
- **La radiación se modela como una impedancia Z_L**

toda la energía que se emite va fuera y no vuelve → sistema exterior muy amplio



sólo hay onda volumétrica saliente

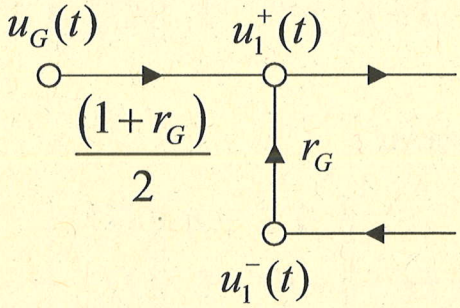


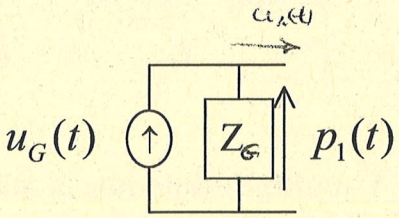
$$r_N = \frac{Z_N - Z_L}{Z_N + Z_L}$$

en general, $Z_L = Z_L(f)$

Condiciones de contorno: Glotis.

- **La generación del sonido se modela como un generador de corriente con impedancia Z_G en paralelo.**





$$r_G = \frac{Z_G - Z_1}{Z_G + Z_1}$$

Respuesta en frecuencia

Análisis armónico:

$$\left.\begin{aligned} u_G(t) &= U_G \exp(j2\pi ft) \\ u_L(t) &= U_L \exp(j2\pi ft) \end{aligned}\right\} \quad \Rightarrow \quad H(f) = \frac{U_G}{U_L}$$

Las relaciones de ondas salientes en función de entrantes es:

$$\begin{bmatrix} u_{i+1}^{+}(t) \\ u_i^{-}(t+\tau_i) \end{bmatrix} = \begin{bmatrix} 1+r_i & r_i \\ -r_i & 1-r_i \end{bmatrix} \begin{bmatrix} u_i^{+}(t-\tau_i) \\ u_{i+1}^{-}(t) \end{bmatrix}$$

Y las del tubo i en función de las del tubo i+1:

$$\begin{bmatrix} u_i^{+}(t-\tau_i) \\ u_i^{-}(t+\tau_i) \end{bmatrix} = \frac{1}{1+r_i} \begin{bmatrix} 1 & -r_i \\ -r_i & 1 \end{bmatrix} \begin{bmatrix} u_{i+1}^{+}(t) \\ u_{i+1}^{-}(t) \end{bmatrix}$$

Que para señales armónicas resulta:

referidos para u_i

$$\begin{bmatrix} \exp(-j2\pi f\tau_i) & 0 \\ 0 & \exp(j2\pi f\tau_i) \end{bmatrix} \begin{bmatrix} U_i^{+} \\ U_i^{-} \end{bmatrix} = \frac{1}{1+r_i} \begin{bmatrix} 1 & -r_i \\ -r_i & 1 \end{bmatrix} \begin{bmatrix} U_{i+1}^{+} \\ U_{i+1}^{-} \end{bmatrix}$$

Y multiplicando por la matriz de exponenciales inversa:

$$\begin{bmatrix} U_i^{+} \\ U_i^{-} \end{bmatrix} = \frac{1}{1+r_i} \begin{bmatrix} \exp(j2\pi f\tau_i) & 0 \\ 0 & \exp(-j2\pi f\tau_i) \end{bmatrix} \begin{bmatrix} 1 & -r_i \\ -r_i & 1 \end{bmatrix} \begin{bmatrix} U_{i+1}^{+} \\ U_{i+1}^{-} \end{bmatrix}$$

Para la Glotis y los Labios tenemos las relaciones:

$$U_G = \frac{2}{1+r_G}\begin{bmatrix}1 & -r_G\end{bmatrix}\begin{bmatrix}U_1^+ \\ U_1^-\end{bmatrix}$$

$$\begin{bmatrix}U_N^+ \\ U_N^-\end{bmatrix} = \frac{1}{1+r_N}\begin{bmatrix}\exp(j2\pi f\tau_N) & 0 \\ 0 & \exp(-j2\pi f\tau_N)\end{bmatrix}\begin{bmatrix}1 \\ -r_N\end{bmatrix}.U_L$$

Y concatenando todas las relaciones obtenemos:

$$U_G = \frac{2}{1+r_G}\begin{bmatrix}1 & -r_G\end{bmatrix}\prod_{i=1}^{N-1}\left\{\frac{1}{1+r_i}.\underline{\underline{\mathbf{\Gamma}}}_i.\underline{\underline{\mathbf{R}}}_i\right\}\frac{1}{1+r_N}.\underline{\underline{\mathbf{\Gamma}}}_N\begin{bmatrix}1 \\ -r_N\end{bmatrix}.U_L$$

Siendo:

$$\underline{\underline{\mathbf{\Gamma}}}_i = \begin{bmatrix}\exp(j2\pi f\tau_i) & 0 \\ 0 & \exp(-j2\pi f\tau_i)\end{bmatrix}; \qquad \underline{\underline{\mathbf{R}}}_i = \begin{bmatrix}1 & -r_i \\ -r_i & 1\end{bmatrix}$$

La expresión de la respuesta en frecuencia resulta ser:

$$H(f) = \frac{U_L}{U_G} = \frac{(1+r_G)\prod_{i=1}^{N}\{1+r_i\}}{2\begin{bmatrix}1 & -r_G\end{bmatrix}\prod_{i=1}^{N-1}\left\{\underline{\underline{\mathbf{\Gamma}}}_i.\underline{\underline{\mathbf{R}}}_i\right\}.\underline{\underline{\mathbf{\Gamma}}}_N\begin{bmatrix}1 \\ -r_N\end{bmatrix}}$$

Respuesta en frecuencia (Modelo de tiempo discreto)

Haciendo el cambio:

$$\exp(j2\pi f\tau) = z^{1/2}$$

en la expresión de la respuesta en frecuencia de los tubos concatenados obtenemos la función del sistema del filtro discreto:

$$H(z) = \frac{(1+r_G)\prod_{i=1}^{N}\{1+r_i\}.z^{-N/2}}{2[1 \quad -r_G]\prod_{i=1}^{N-1}\{\underline{\underline{\mathbf{Q}}}_i\}\begin{bmatrix}1\\ -r_N.z^{-1}\end{bmatrix}}$$

donde la matriz $\mathbf{Q}_i$ es:

$$\underline{\underline{\mathbf{Q}}}_i = z^{-1/2}.\underline{\underline{\mathbf{\Gamma}}}.\underline{\underline{\mathbf{R}}}_i = z^{-1/2}.\begin{bmatrix}z^{1/2} & 0\\ 0 & z^{-1/2}\end{bmatrix}.\begin{bmatrix}1 & -r_i\\ -r_i & 1\end{bmatrix} = \begin{bmatrix}1 & -r_i\\ -r_i z^{-1} & z^{-1}\end{bmatrix}$$

Se trata de un sistema de solo polos (N): (cuadro AR)

$$H(z) = \frac{b.z^{-N/2}}{1-\sum_{k=1}^{N} a_k.z^{-k}}$$

2.21

ENERGÍA LOCALIZADA

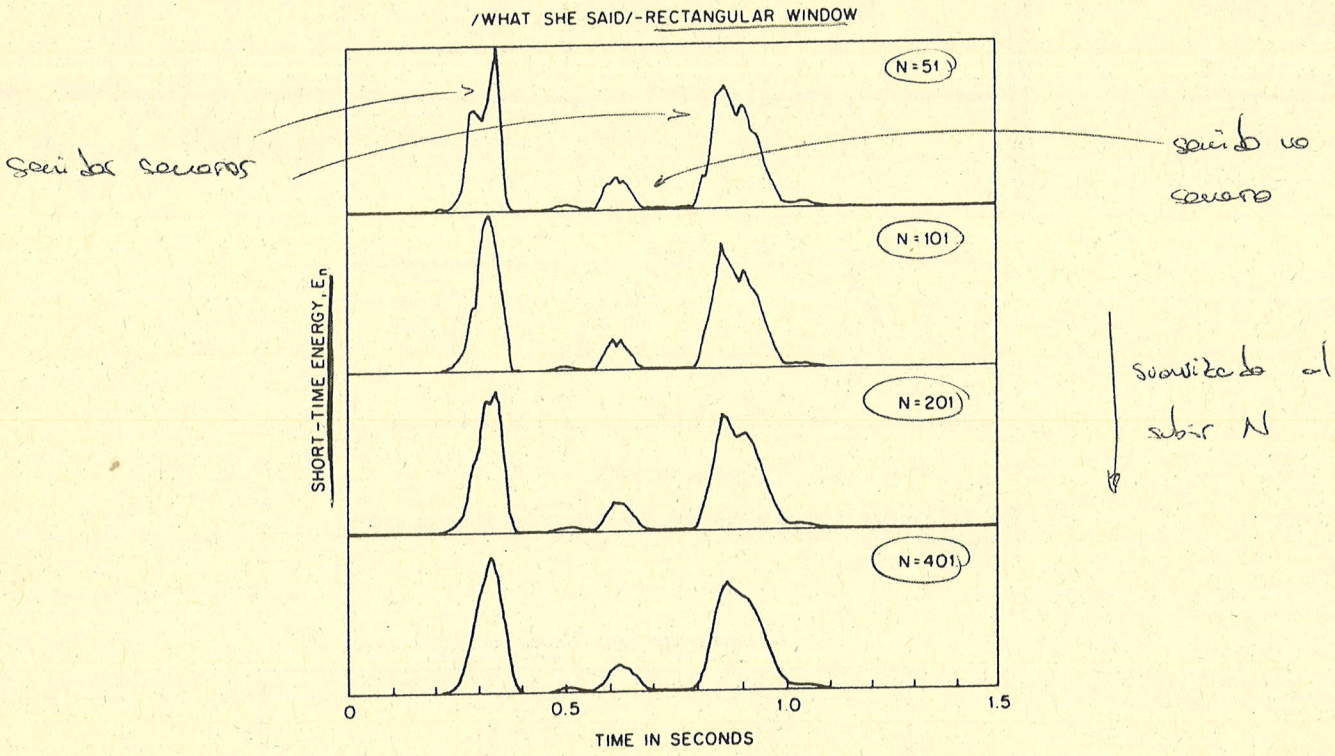
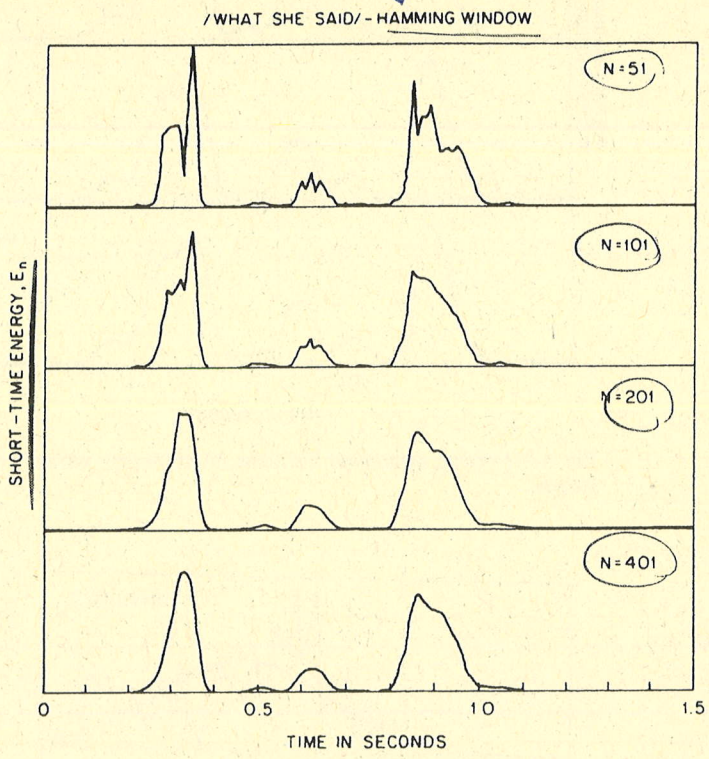


Fig. 4.6 Short-time energy functions for rectangular windows of various lengths.

ventana de Hamming: $h(n) = \begin{cases} 0{,}54 - 0{,}46\cos\frac{2\pi n}{N-1} & (0 \le n \le N-1) \\ 0 & \text{resto} \end{cases}$



mayor ancho de banda para igual duración de ventana que la rectangular

Fig. 4.7 Short-time energy functions for Hamming windows of various lengths.

picos y valles se ven acentuados que en la energía local

⇓

menor dificultad para representarlo en puto fijo

MAGNITUD LOCAL

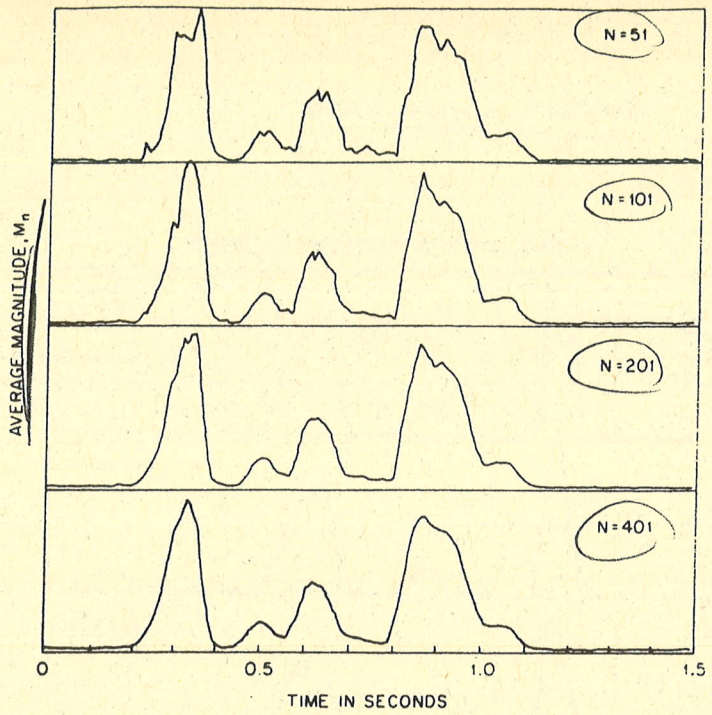


Fig. 4.8 Average magnitude functions for rectangular windows of various lengths.

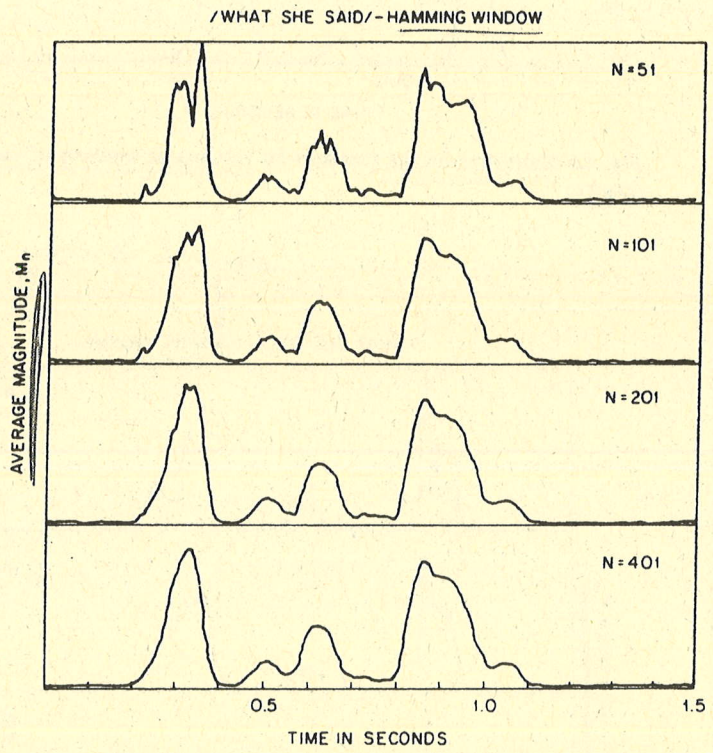
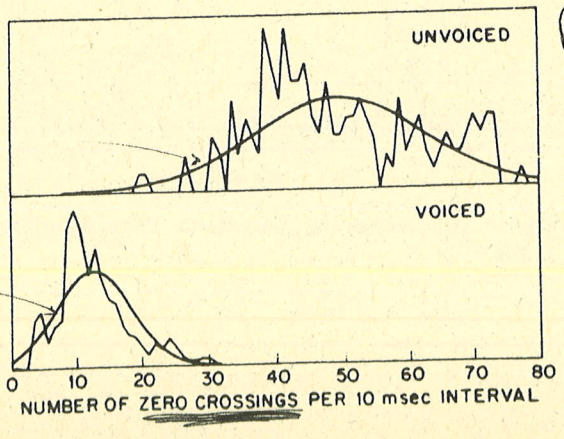


Fig. 4.9 Average magnitude functions for Hamming windows of various lengths.

CRUCES POR CERO

se intenta aproximar por una campana



frecuencias no sonoras

frecuencias sonoras

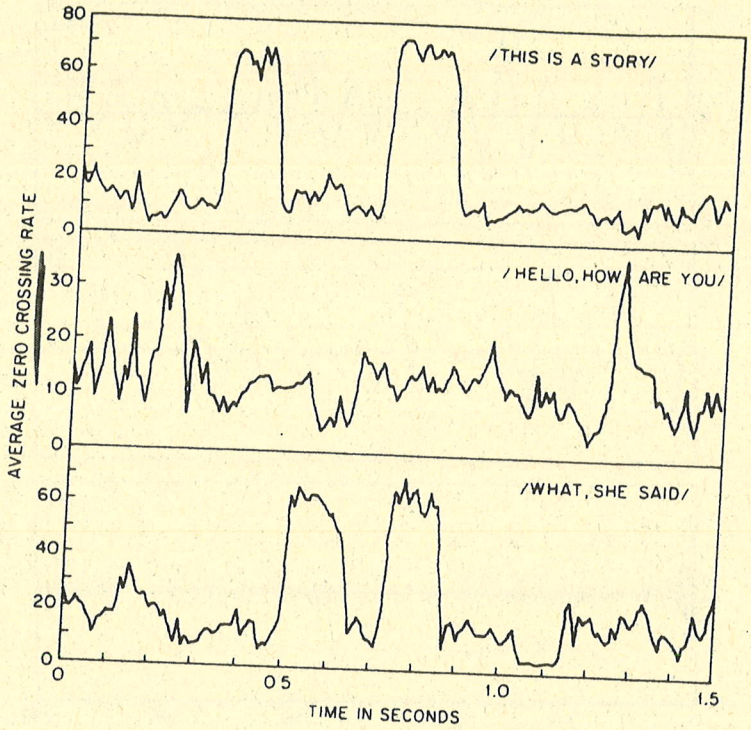


Fig. 4.12 Average zero-crossing rate for three different utterances.

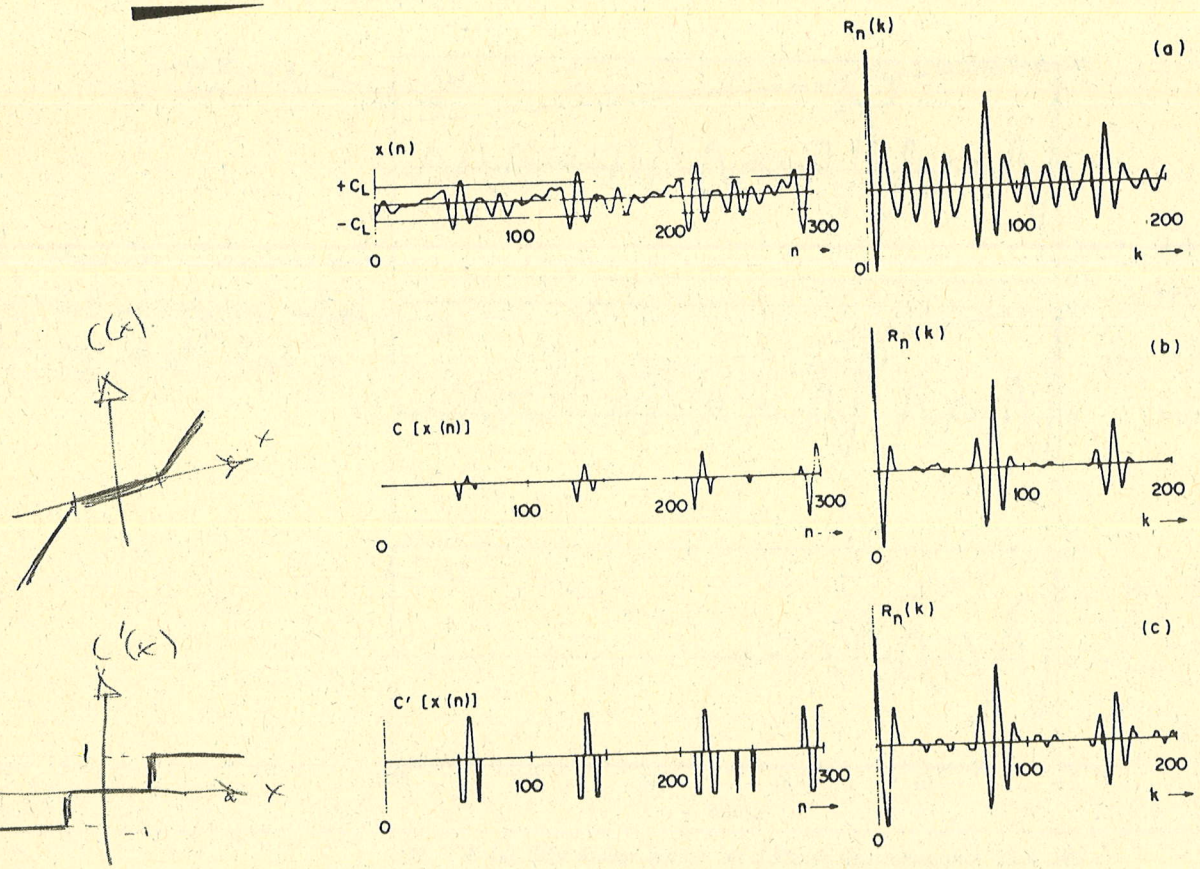


Fig. 4.33 Example of waveforms and correlation function; (a) no clipping; (b) center clipped; (c) 3-level center clipped. (All correlation functions normalized to 1.0.) (After Rabiner [18].)

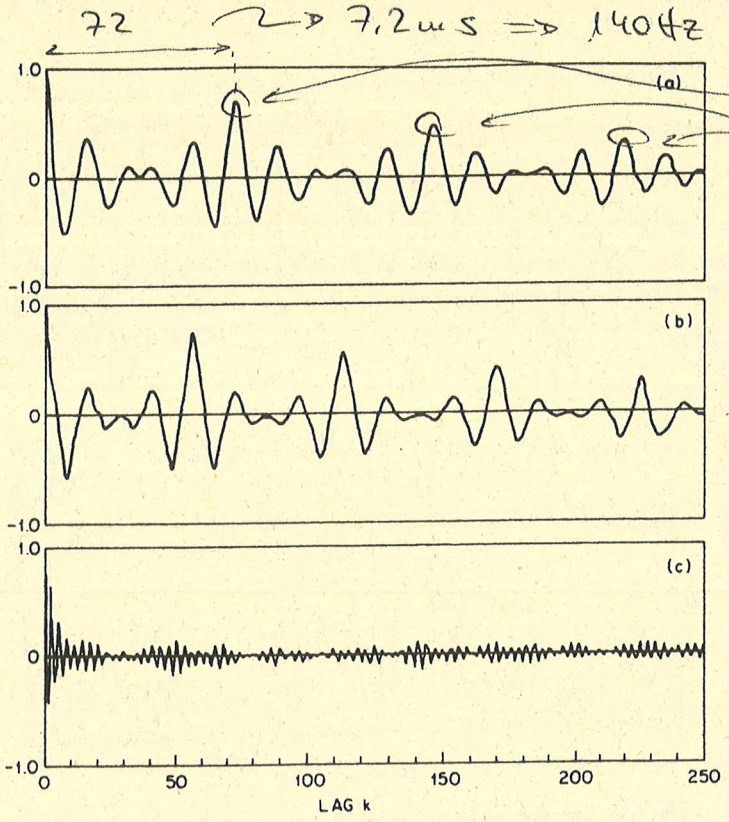


Fig. 4.24 Autocorrelation function for (a) and (b) voiced speech; and (c) unvoiced speech, using a rectangular window with $N = 401$.

Señal de voz muestreada a 10kHz

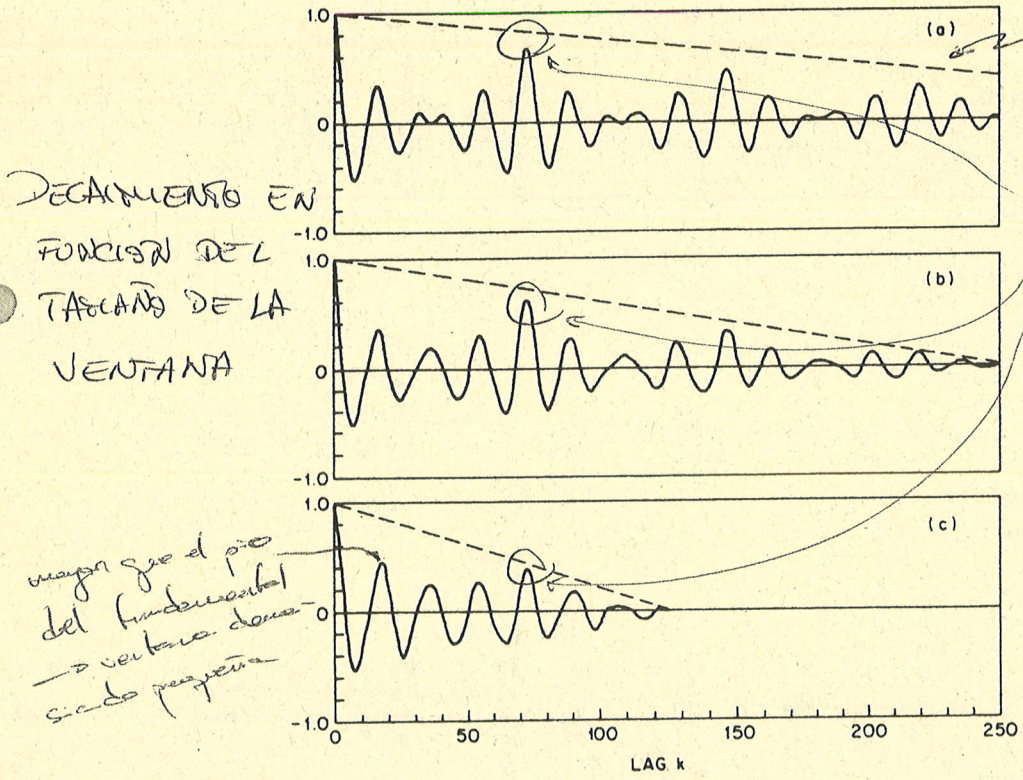


Fig. 4.26 Autocorrelation function for voiced speech with (a) $N = 401$; (b) $N = 251$; and (c) $N = 125$. Rectangular window used in all cases.

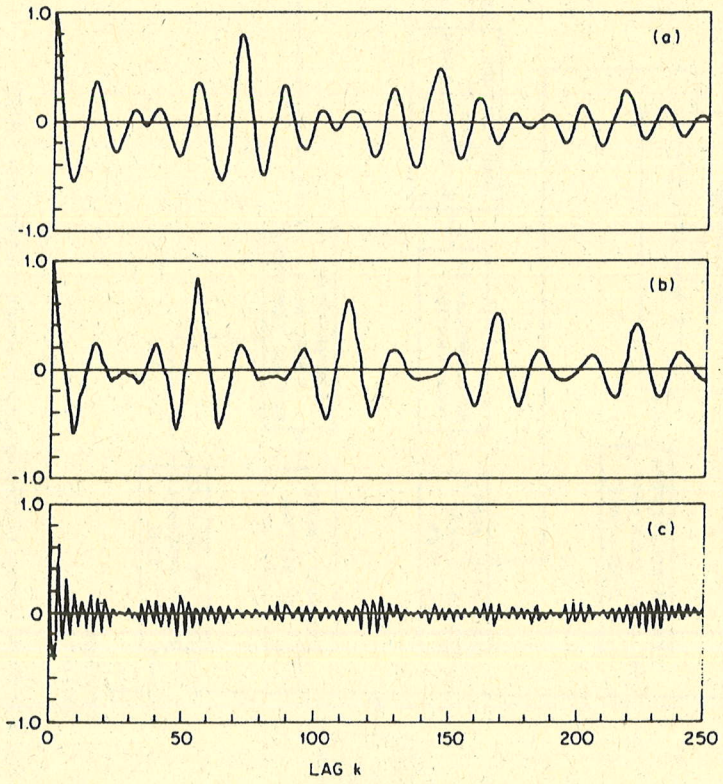


Fig. 4.28 Modified autocorrelation function for speech segments of Fig. 4.24 with $N = 401$.

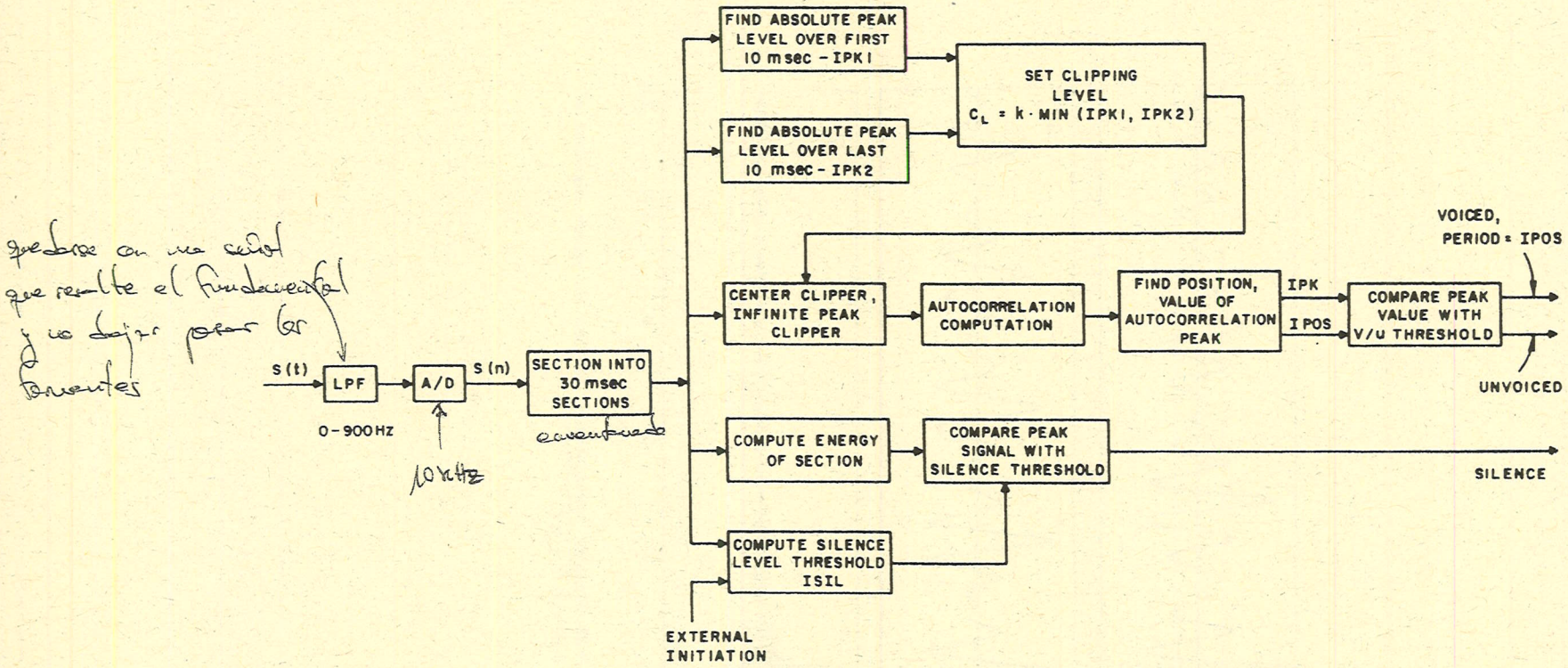


Fig. 4.36 Block diagram of clipping autocorrelation pitch detector. (After Dubnowski et al. [19].)

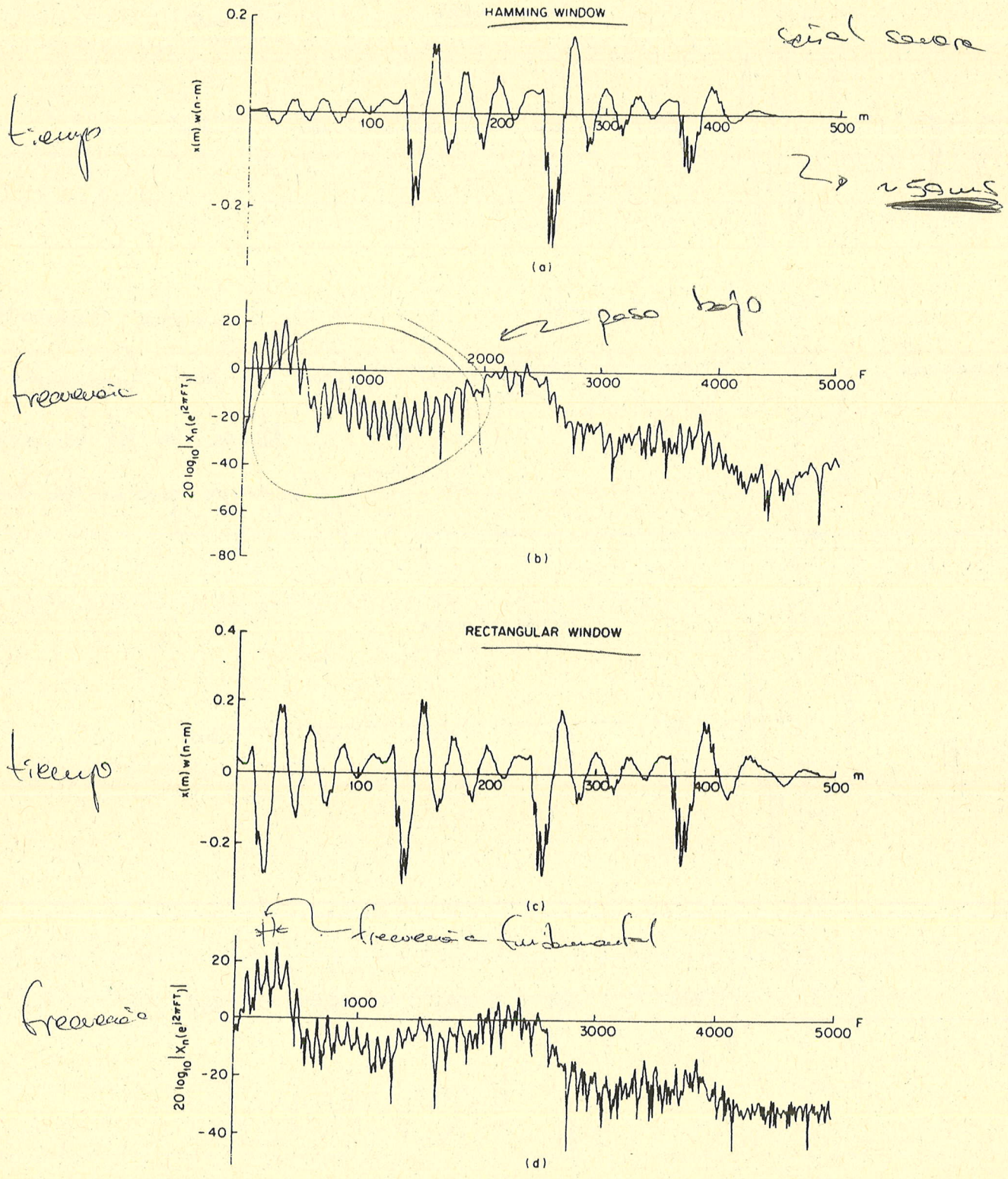


Fig. 6.2 Spectrum analysis for voiced speech using a 50 msec (a,b) Hamming window; (c,d) rectangular window. Parts (a) and (c) show time waveforms; parts (b) and (d) show corresponding spectra.

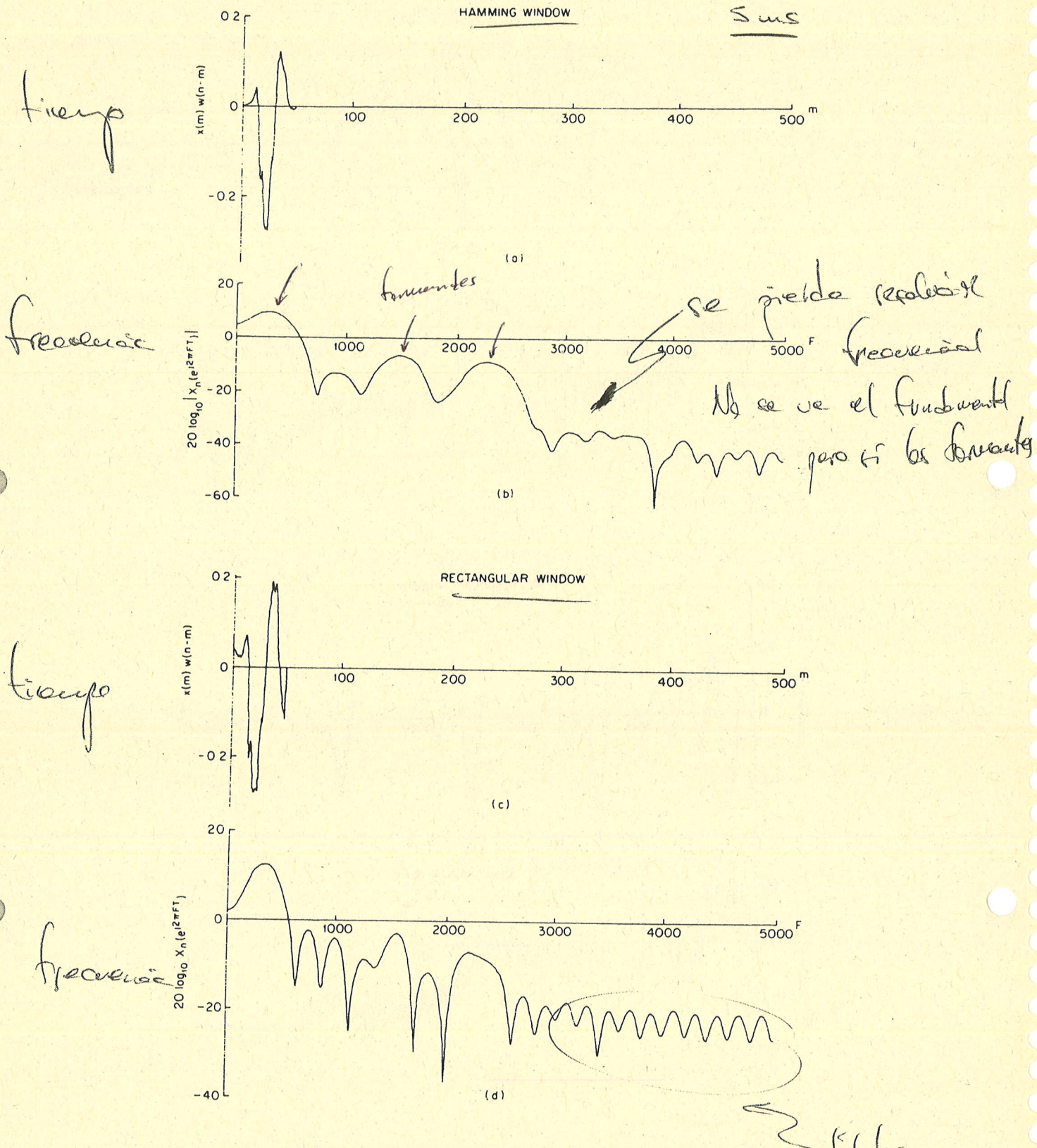


Fig. 6.3 Spectrum analysis of voiced speech using a 5 msec (a,b) Hamming window; (c,d) rectangular window. Parts (a) and (c) show time waveforms; parts (b) and (d) show corresponding spectra.

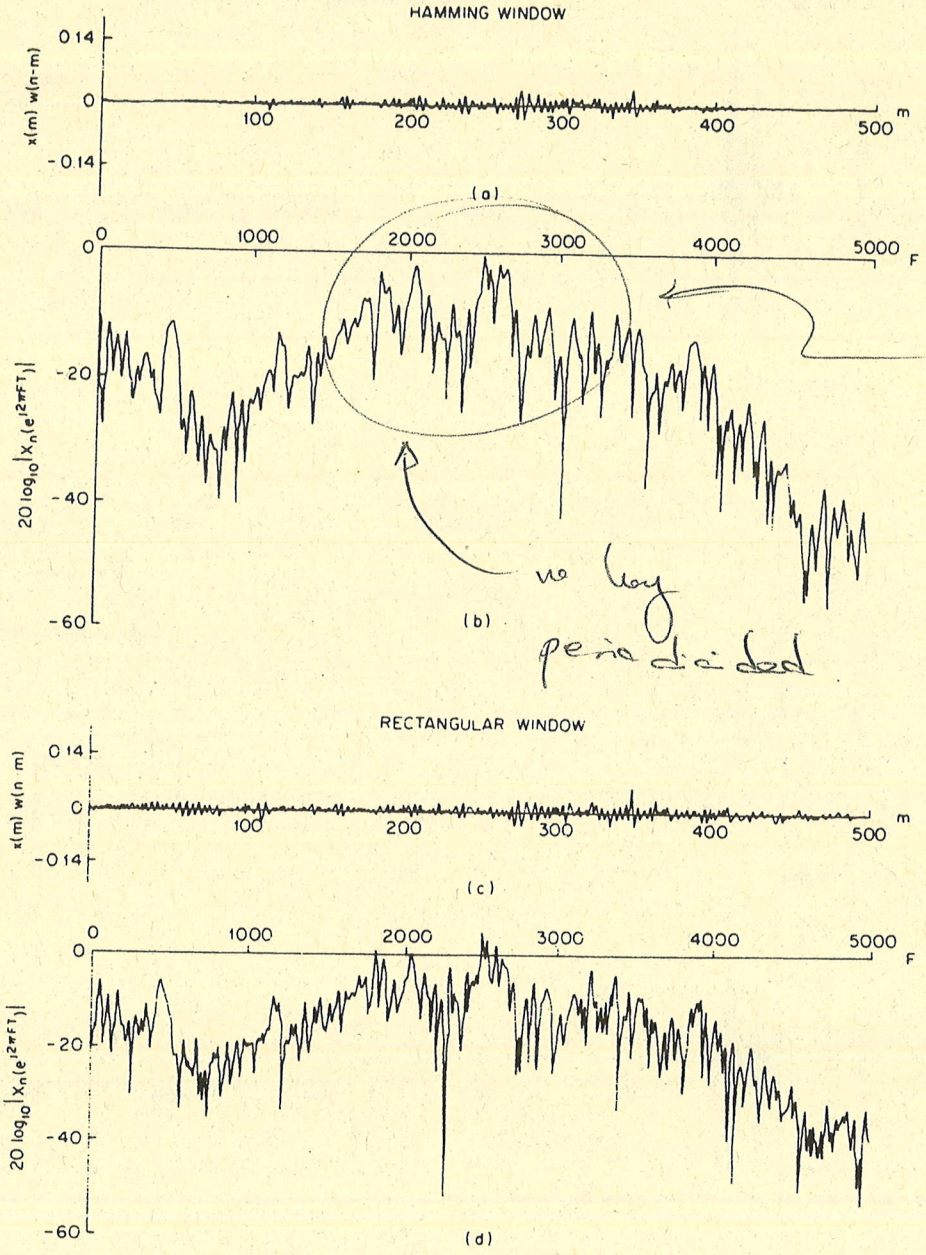
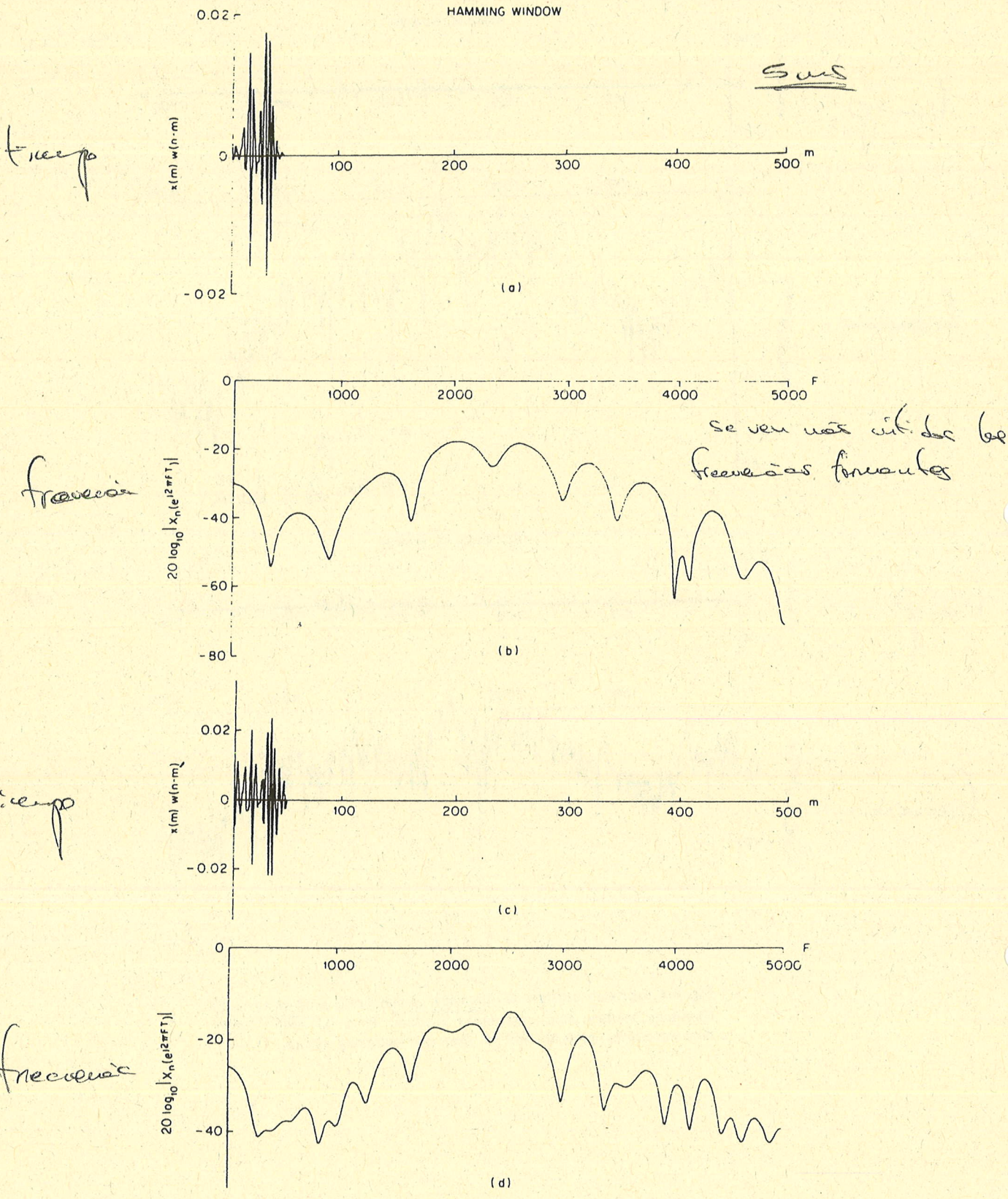


Fig. 6.4 Spectrum analysis of unvoiced speech using a 50 msec (a,b) Hamming window; (c,d) rectangular window. Parts (a) and (c) show time waveforms; parts (b) and (d) show corresponding spectra.

SONIDO FRICATIVO NO SONORO



Sus

tiempo

frecuencia

Se ven más definidas las frecuencias formantes

tiempo

frecuencia

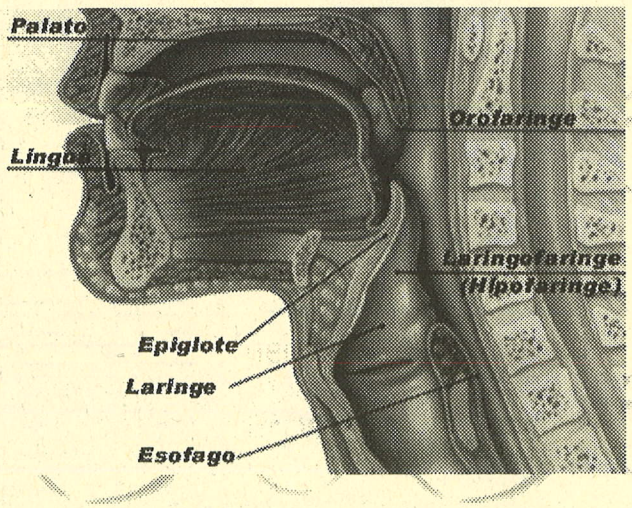
Fig. 6.5 Spectrum analysis of unvoiced speech using a 5 msec (a,b) Hamming window; (c,d) rectangular window. Parts (a) and (c) show time waveforms; parts (b) and (d) show corresponding spectra.

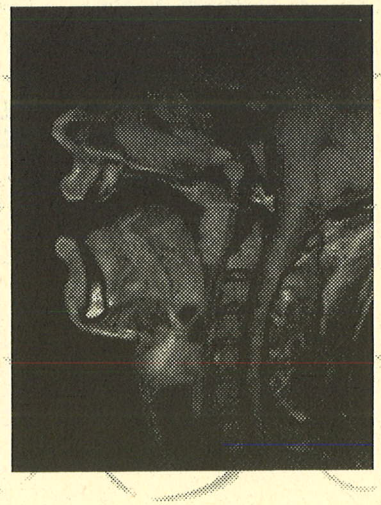
La señal de voz

Tratamiento Digital de la Señal II

Enrique Nava

La voz: fisiología

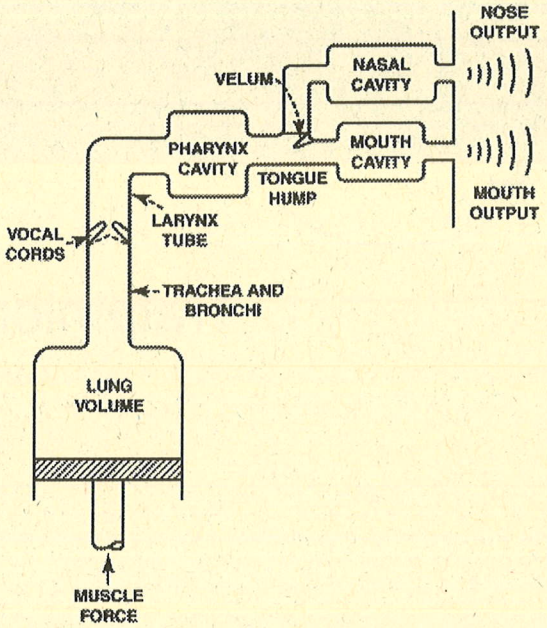




Fonemas

- s(t) = s(p,v; t)

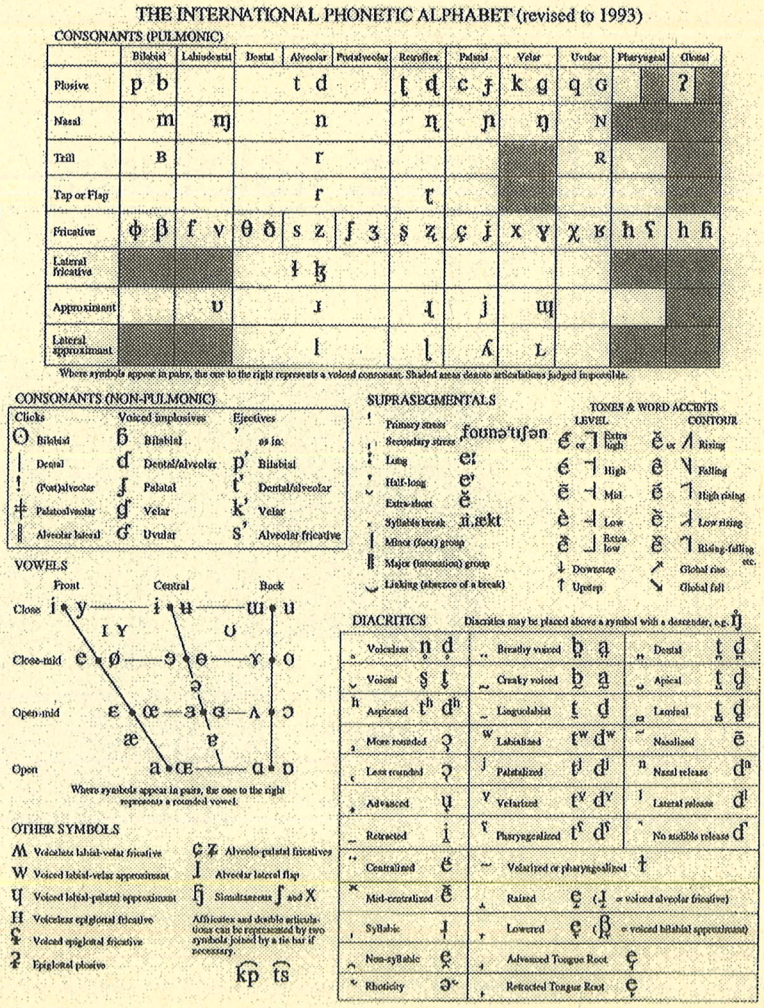
  sonido = ondas de presión acústica acopladas al movimiento de un fluido



Flanagan: "Speech Analysis and Perception", Springer-Verlag 1965

Fonemas

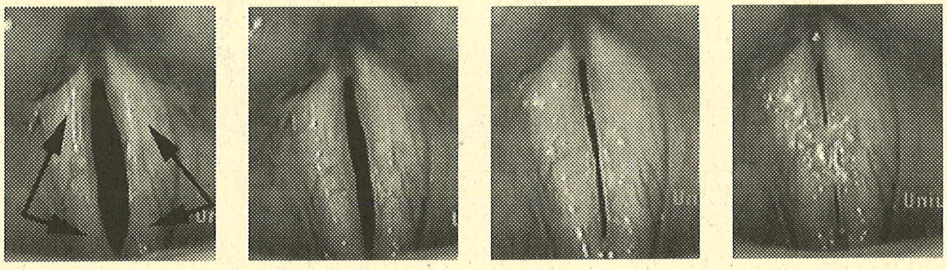
- Alfabeto fonético internacional (IPA)



International Phonetic Association
(c/o Department of Linguistics,
University of Victoria, Victoria,
British Columbia, Canada)

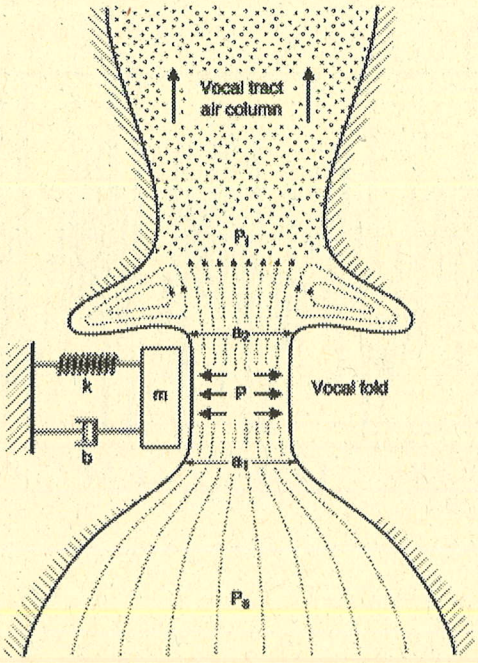
Fonemas: tipos de sonidos

- Sonoros: las cuerdas vocales vibran:
  - Sonidos casi-periódicos (vocales)



- No sonoros (sordos)
  - Cuerdas vocales abiertas, sonidos "tipo ruido" /s/

Sonido sonoro



Frecuencia fundamental

Tensión longitudinal

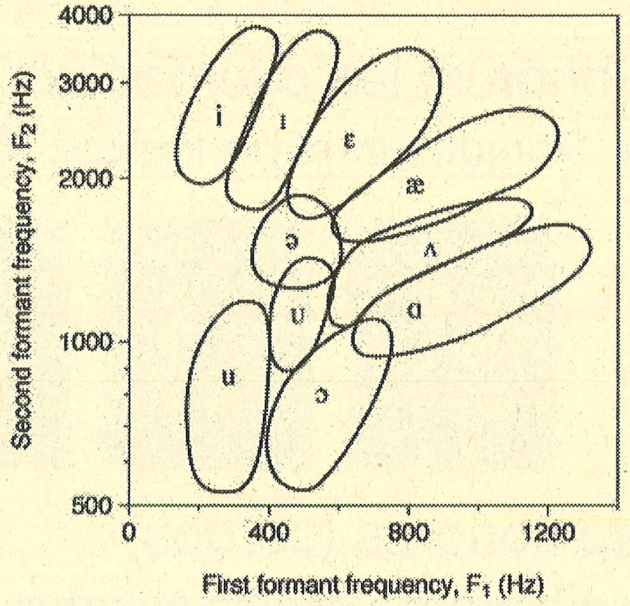
$$F_0 = \frac{1}{2L}\sqrt{\frac{\sigma}{\rho}}$$

Densidad del tejido

Longitud de las cuerdas

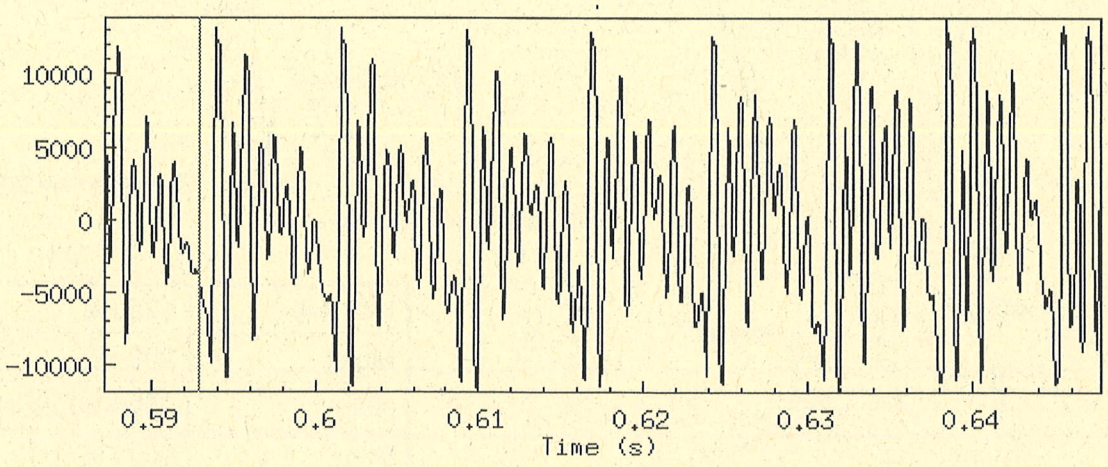
Hombre	125 Hz
Mujer	200 Hz
Niño	250-400 Hz
¡Bebé	500 Hz

Vocales



Peterson, Barney 1952

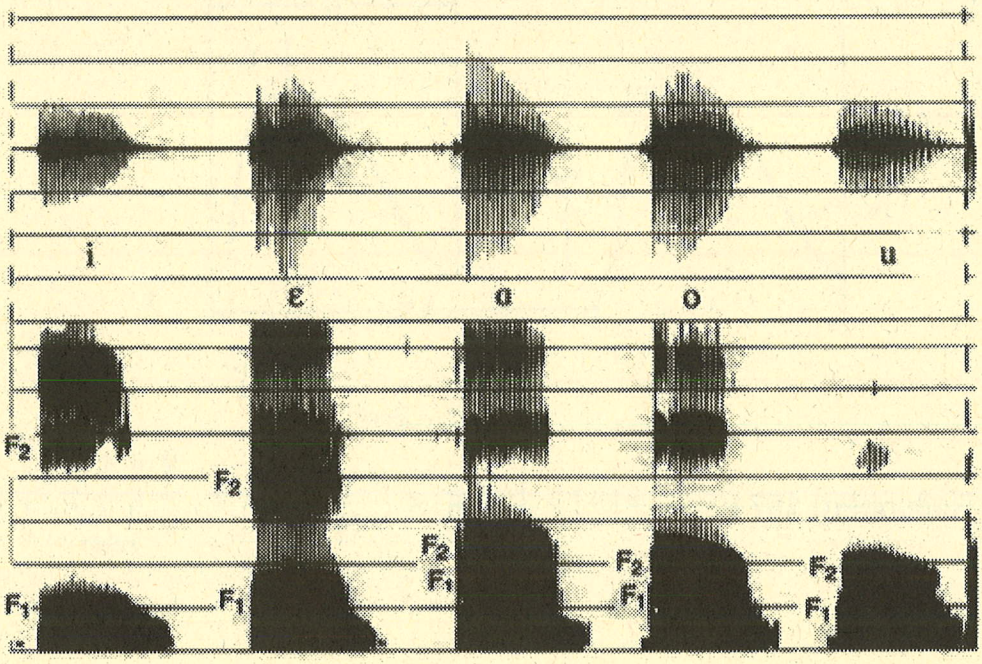
Vocales



L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978

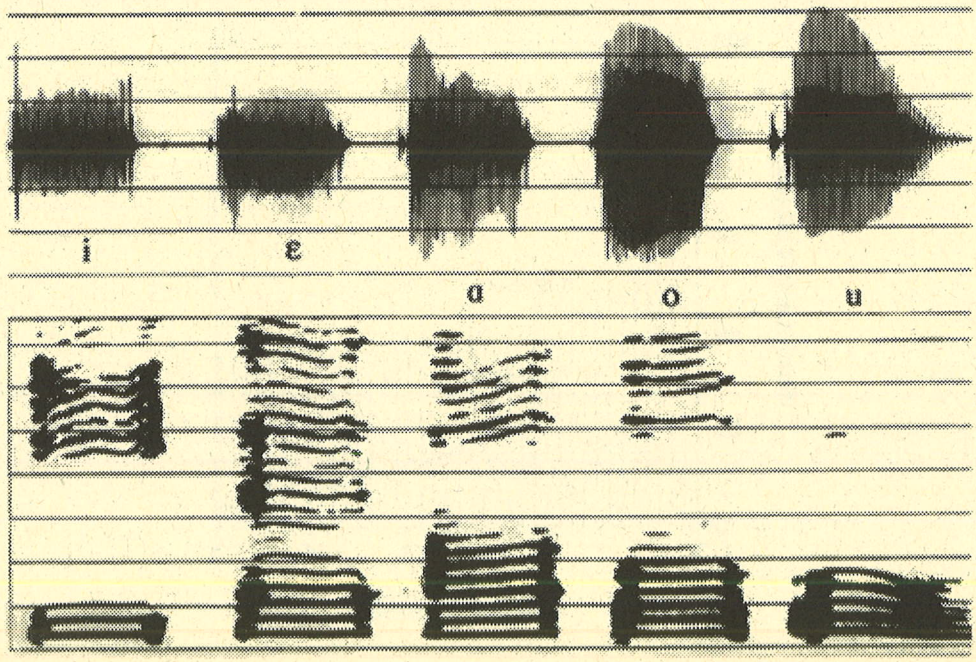
Espectrograma

Time waveform and broadband spectrogram of [i ɛ a o u].

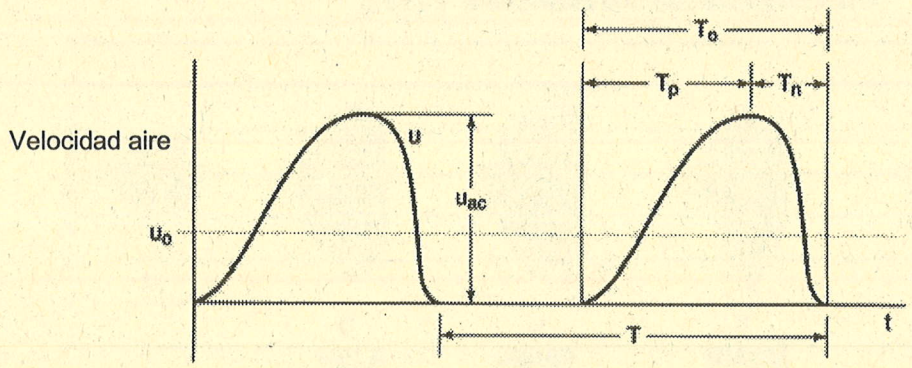


Espectrograma

Time waveform and narrowband spectrogram of [i ɛ a o u].



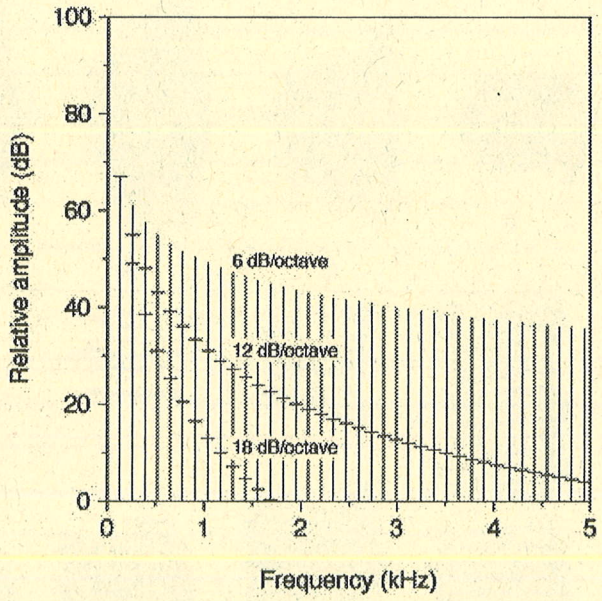
Pulsos glotales



Rosenberg 1971:

$$u[n] = \begin{cases} \frac{1}{2}\left(1 - \cos\left(\frac{\pi n}{T_p}\right)\right) & 0 \le n \le T_p \\ \cos\left(\pi \frac{n - T_p}{2T_n}\right) & T_p \le n \le T_p + T_n \\ 0 & \textit{resto} \end{cases}$$

Efecto de radiación



Modelo de voz

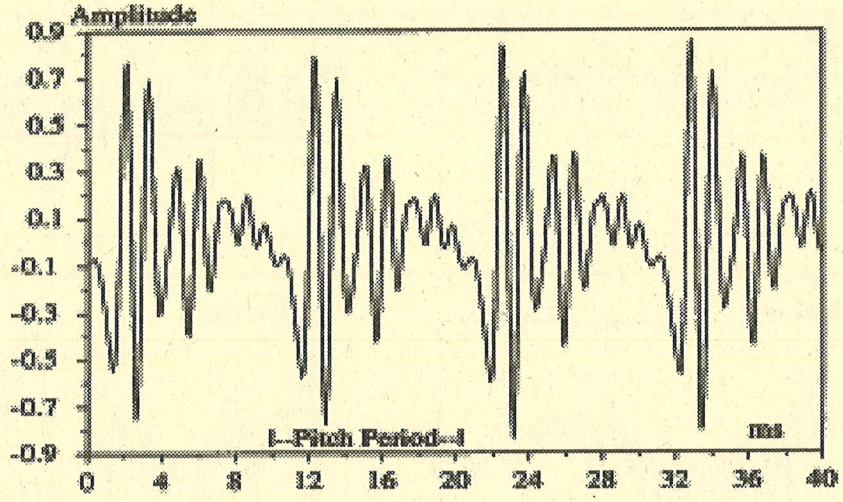


FIGURE 2.3
Time-domain waveform of a short segment of voiced speech, x-axis units in ms, y axis is relative amplitude of sound pressure.

Modelo de voz

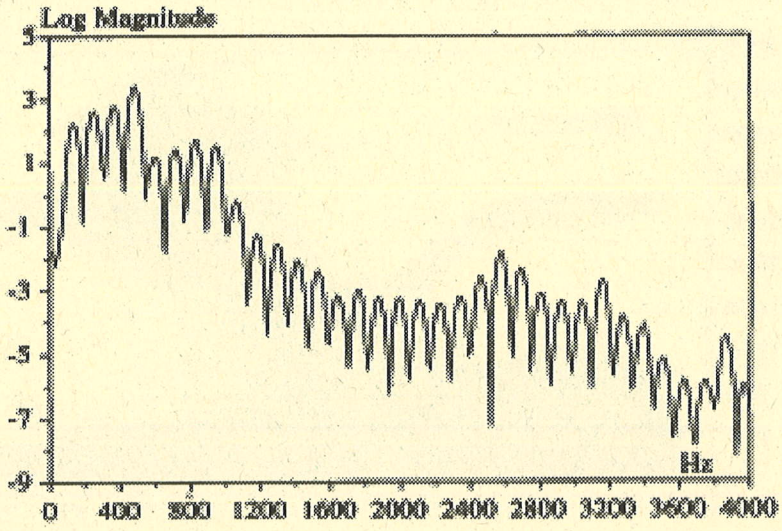


FIGURE 2.4
Log magnitude spectrum of a short segment of voiced speech, X axis units in Hz.

Fonemas

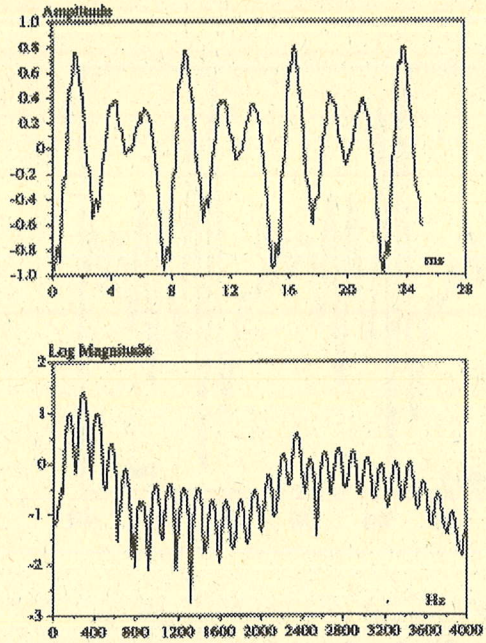


FIGURE 2.5
Time waveform and log magnitude spectrum of /I/, as in the word "hit."

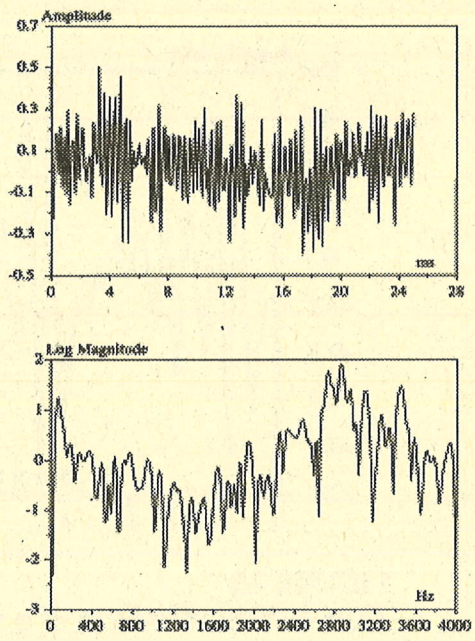


FIGURE 2.7
Time waveform and log magnitude spectrum of /sh/, as in the beginning of the word "shop."

Sonograma

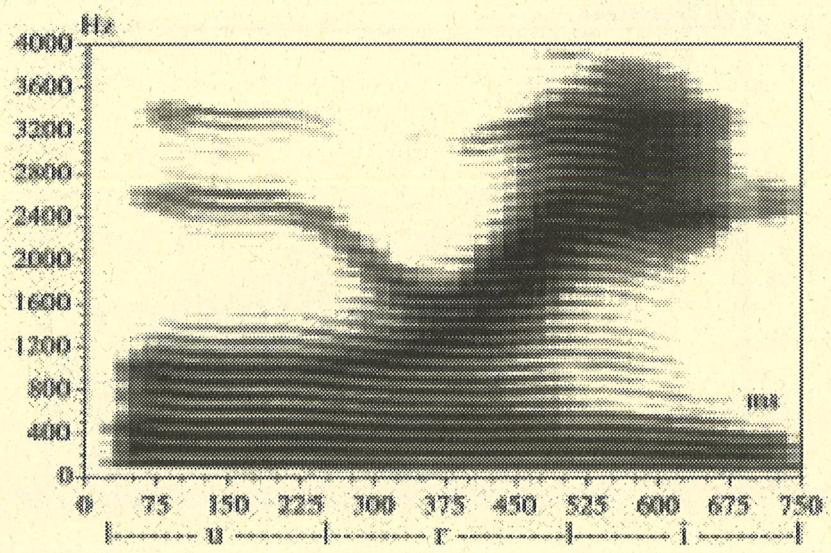
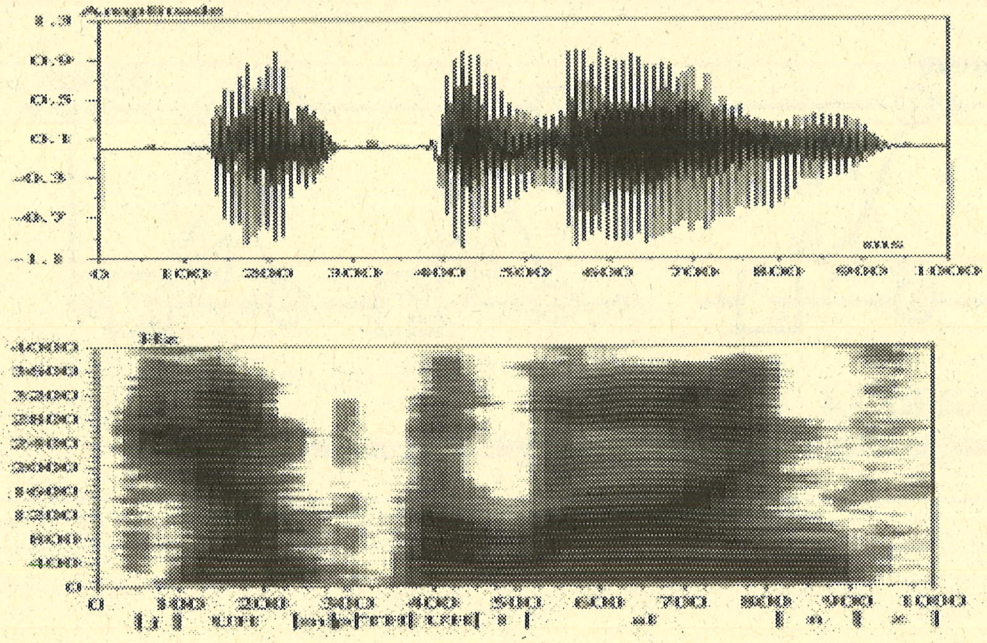


FIGURE 2.11
Spectrogram of nonword utterance /u-r-i/.

Sonogramas



Modelo de tubos

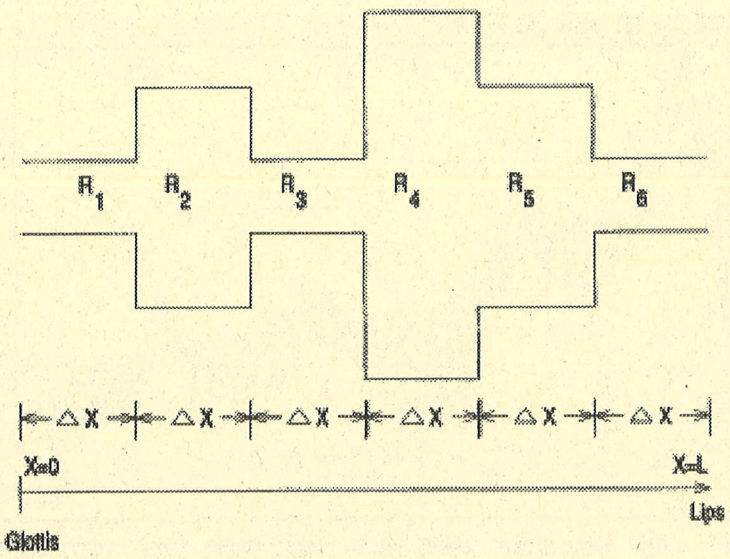


FIGURE 4.3
Multiple concatenated tube model.

Modelo de tubos

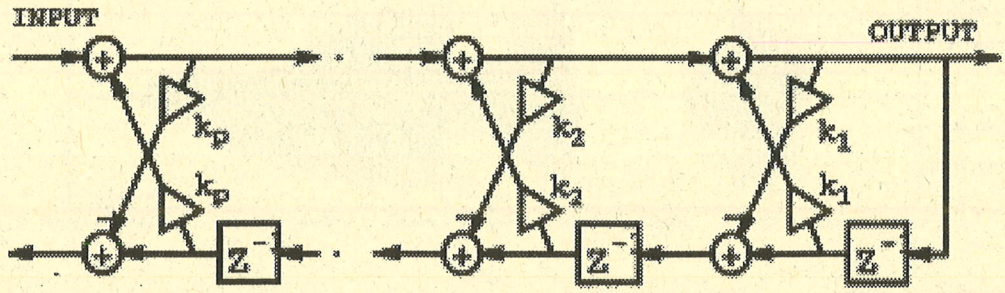


FIGURE 4.4
Lattice filter realization of multiple-tube model.

Modelo de tubos

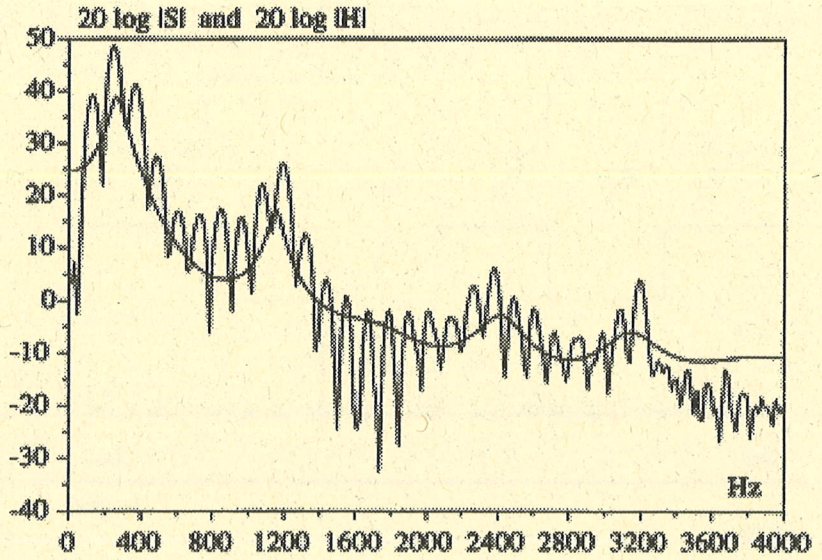


FIGURE 4.6
Log magnitude of DFT and LP spectra for a segment of voiced speech.

Modelo de tubos

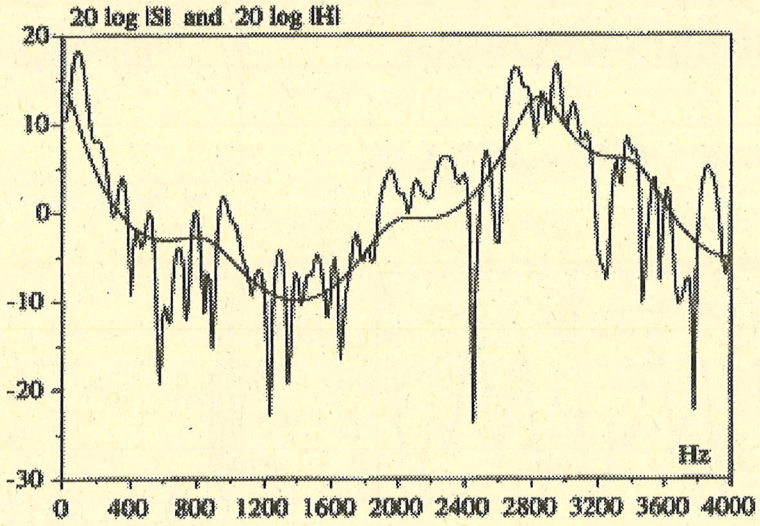


FIGURE 4.7
Log magnitude of DFT and LP spectra for a segment of unvoiced speech.

Detección del tono fundamental

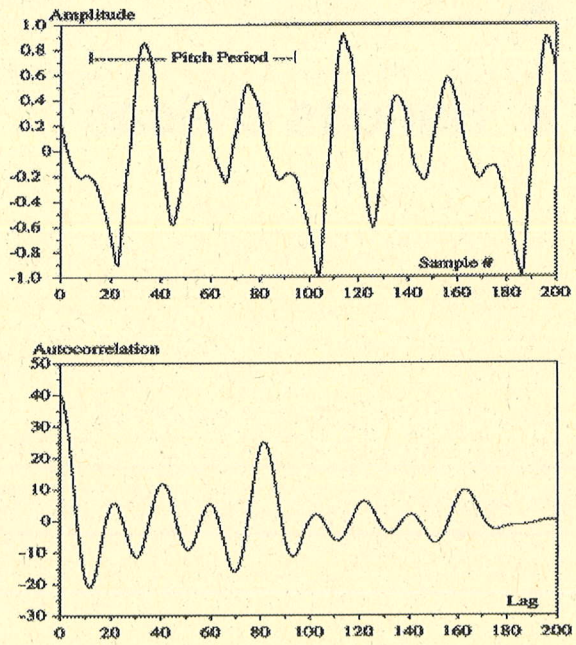
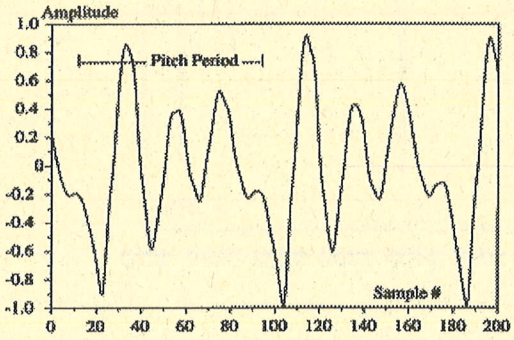


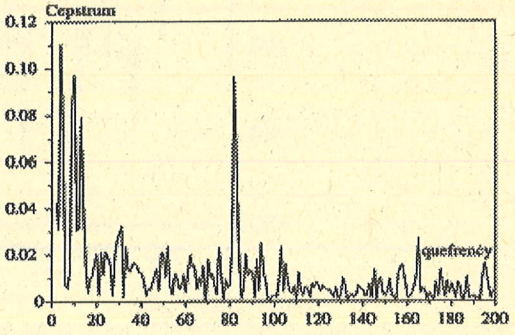
FIGURE 5.1
Time-domain waveform and autocorrelation of a short segment of viced speech.

Detección del tono fundamental

- Análisis cepstral

$$\text{Cepstrum}(d) = \text{IFFT}(\log_{10}|\text{FFT}(s(n))|)$$





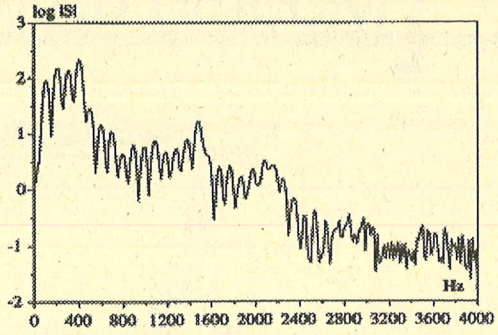


FIGURE 5.6
Log magnitude of DFT and cepstrum of speech segment of Figure 5.1.

Percepción humana

- Bandas críticas

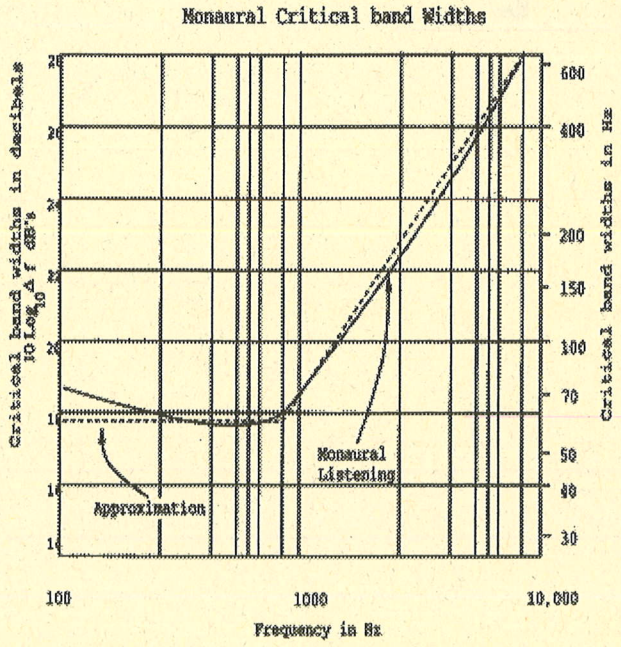


FIGURE 6.1
Frequency width of critical bands as a function of the band center frequency.

Escalas de frecuencia

Critical Band No. (Barks)	Frequency (Hz)	Mels
1	20-100	0-150
2	100-200	150-300
3	200-300	300-400
4	300-400	400-500
5	400-510	500-600
6	510-630	600-700
7	630-770	700-800
8	770-920	800-950
9	920-1080	950-1050
10	1080-1270	1050-1150
11	1270-1480	1150-1300
12	1480-1720	1300-1400
13	1720-2000	1400-1550
14	2000-2320	1550-1700
15	2320-2700	1700-1850
16	2700-3150	1850-2000
17	3150-3700	2000-2150
18	3700-4400	2150-2300
19	4400-5300	2300-2500
20	5300-6400	2500-2700
21	6400-7200	2700-2850
22	7200-9500	2850-3050

Table 6.1 The relationship between the frequency units: Barks, Hertz, and Mels.

Audibilidad

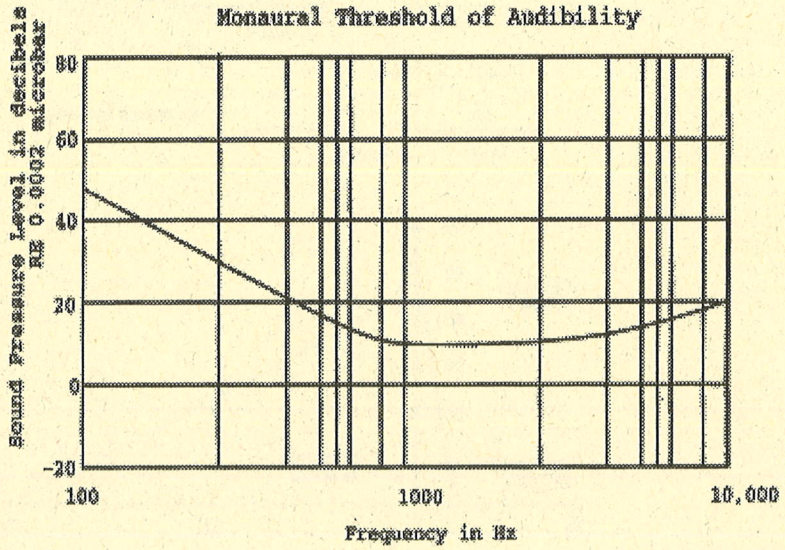


FIGURE 6.3
Threshold of audibility for a pure tone in silence.

Enmascaramiento espectral

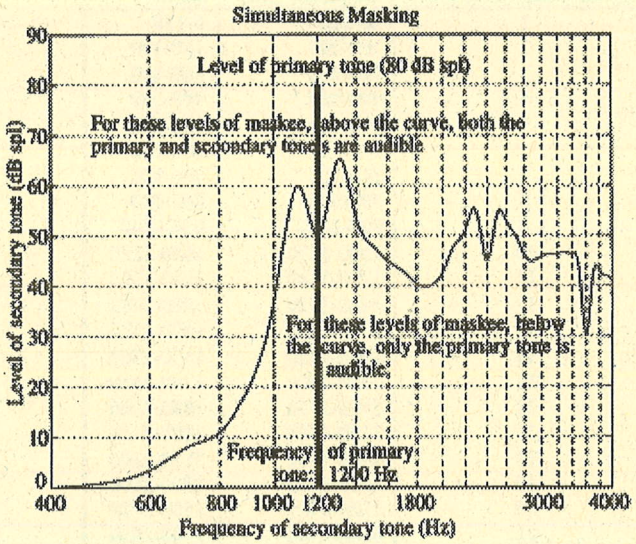


FIGURE 6.4
Simultaneous masking in frequency of one tone on another tone (data adapted from [81]).

Enmascaramiento temporal

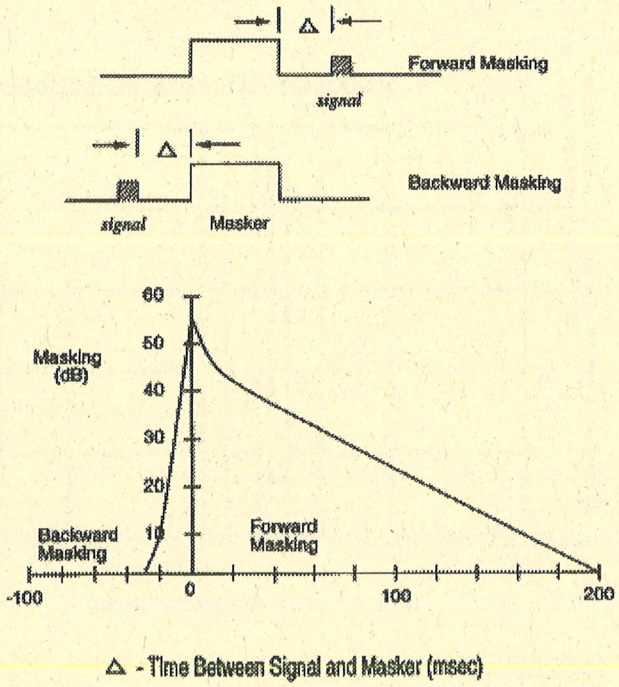


FIGURE 6.5
Illustration of the effect of temporal masking.

Enmascaramiento psicoacústico

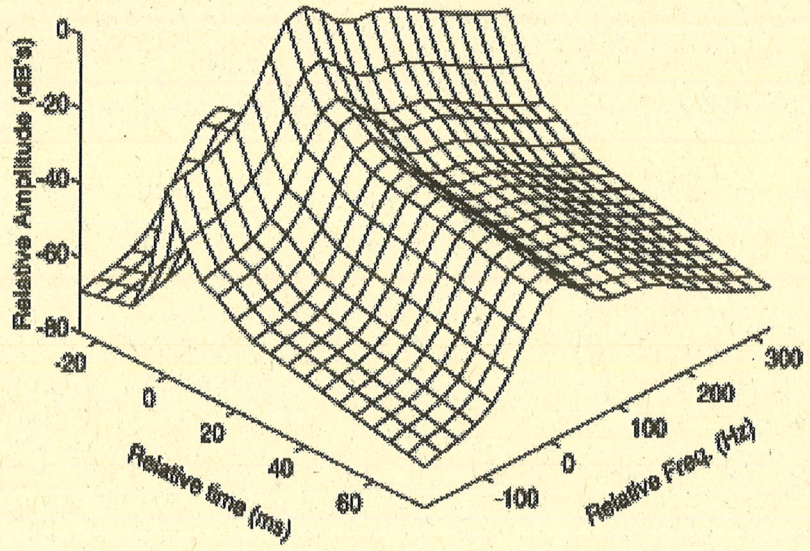


FIGURE 12.3
Psycho-acoustic masking data, both temporal and frequency.