Procesamiento de la señal de voz

Leandro Vignolo

Procesamiento Digital de Señales
Ingeniería Informática FICH-UNL

29 de mayo de 2014

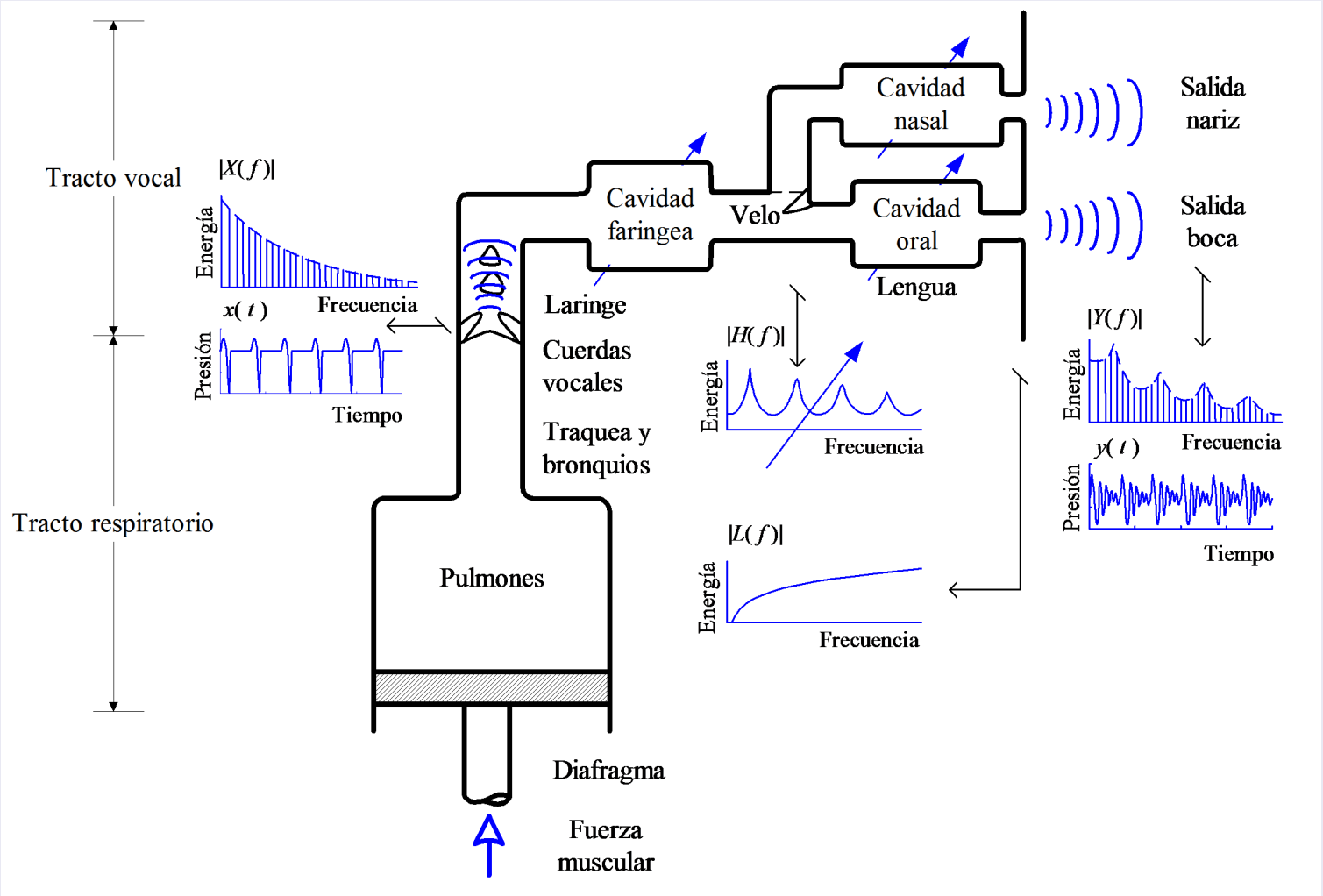
Organización de la clase

1. Producción y percepción de la voz
   - Generalidades del aparato fonador
   - Fuentes y modificadores del sonido de la voz
   - Generalidades del oído y la percepción
2. Análisis por tramos
   - Niveles estructurales del habla
   - Análisis por tramos
3. Procesamiento homomórfico
   - Definición de los coeficientes cepstrales
   - Procesamiento homomórfico de la voz
4. Estimación de la F0
   - Estimación de F0 por cepstrum
   - Estimación de F0 por autocorrelación

Organización de la clase

Modelo lineal de producción de la voz

Diagrama esquemático del aparato fonador



Modelo lineal de producción de la voz

- Se supone que la señal es la salida de un sistema lineal
- La señal de voz es el resultado de la convolución entre una señal de excitación y la respuesta al impulso del tracto vocal

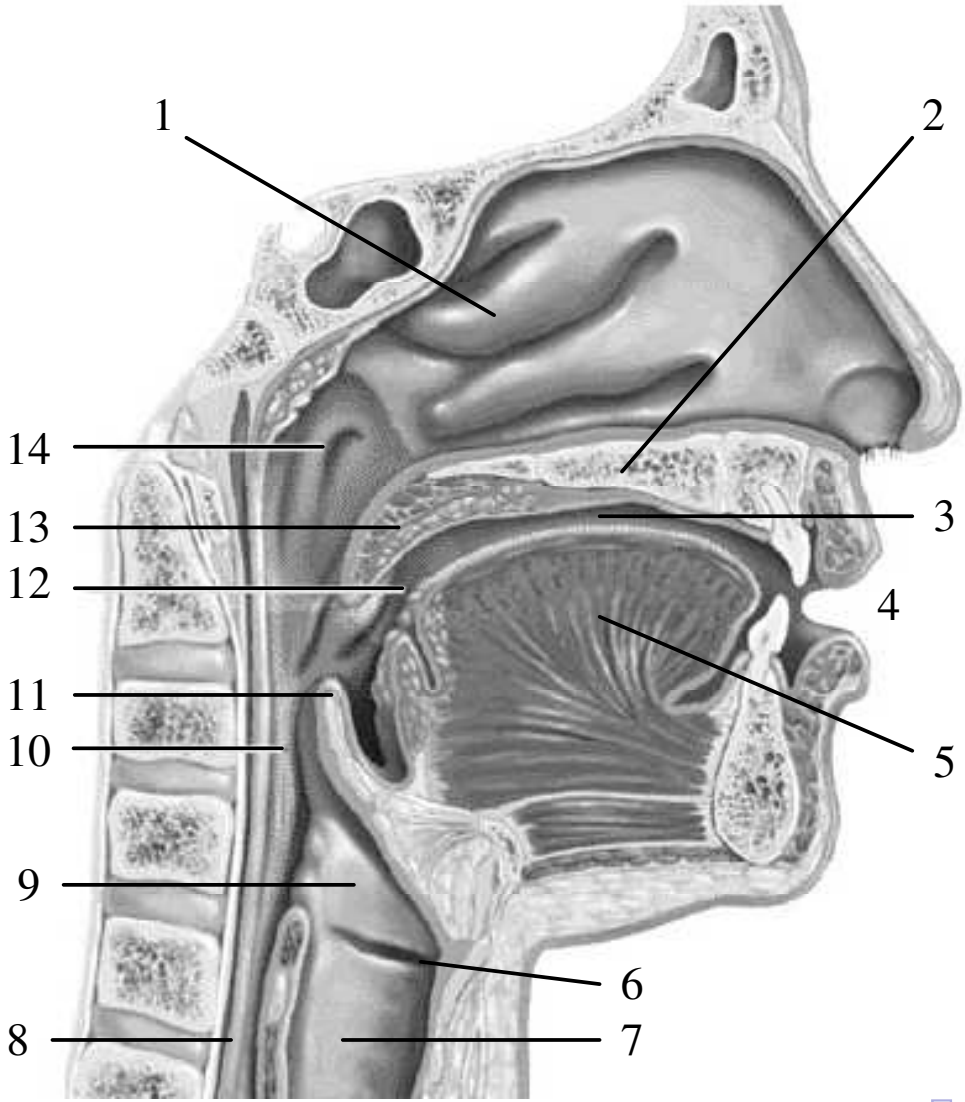
$$y(t) = x(t) * h(t)$$

- Sólo se conoce $y(t)$ y es de interés analizarla para estimar las características de la respuesta al impulso del tracto vocal $h(t)$.
- En el dominio frecuencial,

$$Y(f) = X(f)H(f)$$

donde $X(f)$ es el espectro de la excitación y $H(f)$ es la respuesta en frecuencias del tracto vocal.

Estructura anatómica del tracto vocal



Fuentes principales del sonido

Tipos de entrada

- Tren de pulsos cuasiperiódicos (sonidos sonoros)
  Frecuencia fundamental (F0)
- Ruido de banda ancha (sonidos sordos)

Modificadores del sonido

- Morfología del tracto vocal
- Circuito nasal
- Restricciones en el flujo de aire
- Radiación en los labios
- Posición de la lengua
- Posición de la mandíbula
- Sistema variante en el tiempo

Fuentes principales del sonido

Tipos de entrada

- Tren de pulsos cuasiperiódicos (sonidos sonoros)
  Frecuencia fundamental (F0)
- Ruido de banda ancha (sonidos sordos)

Modificadores del sonido

- Morfología del tracto vocal
- Circuito nasal
- Restricciones en el flujo de aire
- Radiación en los labios
- Posición de la lengua
- Posición de la mandíbula
- Sistema variante en el tiempo

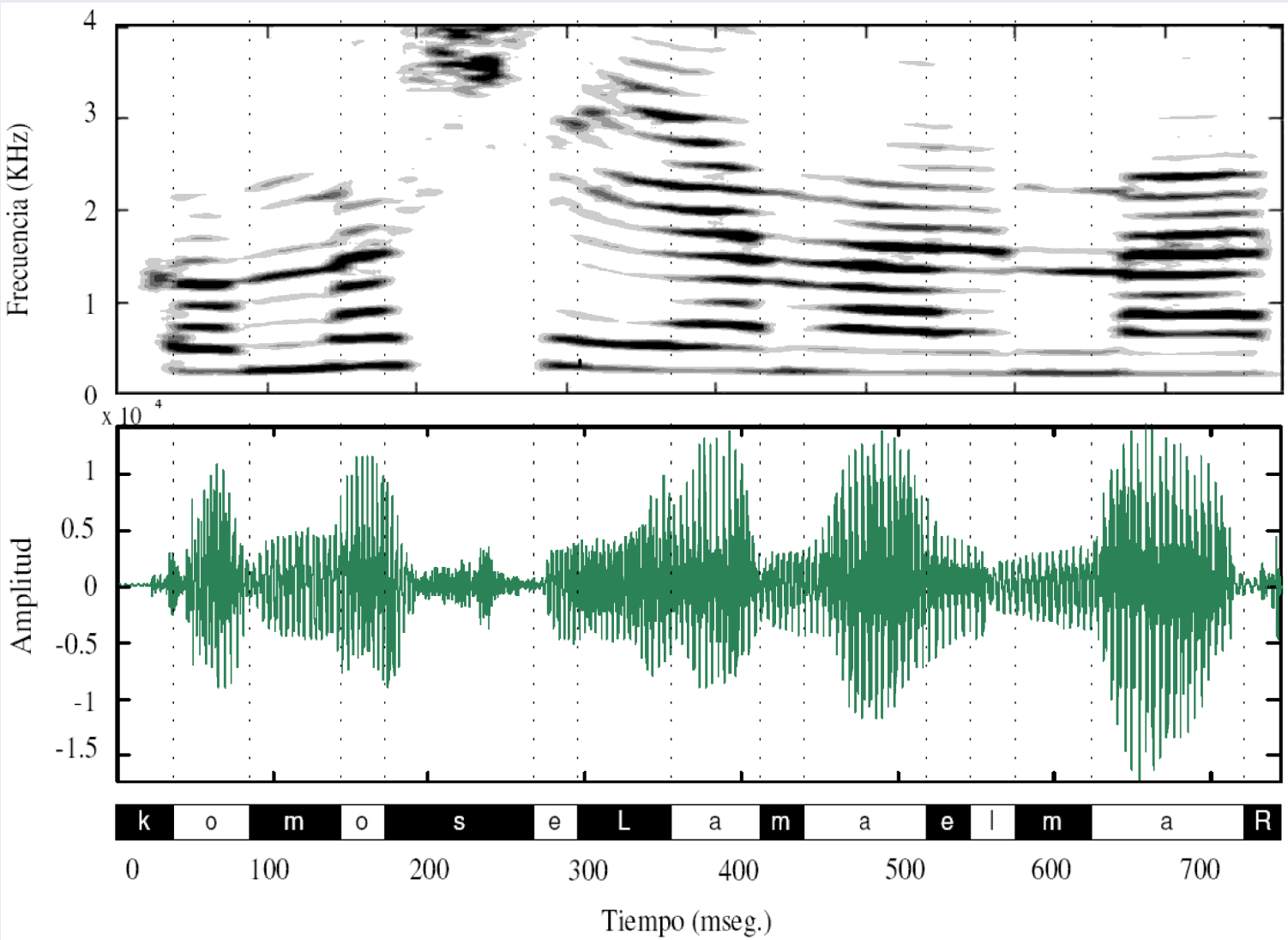
Análisis de la señal de voz

Período y Frecuencia fundamental (F_0) - Formantes

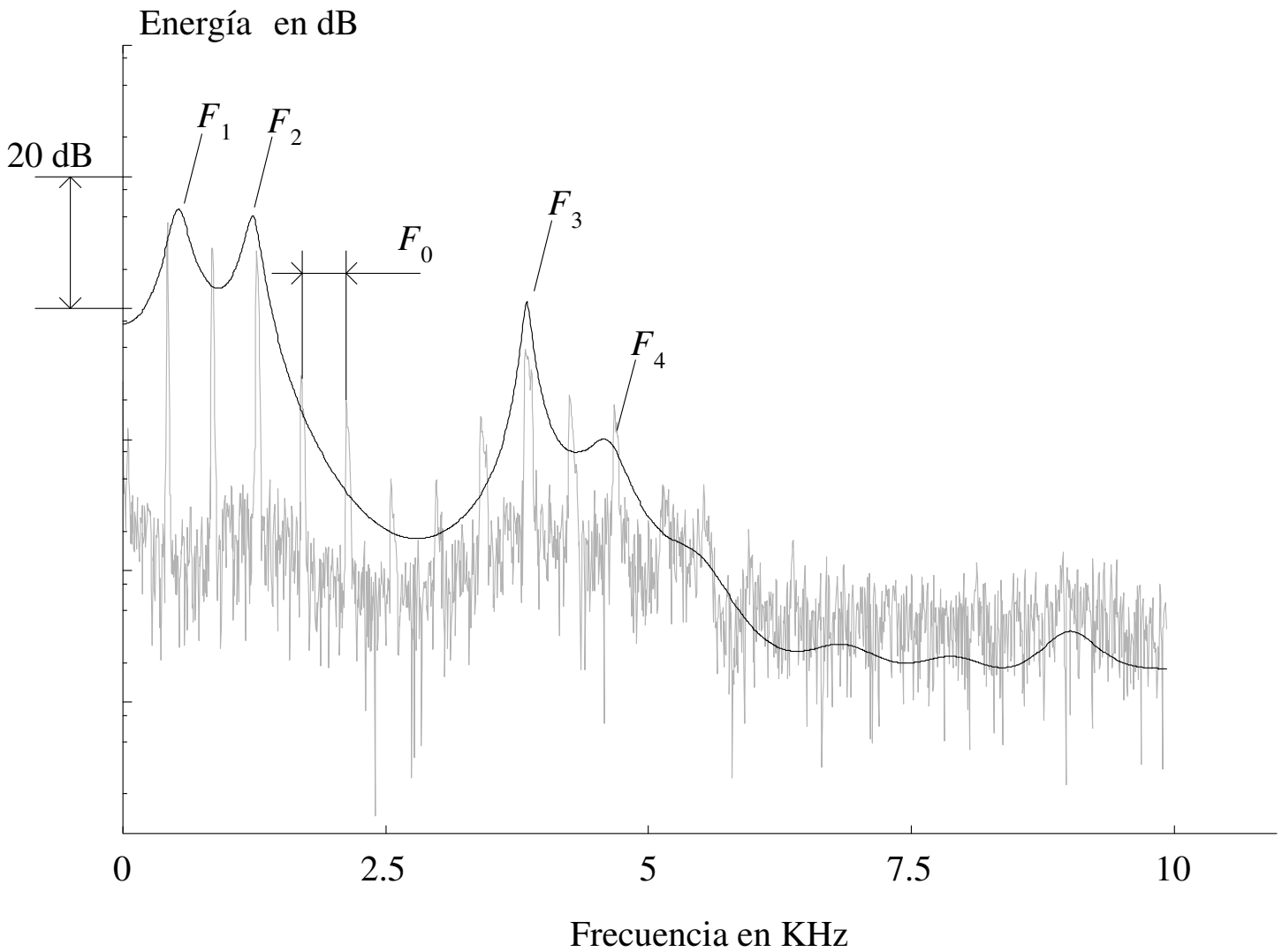
- La frecuencia fundamental F_0 corresponde a la frecuencia glótica, presente en los fonemas sonoros, y es una componente importante de la entonación en el habla.
- Período fundamental: $T_0 = \frac{1}{F_0}$
- Las frecuencias formantes $(F_1, F_2, F_3, \cdots)$ permiten discriminar entre las vocales.

Análisis de la señal de voz

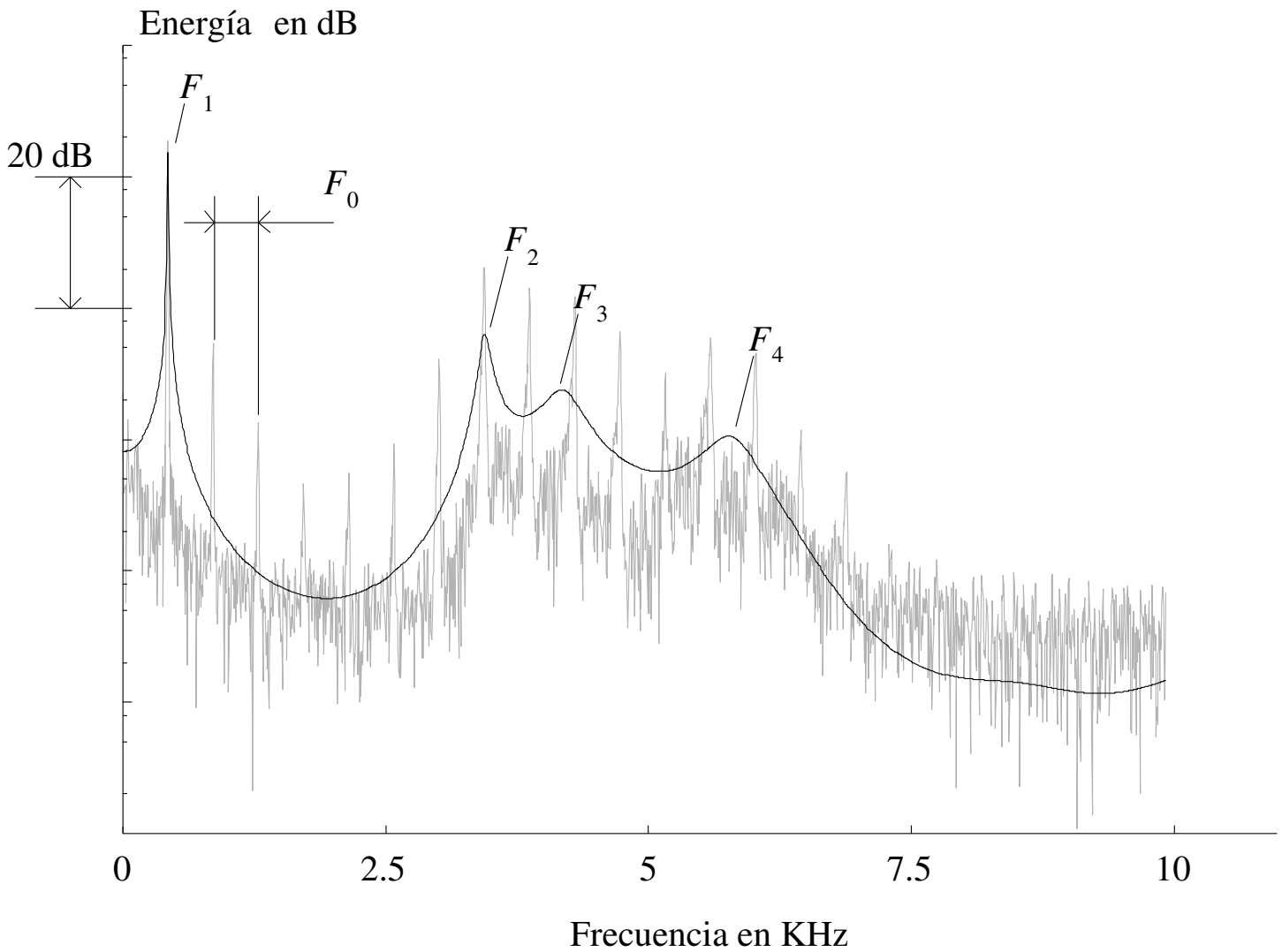
Sonograma y espectrograma



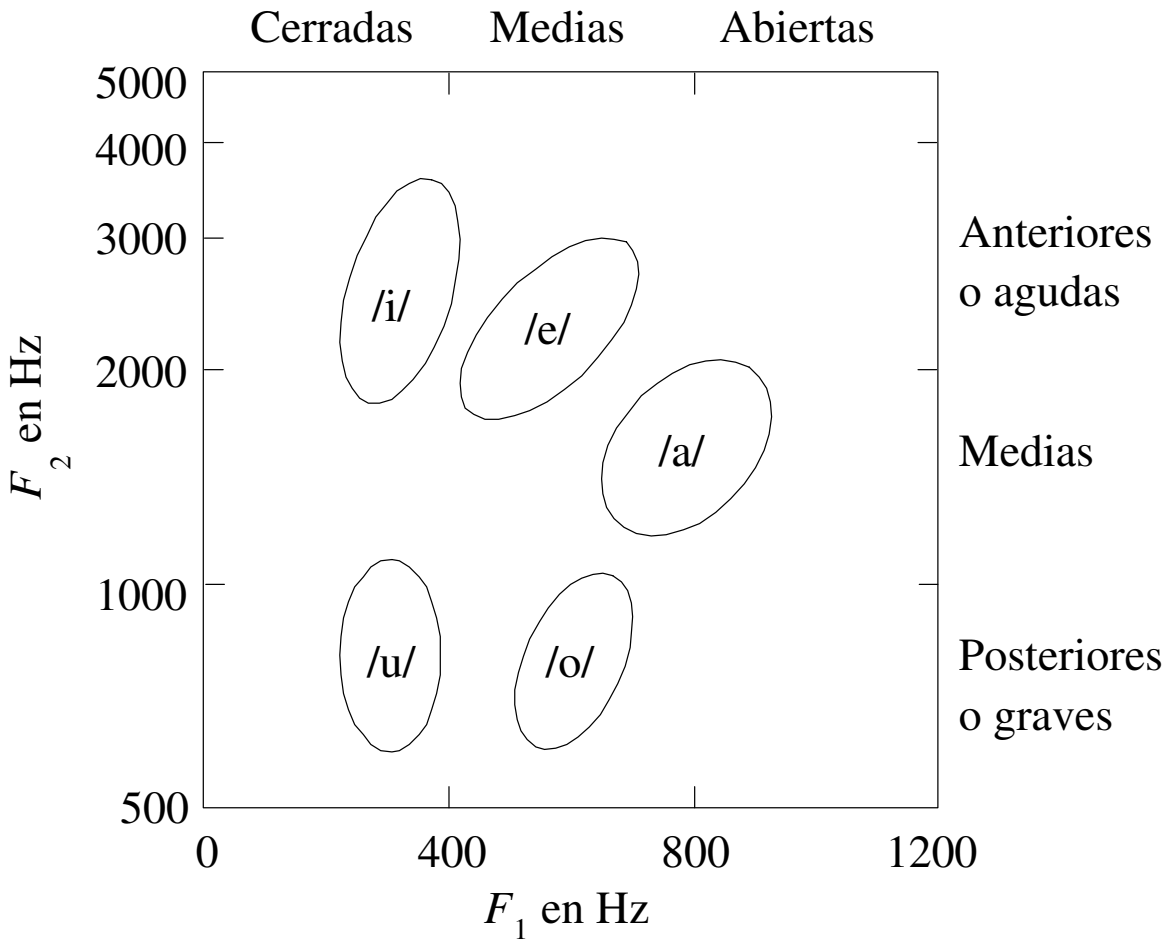
Espectro de una vocal



Espectro de una vocal

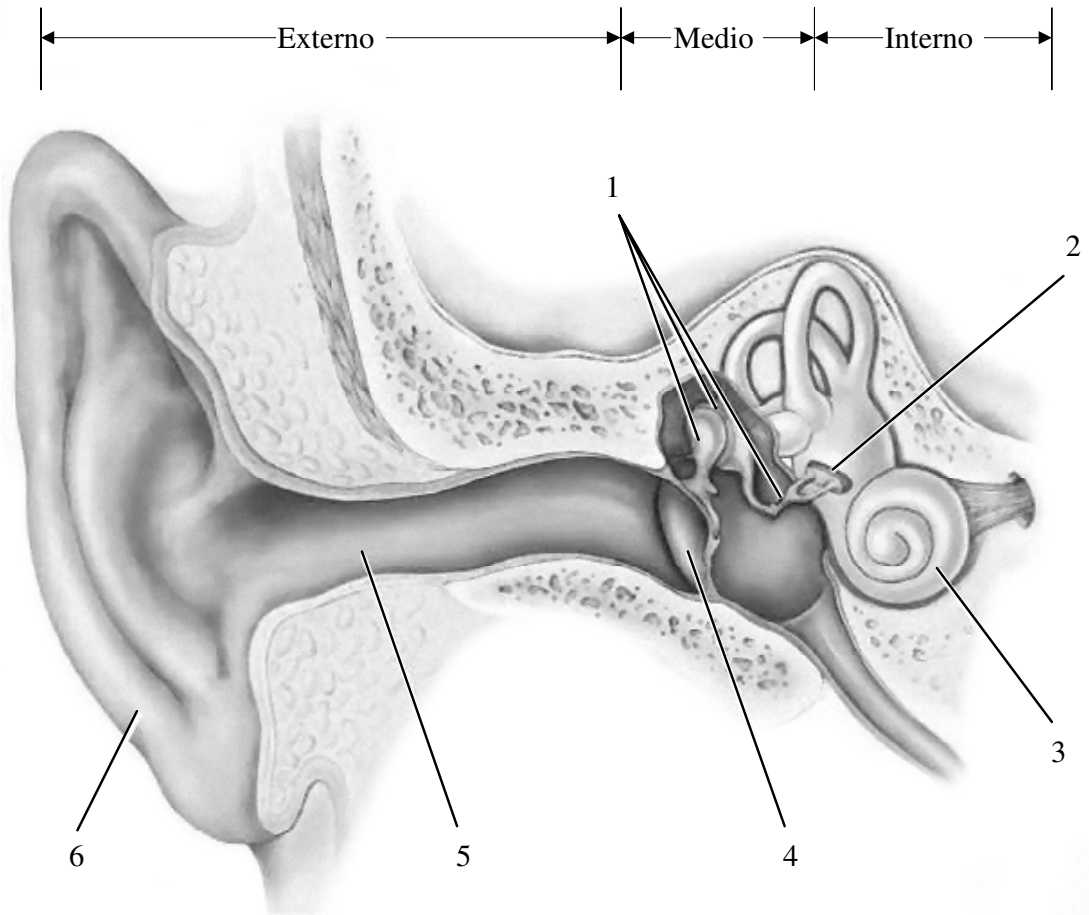


Triángulo de las vocales

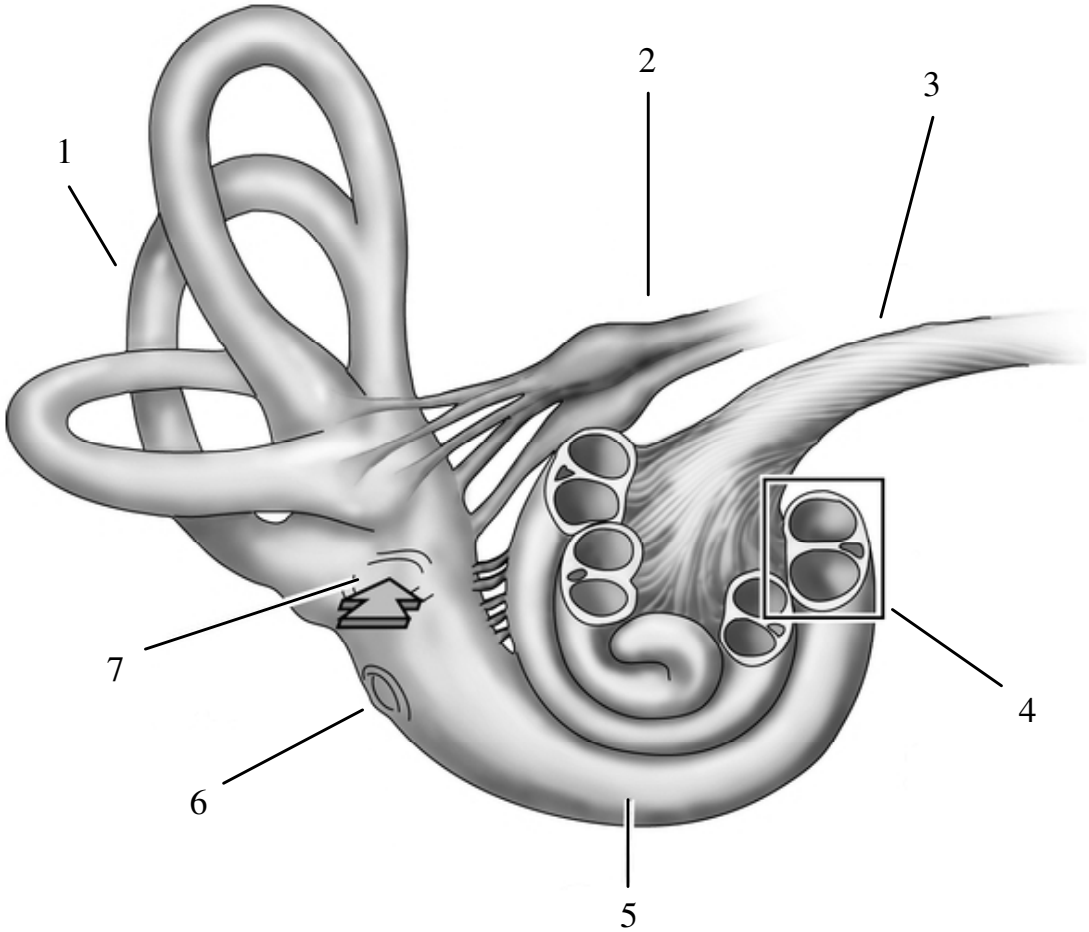


Percepción de la voz...

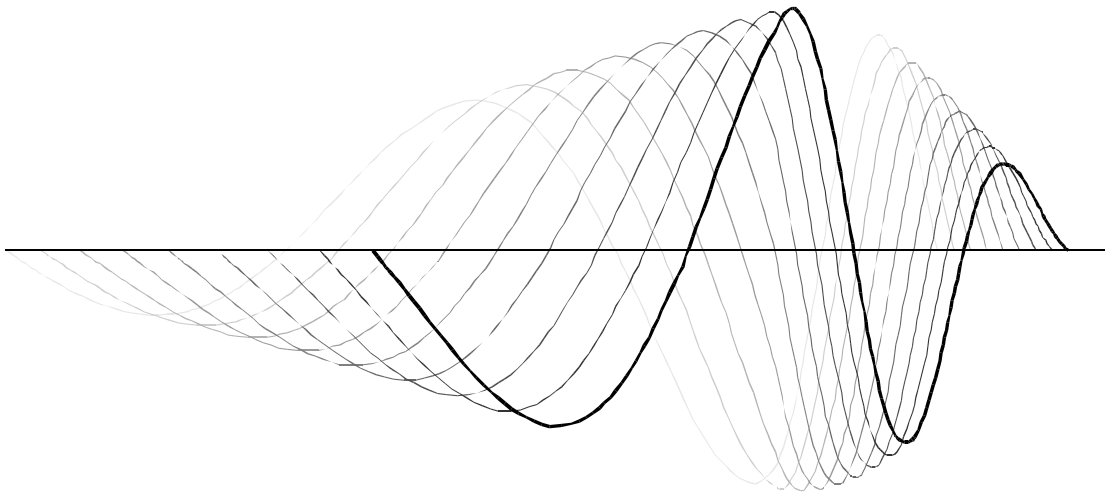
Partes del oído



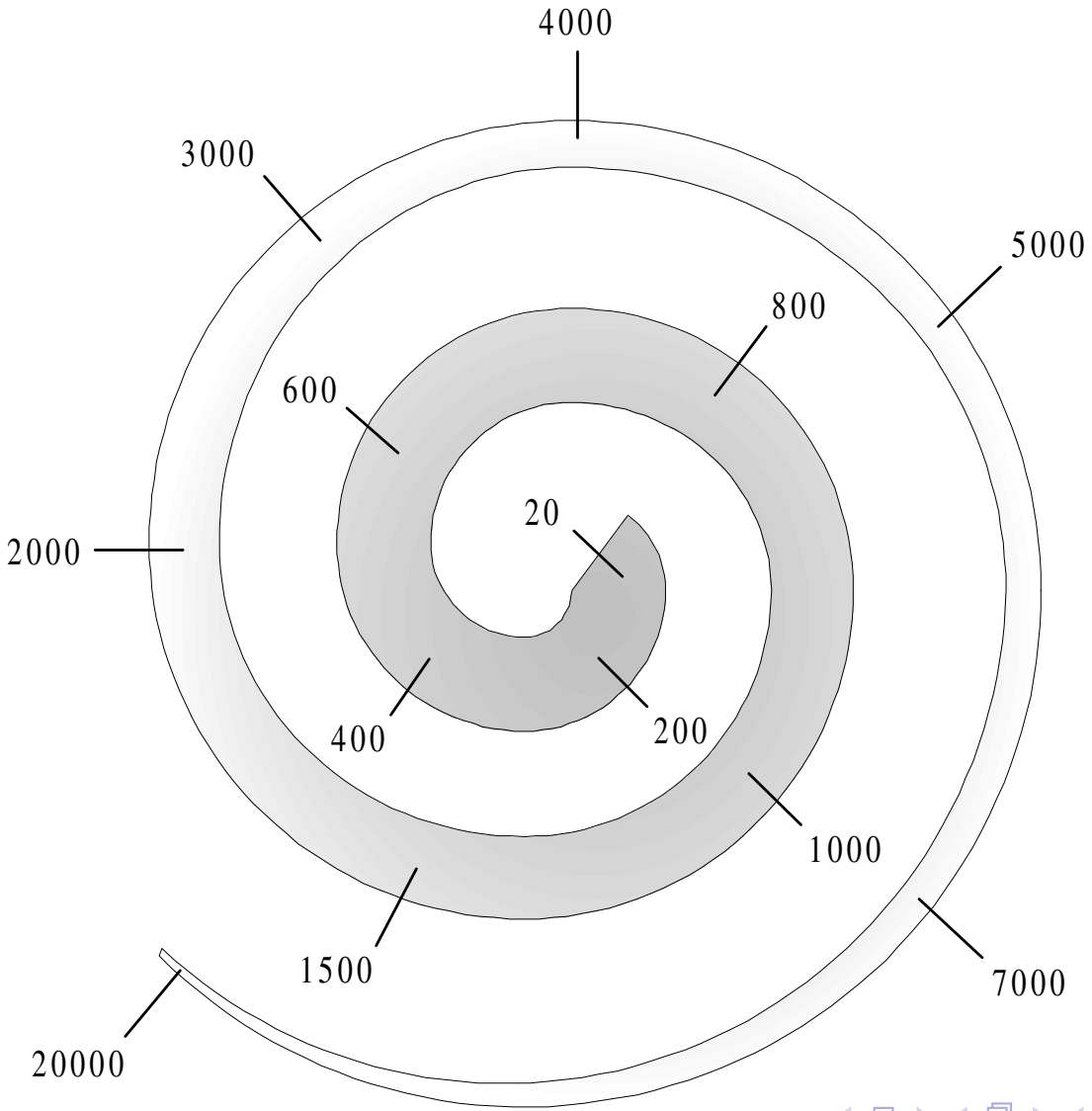
Cóclea



Onda viajera



Principio de tonotopía



Frecuencia y Pitch

F0 y Pitch

- A menudo confundidos en la literatura, el pitch no es igual a la frecuencia fundamental.
- La frecuencia, intensidad y las propiedades espectrales de un sonido interactúan en formas muy complejas para dar una percepción de pitch que puede ser un reflejo muy pobre de la F_0. El pitch percibido cambia con la intensidad.
- El pitch se refiere a un atributo perceptual del sonido, mientras que a frecuencia es un atributo físico de las señales.

Escala de mel

Mel

La unidad del pitch percibido de un tono puro es el mel.
No se corresponde linealmente con la frecuencia física del tono.
Stevens y Volkman (1940) establecieron: 1000 Hz $= 1000$ mel.

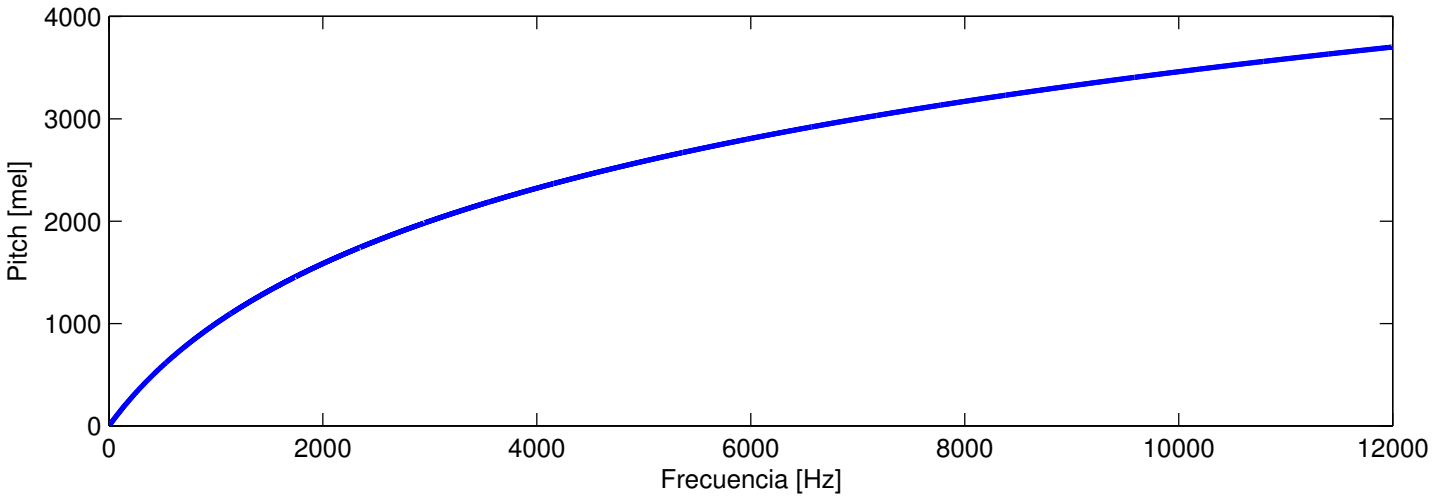
Escala de mel

$$F_{mel} = \frac{1000}{\log(2)} \log\left(1 + \frac{F_{Hz}}{1000}\right) \qquad \text{(Fant, 1973)}$$

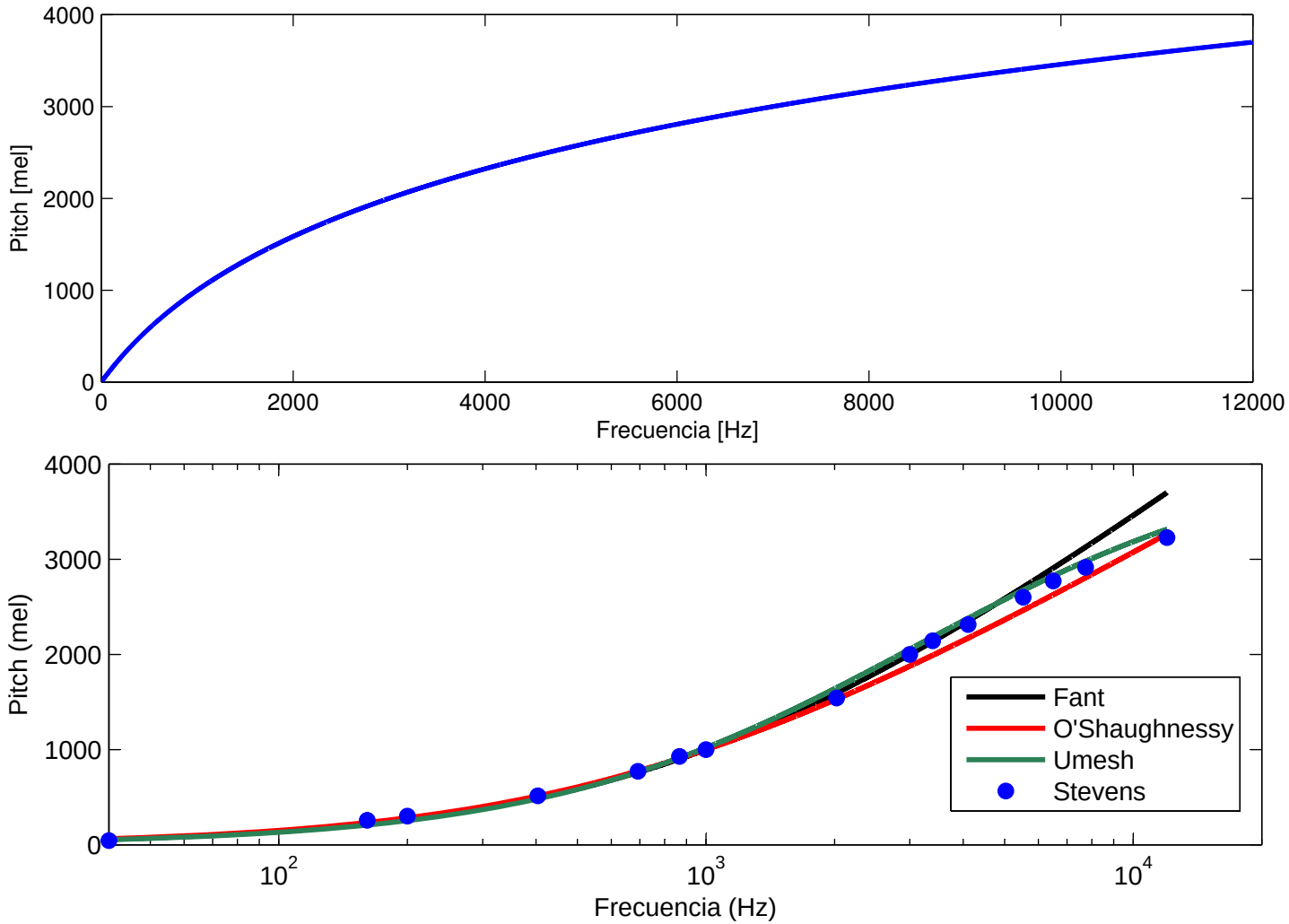
Es una aproximación y existen otras variantes

- O'Shaugnessy (1987)
- Umesh (1999)

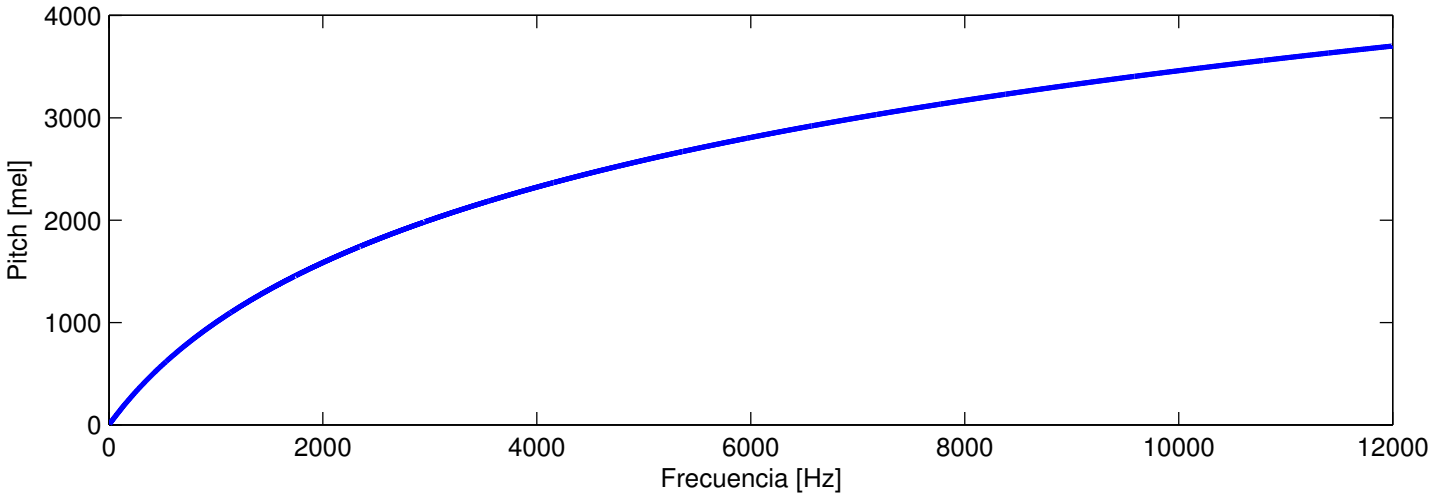
Banco de filtros en escala de mel

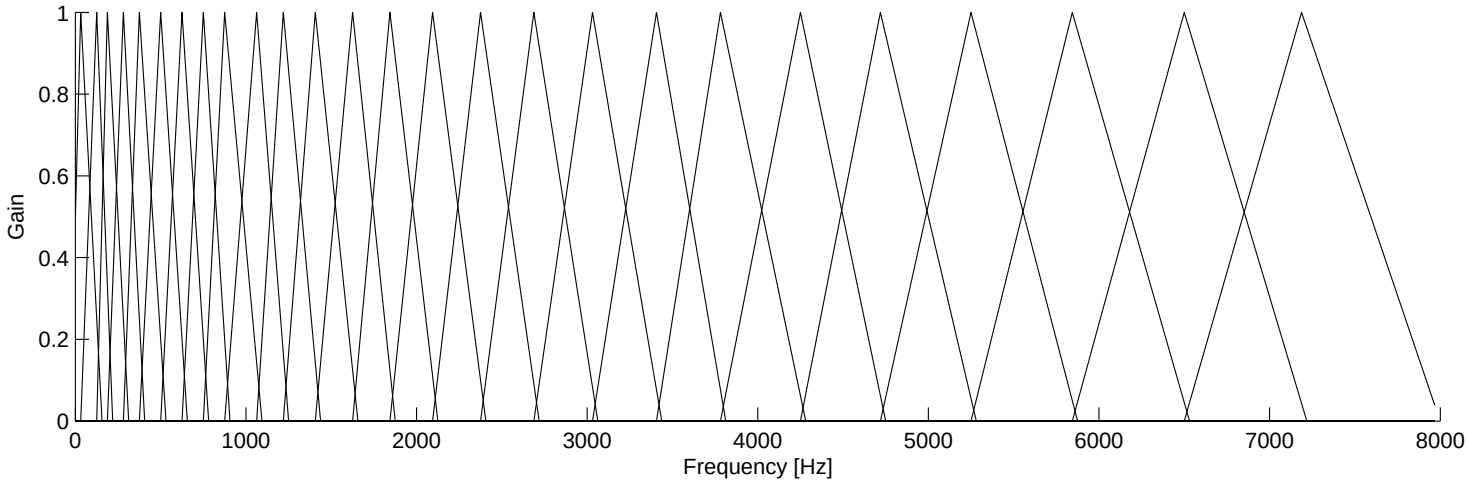


Banco de filtros en escala de mel



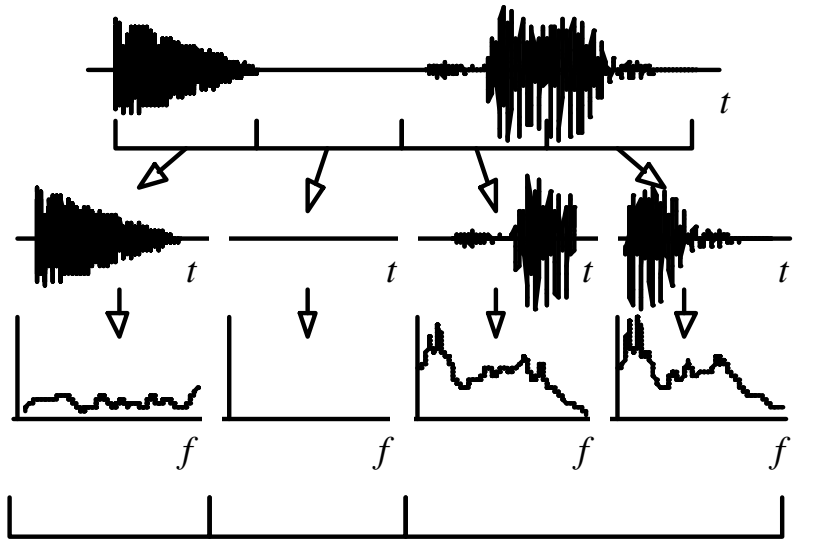
Banco de filtros en escala de mel





Organización de la clase

Primeros niveles estructurales del habla



ruidos silencios habla → Sonidos

/a/-/k/-/l/-/a/-/r/-/o/ **/k/-/e/** /u/-/N/ → Fonos y fonemas

/A/ /A/ /T/ /A/ /A/ → Suprasegmentos

/a/-/cla/-/ró/-/que/-/un/ → Sílabas

/Aclar/-/ó/-/que/-/un/ → Morfemas

/Aclaró/ - /que/ - /un/ - /Aleph/ → Palabras

/verbo/ - /nexo. inc./ - /numeral/ - /sust./ → Unidades sintácticas

Emisión completa

Tramos de análisis

Tramos procesados

Física acústica. Procesamiento de señales.

Fonética

Fonología

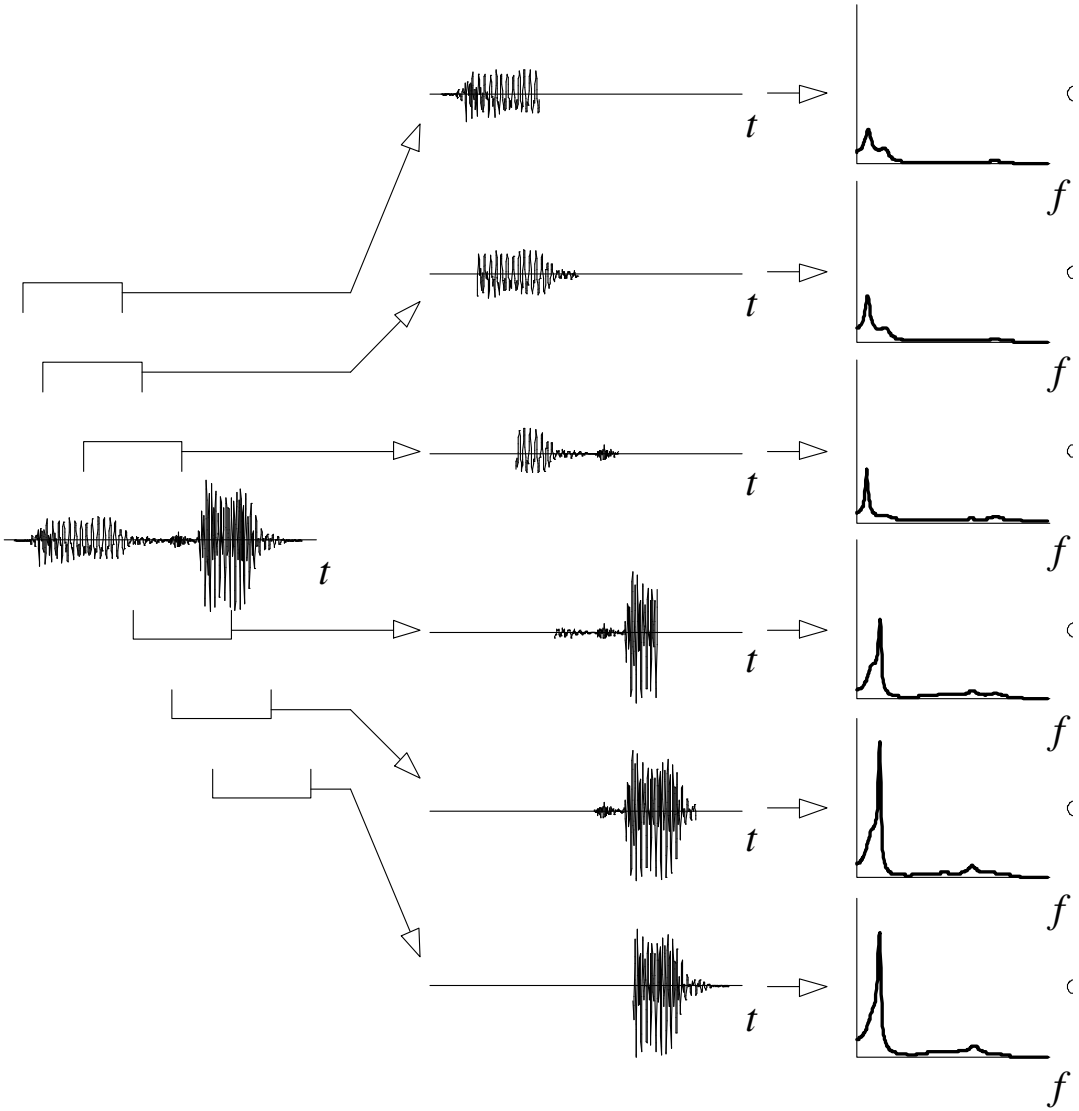
Morfología

Léxico

Sintaxis

Análisis por tramos

- Necesidad: señal no estacionaria
- Estacionariedad por tramos
- Tipos de ventanas (cuadrada, Hamming, etc.)
- Técnicas de ventaneo
- Solapado en el tiempo
- Análisis de las ventanas independientes


t
f
t
f
t
f
t
f
t
f
t
f
t
f

Análisis por tramos

Ventaneo

$$v(t,n) = \omega(n, N_\omega) x(tN_d + n), \quad 0 < n \leq N_\omega$$

t: índice de la ventana

n: índice de la muestra

Hamming

$$\omega_H(m, N_\omega) = \frac{27}{50} - \frac{23}{50}\cos(2\pi m / N_\omega)$$

Transformaciones de dominio sobre tramos individuales

$$V(t,k) = \mathcal{T}(k)\{v(t,n)\}, \quad 0 < k \leq N_x$$

- CE: $\mathbf{u}_t \leftarrow u(t,k) = \mathcal{T}_F(k)\{v(t,n)\}$
- CPL: $\mathbf{a}_t \leftarrow a(t,k) = \mathcal{T}_L(k)\{v(t,n)\}$

Análisis por tramos

Ventaneo

$$v(t,n) = \omega(n, N_\omega) x(tN_d + n), \quad 0 < n \leq N_\omega$$

t: índice de la ventana

n: índice de la muestra

Hamming

$$\omega_H(m, N_\omega) = \frac{27}{50} - \frac{23}{50} \cos(2\pi m / N_\omega)$$

Transformaciones de dominio sobre tramos individuales

$$V(t,k) = \mathcal{T}(k)\{v(t,n)\}, \quad 0 < k \leq N_x$$

- CE: $\mathbf{u}_t \leftarrow u(t,k) = \mathcal{T}_F(k)\{v(t,n)\}$
- CPL: $\mathbf{a}_t \leftarrow a(t,k) = \mathcal{T}_L(k)\{v(t,n)\}$

Análisis por tramos

Ventaneo

$$v(t,n) = \omega(n, N_\omega) x(tN_d + n), \quad 0 < n \leq N_\omega$$

t: índice de la ventana

n: índice de la muestra

Hamming

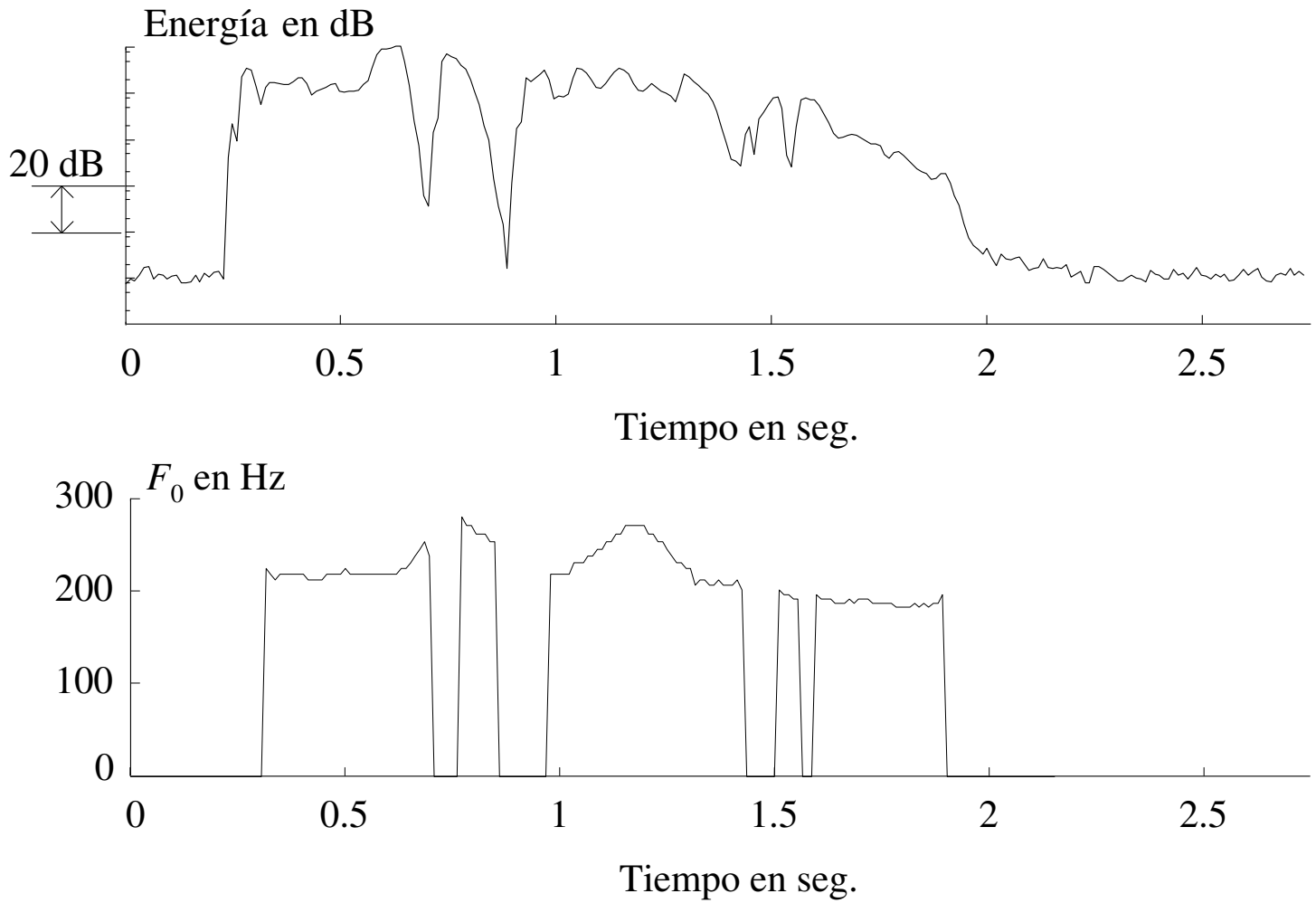
$$\omega_H(m, N_\omega) = \frac{27}{50} - \frac{23}{50}\cos(2\pi m / N_\omega)$$

Transformaciones de dominio sobre tramos individuales

$$V(t,k) = \mathcal{T}(k)\left\{v(t,n)\right\}, \quad 0 < k \leq N_x$$

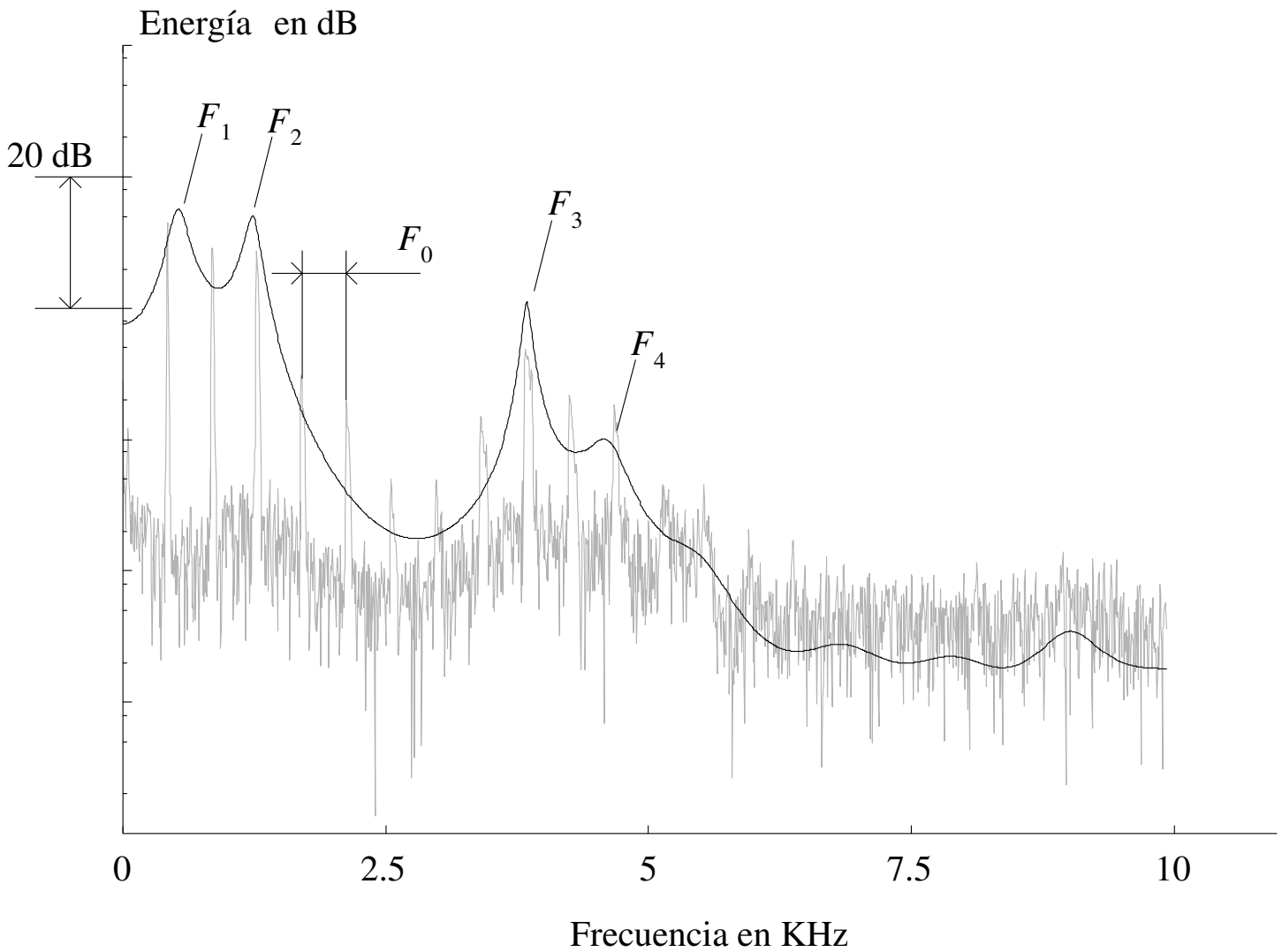
- CE: $\mathbf{u}_t \leftarrow u(t,k) = \mathcal{T}_F(k)\left\{v(t,n)\right\}$
- CPL: $\mathbf{a}_t \leftarrow a(t,k) = \mathcal{T}_L(k)\left\{v(t,n)\right\}$

Energía y entonación (F0) por tramos

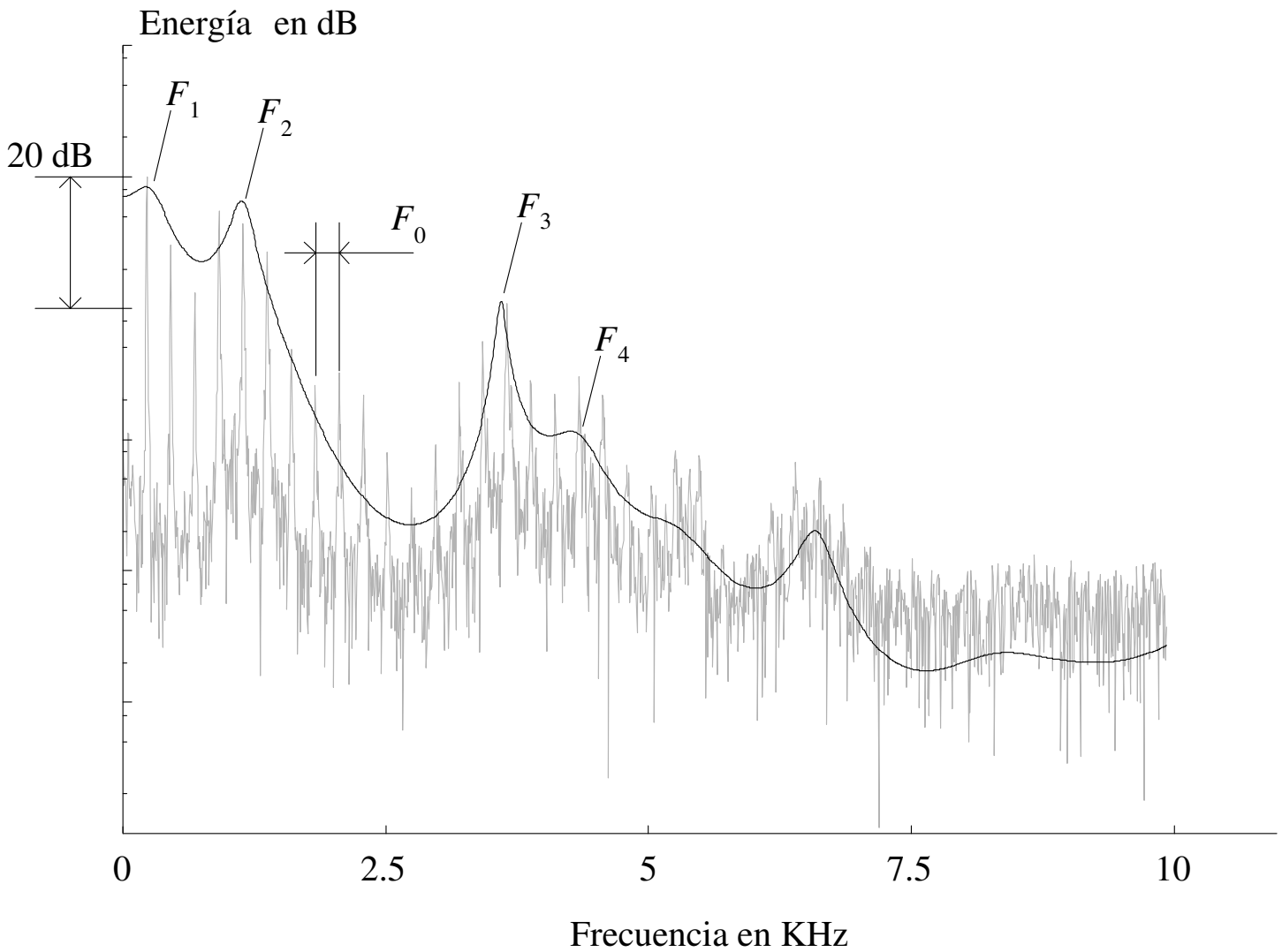


Organización de la clase

Espectro de una vocal



Otra elocución de la misma vocal



Coeficientes cepstrales

$$c(m) = \mathcal{T}_F^{-1}\left\{\log\left|\mathcal{T}_F\left\{v(m)\right\}\right|\right\}$$

- Espectral $\rightarrow$ Cepstral
- Espectro $\rightarrow$ Cepstro
- Frcuencias $\rightarrow$ Cuefrencias
- Filtro, filtrado $\rightarrow$ Liftro, liftrado
- Armónicas $\rightarrow$ Ramónicas

Coeficientes cepstrales

$$c(m) = \mathcal{T}_F^{-1} \{\log |\mathcal{T}_F \{v(m)\}|\}$$

Espectral $\rightarrow$ Cepstral

Espectro $\rightarrow$ Cepstro

Frcuencias $\rightarrow$ Cuefrencias

Filtro, filtrado $\rightarrow$ Liftro, liftrado

Armónicas $\rightarrow$ Ramónicas

Separación de fuentes y modificadores del sonido

$$\hat{v}(n) = g(n) * h(n)$$

$$\hat{V}(k) = G(k) \times H(k)$$

$$\hat{\log}\,|V(k)| = \log|G(k) \times H(k)|$$

$$\hat{\log}\,|V(k)| = \log|G(k)| + \log|H(k)|$$

$$\hat{v}(m) = \mathcal{T}_F^{-1}\{\log|G(k)|\} + \mathcal{T}_F^{-1}\{\log|H(k)|\}$$

Separación de fuentes y modificadores del sonido

$$\hat{v}(n) = g(n) * h(n)$$

$$\hat{V}(k) = G(k) \times H(k)$$

$$\hat{\log}\,|V(k)| = \log|G(k) \times H(k)|$$

$$\hat{\log}\,|V(k)| = \log|G(k)| + \log|H(k)|$$

$$\hat{v}(m) = \mathcal{T}_F^{-1}\{\log|G(k)|\} + \mathcal{T}_F^{-1}\{\log|H(k)|\}$$

Separación de fuentes y modificadores del sonido

$$\hat{v}(n) = g(n) * h(n)$$

$$\hat{V}(k) = G(k) \times H(k)$$

$$\hat{\log}\,|V(k)| = \log|G(k) \times H(k)|$$

$$\hat{\log}\,|V(k)| = \log|G(k)| + \log|H(k)|$$

$$\hat{v}(m) = \mathcal{T}_F^{-1}\{\log|G(k)|\} + \mathcal{T}_F^{-1}\{\log|H(k)|\}$$

Separación de fuentes y modificadores del sonido

$$\hat{v}(n) = g(n) * h(n)$$

$$\hat{V}(k) = G(k) \times H(k)$$

$$\hat{\log}\,|V(k)| = \log|G(k) \times H(k)|$$

$$\hat{\log}\,|V(k)| = \log|G(k)| + \log|H(k)|$$

$$\hat{v}(m) = \mathcal{T}_F^{-1}\{\log|G(k)|\} + \mathcal{T}_F^{-1}\{\log|H(k)|\}$$

Separación de fuentes y modificadores del sonido

$$\hat{v}(n) = g(n) * h(n)$$

$$\hat{V}(k) = G(k) \times H(k)$$

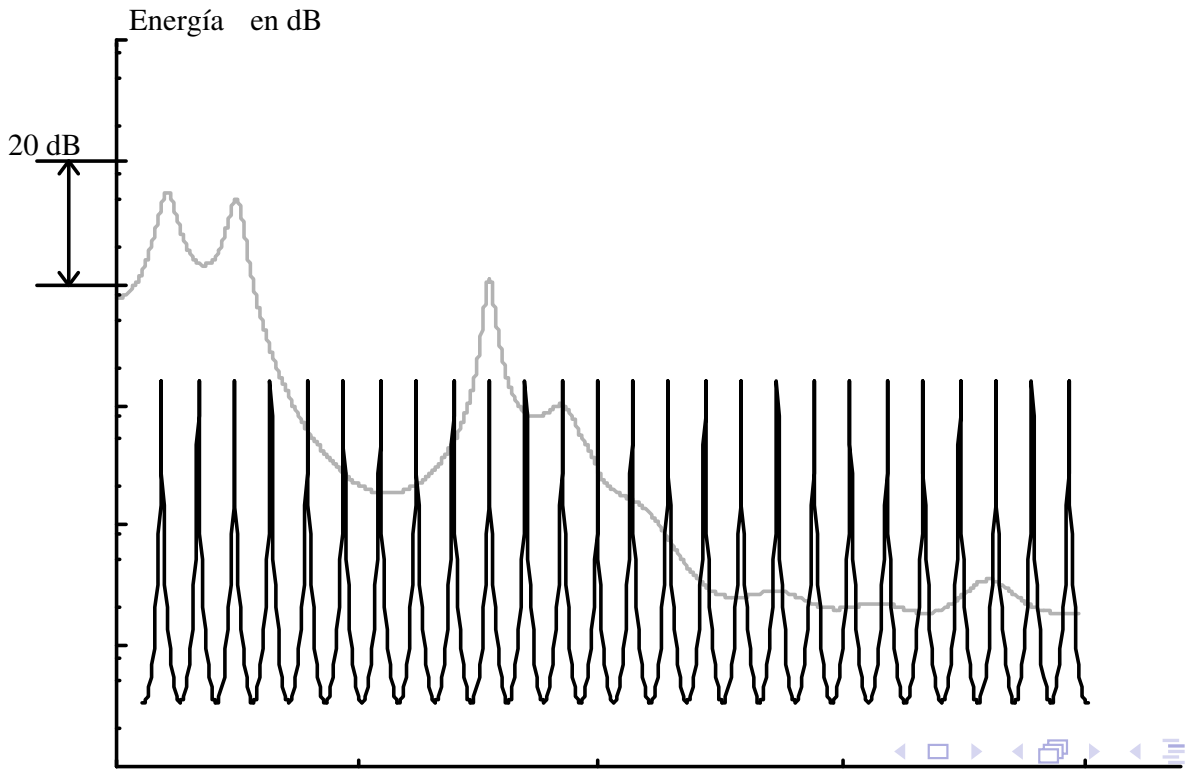
$$\hat{\log}\,|V(k)| = \log|G(k) \times H(k)|$$

$$\hat{\log}\,|V(k)| = \log|G(k)| + \log|H(k)|$$

$$\hat{v}(m) = \mathcal{T}_F^{-1}\{\log|G(k)|\} + \mathcal{T}_F^{-1}\{\log|H(k)|\}$$

Separación de fuentes y modificadores del sonido

$$\hat{v}(m) = \mathcal{T}_F^{-1}\left\{\log|G(k)|\right\} + \mathcal{T}_F^{-1}\left\{\log|H(k)|\right\}$$

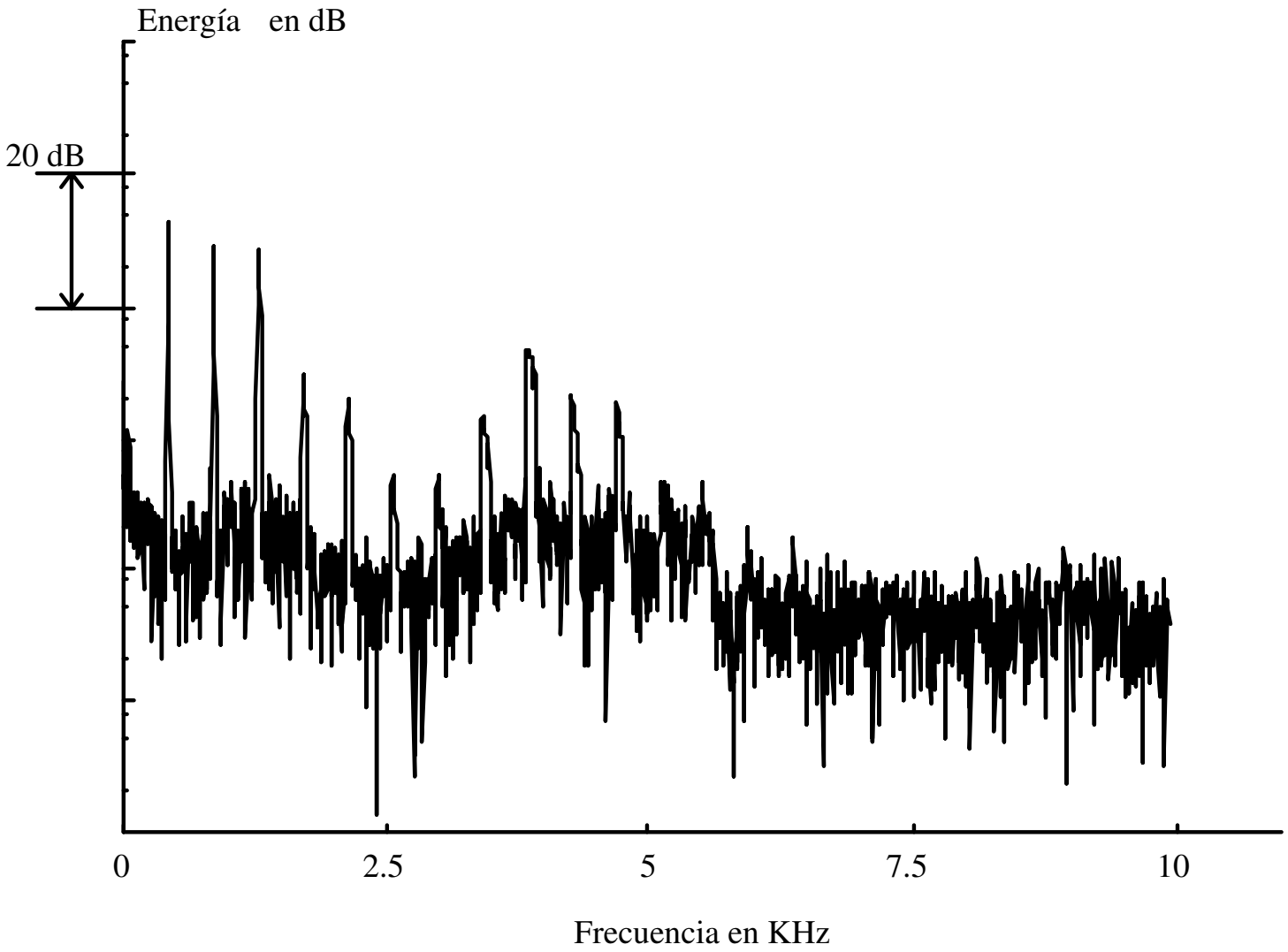


Separación de fuentes y modificadores del sonido

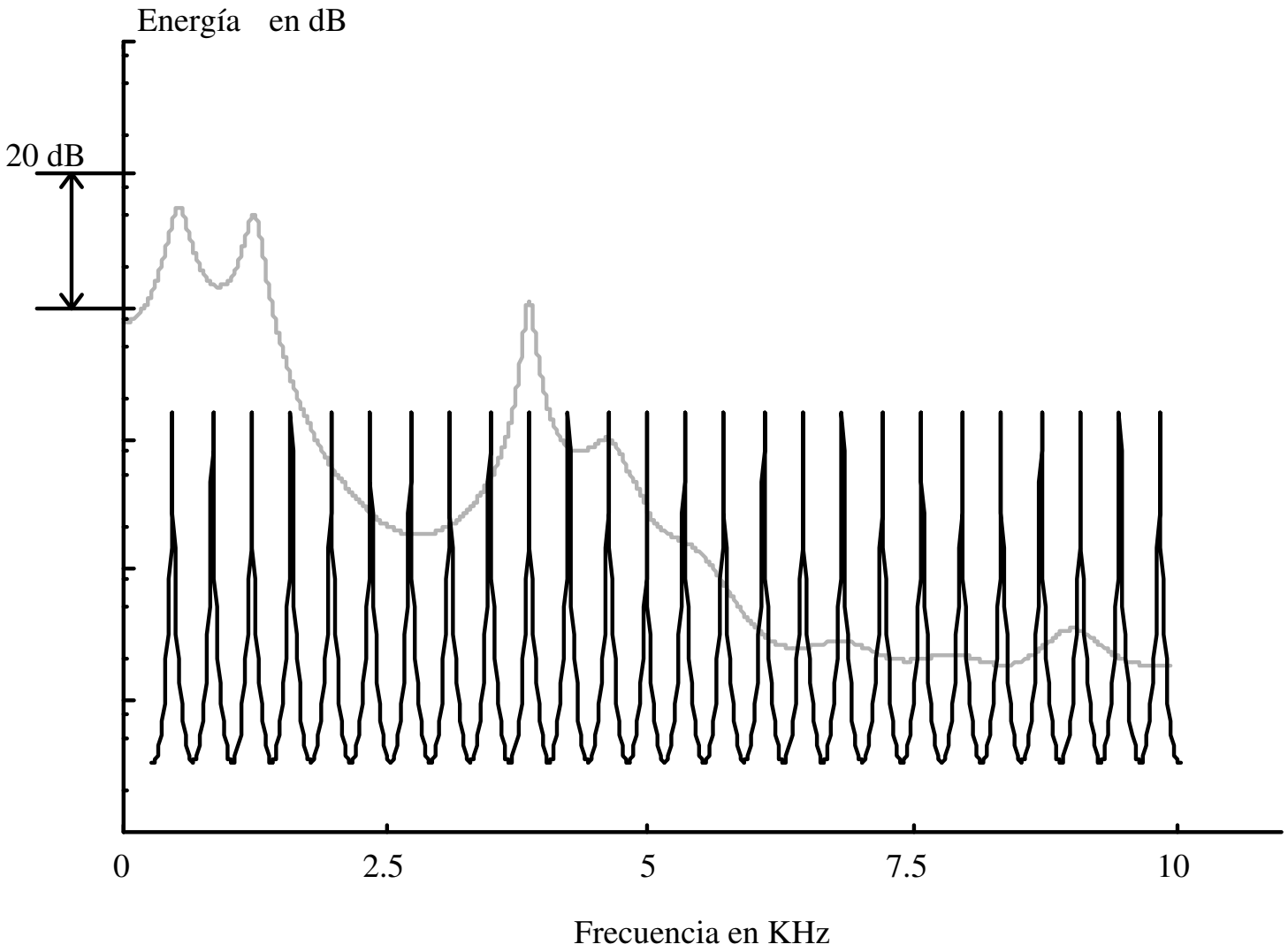
$$\hat{v}(m) = \mathcal{T}_F^{-1}\{\log|G(k)|\} + \mathcal{T}_F^{-1}\{\log|H(k)|\}$$

G y H ocupan partes diferentes del eje de cuefrencias. Podemos separar la parte que varía rápidamente (correspondiente a la excitación del tracto vocal) de la que varía lentamente (la respuesta en frecuencia del tracto).

Fuentes y modificadores de sonido en el espectro

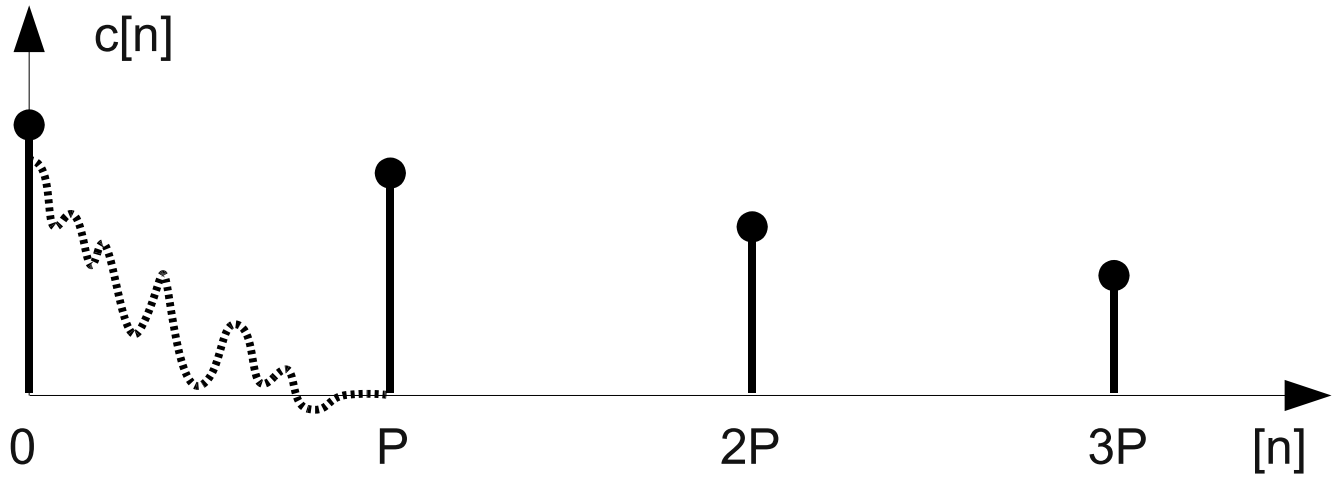


Fuentes y modificadores de sonido en el espectro

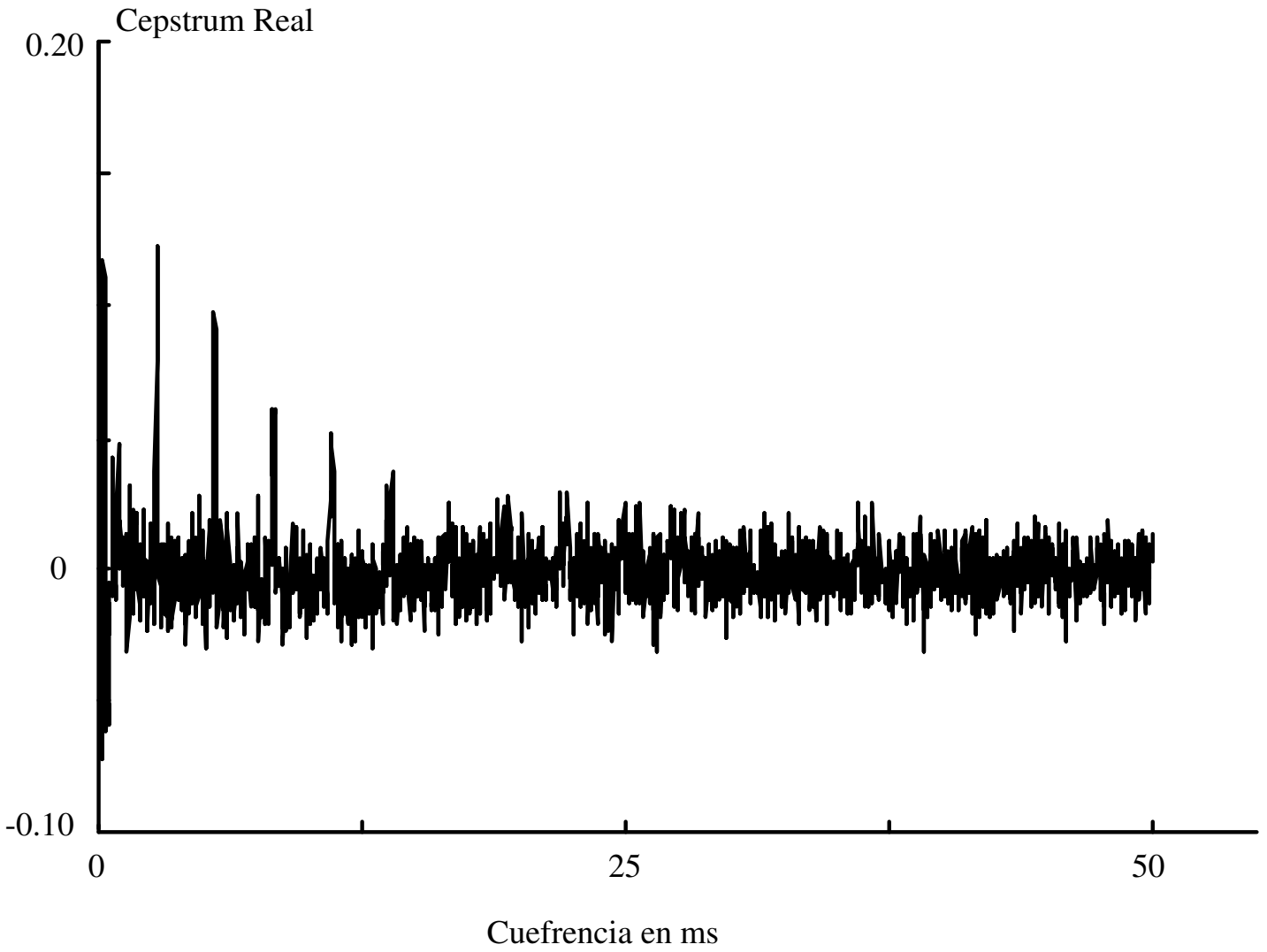


Cepstrum de una vocal

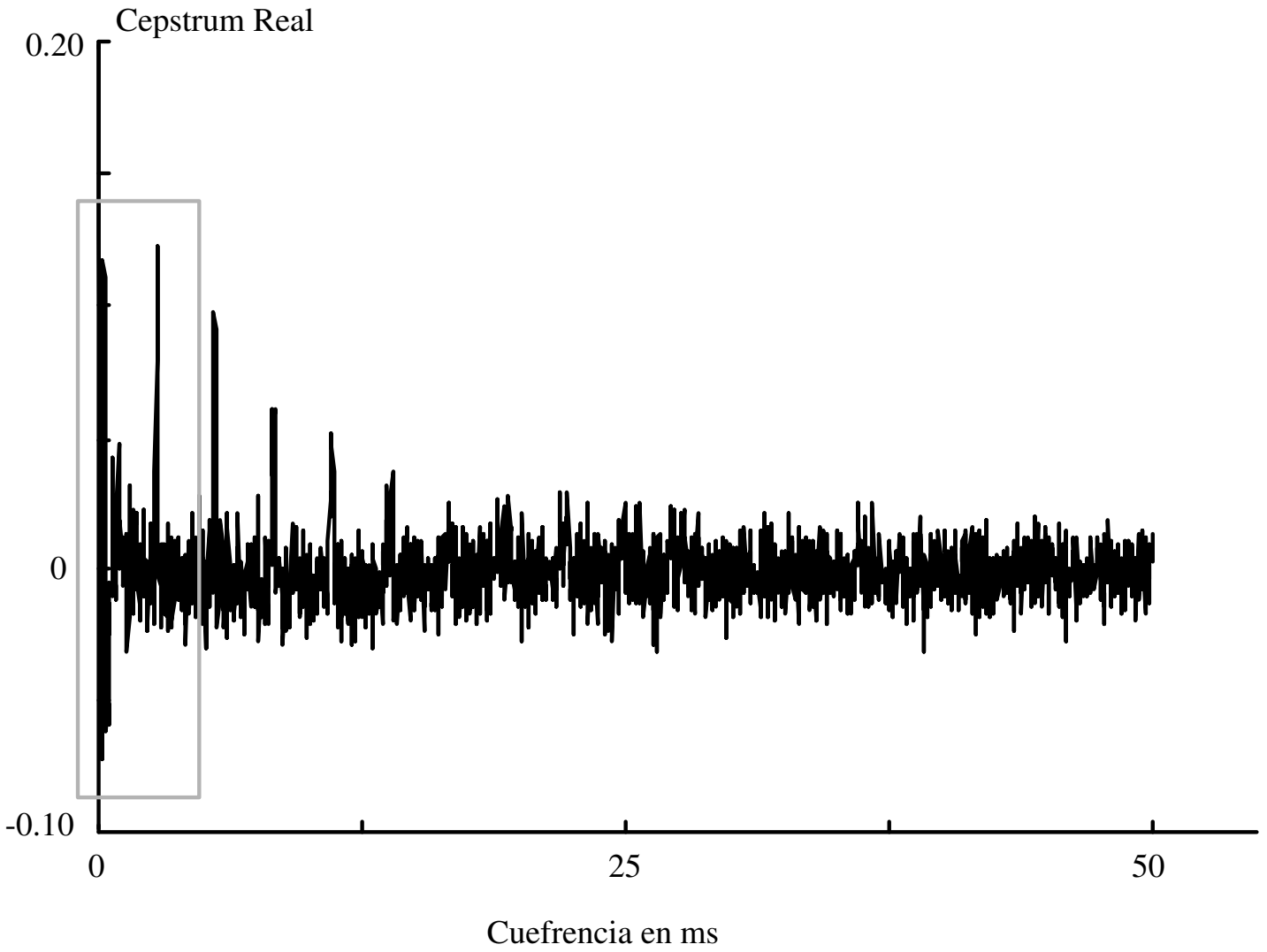
(esquema representativo)



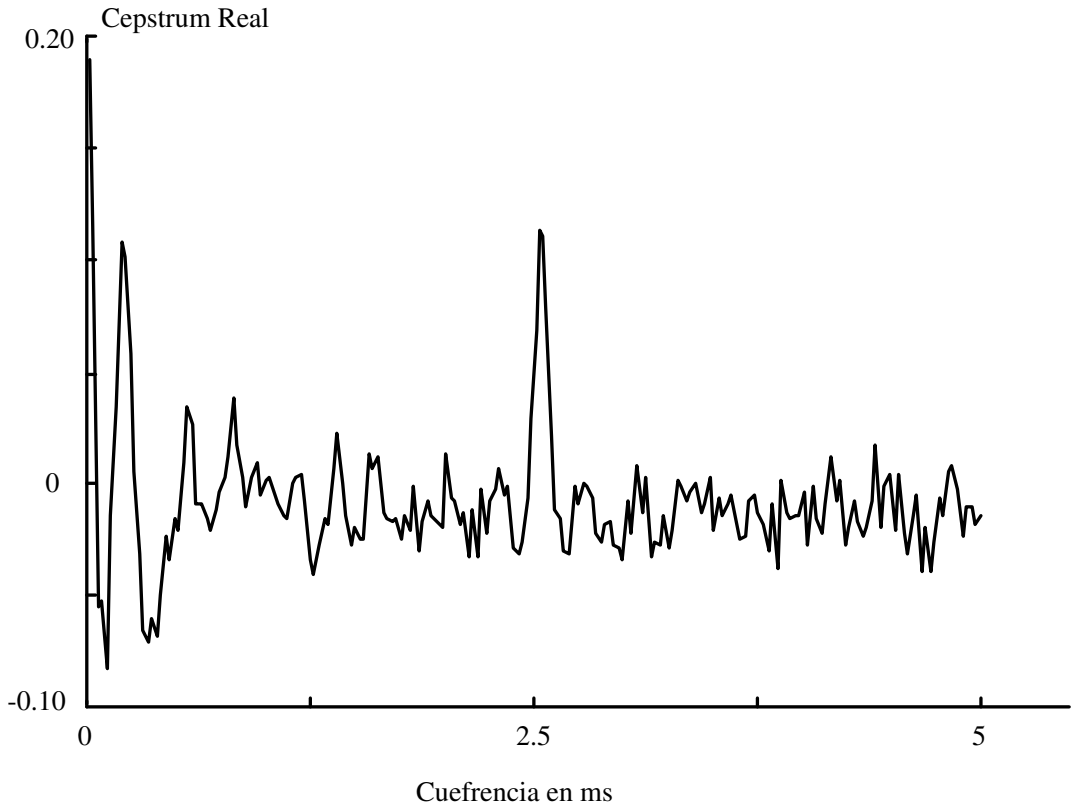
Cepstrum de una vocal



Cepstrum de una vocal



Cepstrum de una vocal



- Detección del pico que determina el período findamental T_0
- Rango de posible de F_0 (100 - 300 Hz) $\rightarrow$ rango posible de T_0

Coeficientes cepstrales en escala de mel

Permiten obtener una representación de la señal de voz emulando el análsis frecuencial que realiza el sistema auditivo.

- Banco de filtros en escala de mel
- Integración por bandas del espectro
- Coeficientes de energía por cada banda
- Transformación inversa

Coeficientes cepstrales en escala de mel

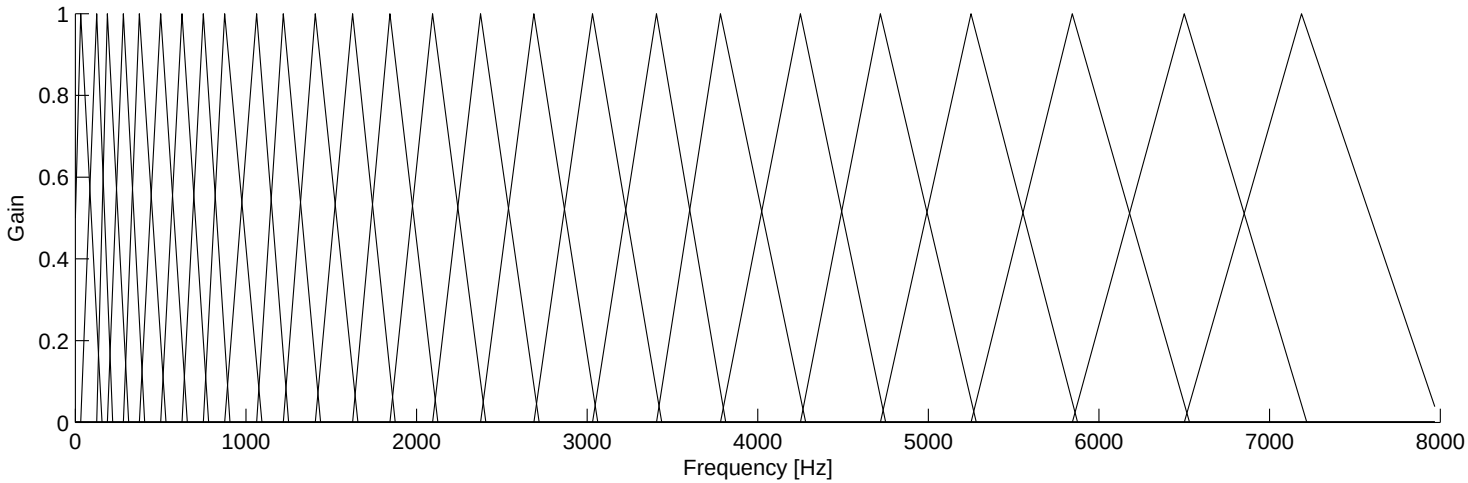
Permiten obtener una representación de la señal de voz emulando el análsis frecuencial que realiza el sistema auditivo.

- Banco de filtros en escala de mel
- Integración por bandas del espectro
- Coeficientes de energía por cada banda
- Transformación inversa

Coeficientes cepstrales en escala de mel

Escala de mel

$$F_{mel} = \frac{1000}{\log(2)} \log\left(1 + \frac{F_{Hz}}{1000}\right)$$



Coeficientes cepstrales en escala de mel

El espectro de magnitud logarítmico

$$X[k] = \log_e |TDF\{x[n]\}|,$$

es integrado en bandas usando filtros W_i, $i = 1...I$

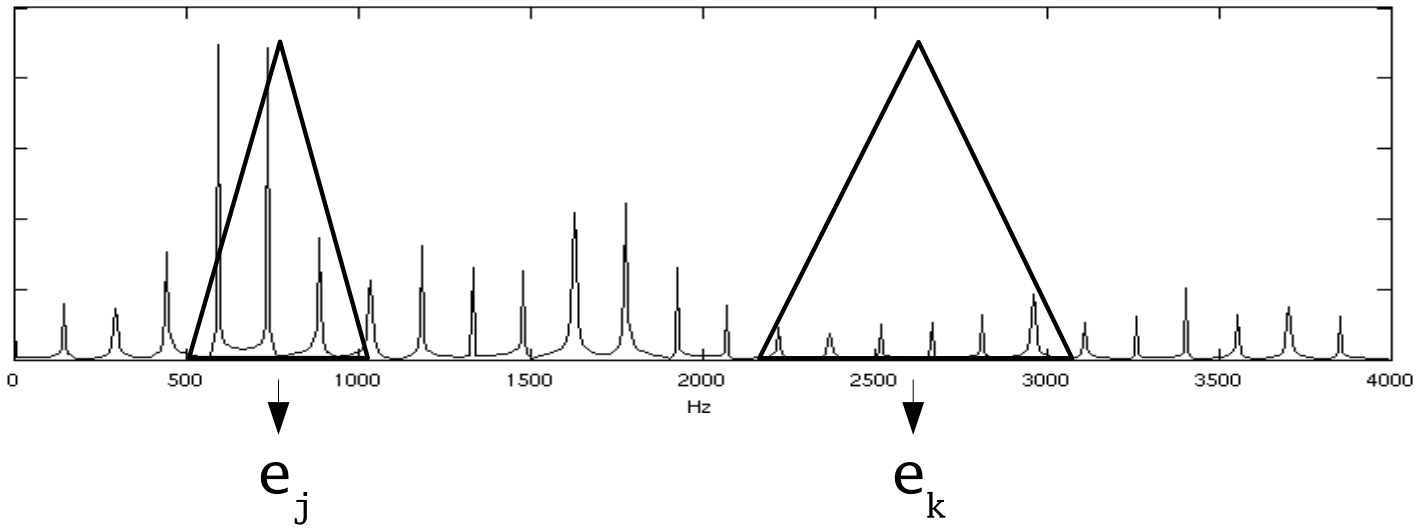
$$U[i] = \sum_k W_i[k]X[k],$$

y luego se calcula la transformada inversa

$$C = TDFI\{U\}.$$

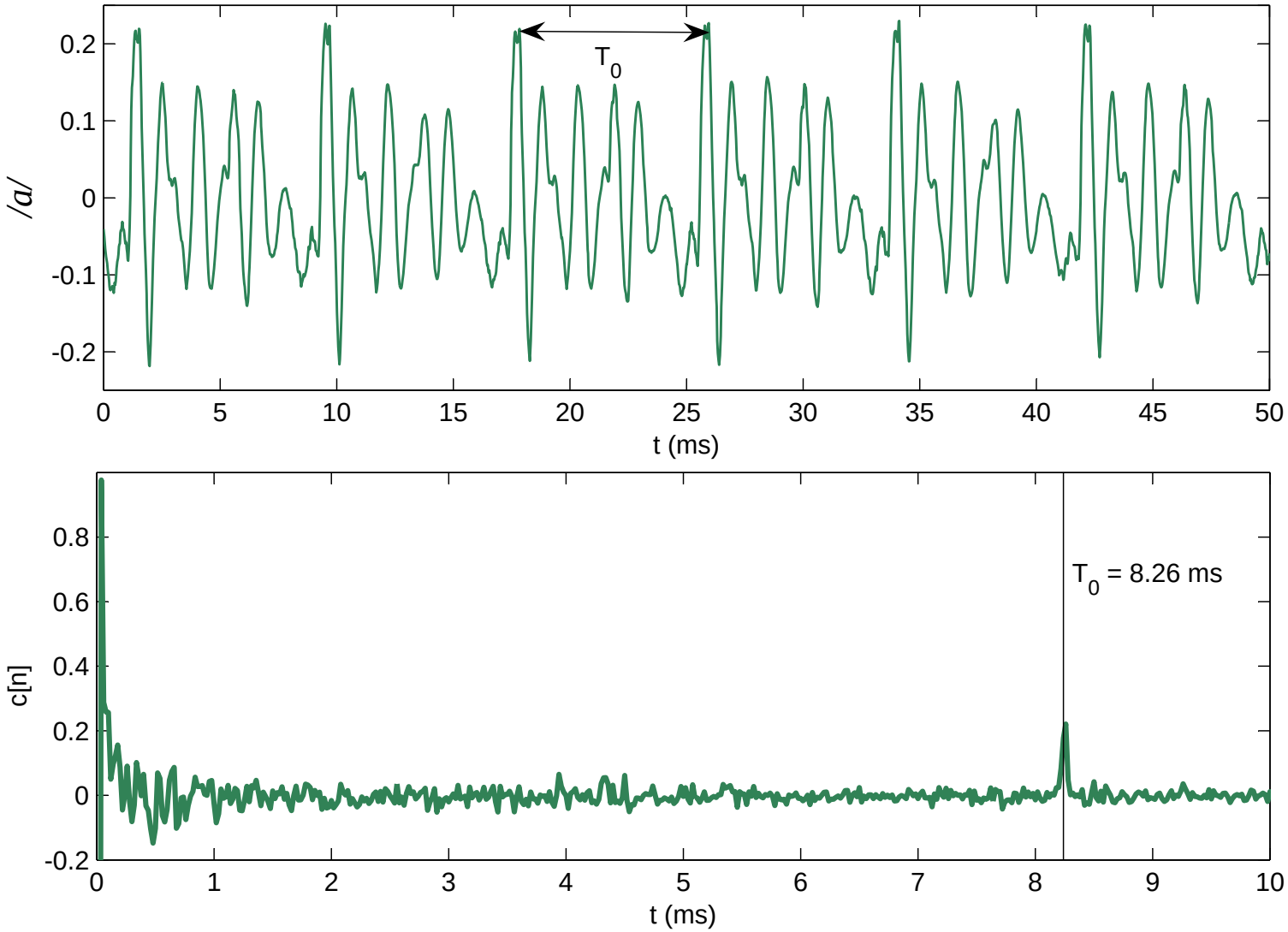
Coeficientes cepstrales en escala de mel

Integración por bandas



Organización de la clase

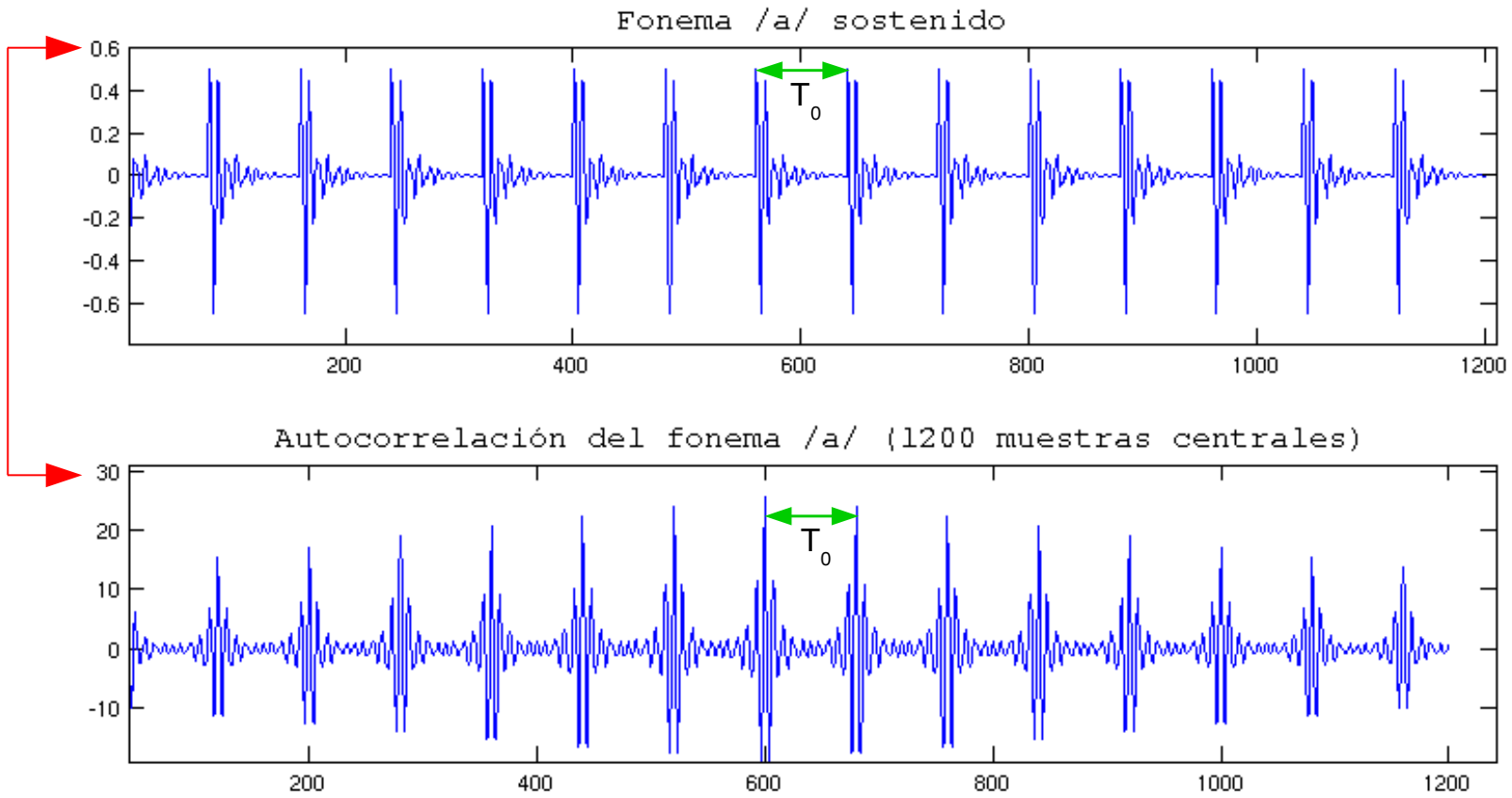
Estimación de F0 por cepstrum



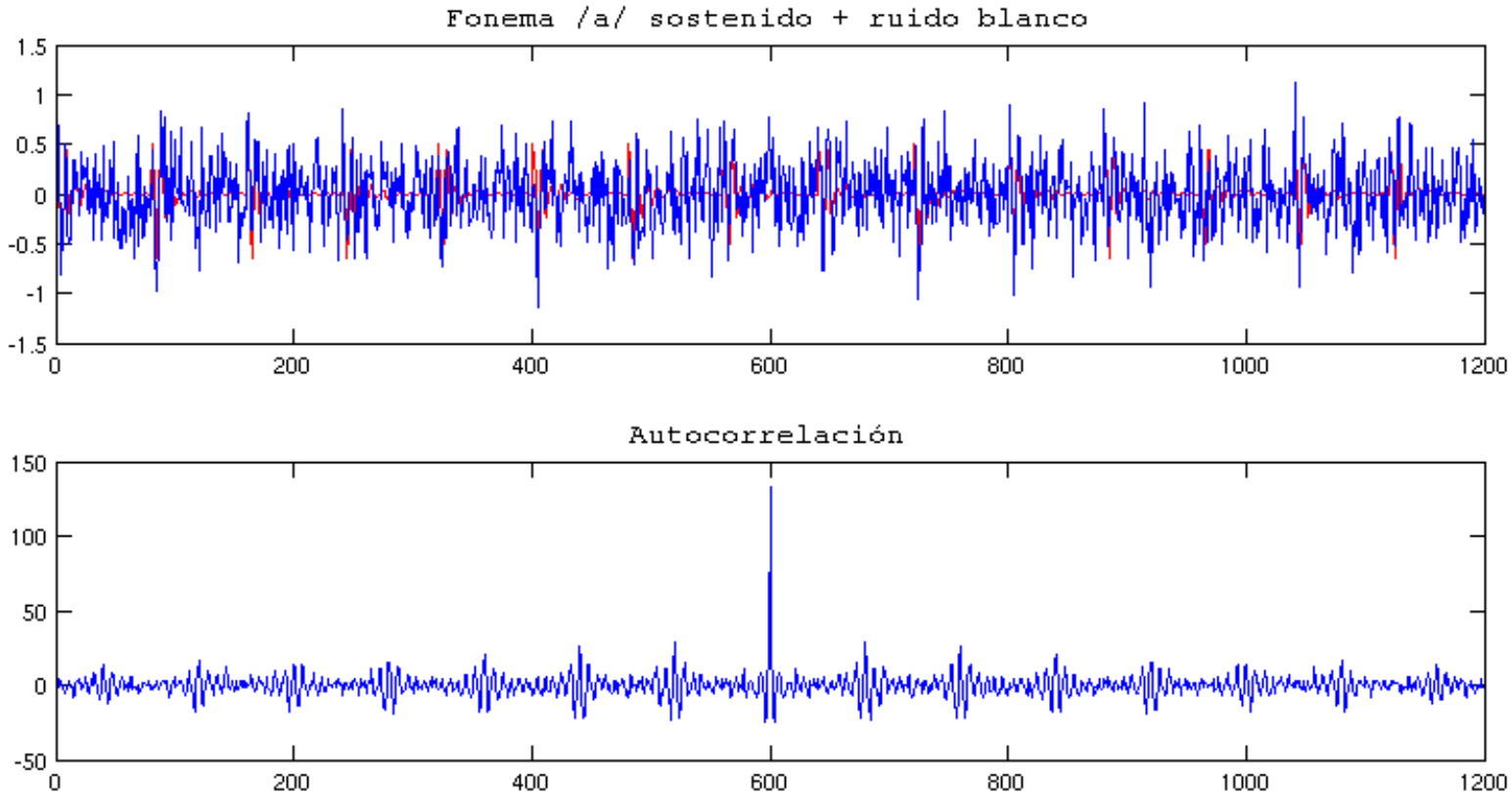
Estimación de F0 por autocorrelación

$$AC_x[j] = \sum_n x_n x_{n-j}$$

Estimación de F0 por autocorrelación



Estimación de F0 por autocorrelación



Bibliografía básica

- L. R. Rabiner y B. Gold, Theory and Application of Digital Signal Processing, Prentice Hall, 1975.
  Secciones: 12.1, 12.2, 12.3 y 12.13.
- J. R. Deller, J. G. Proakis, J. H. Hansen, Discrete-Time Processing of Speech Signals, Prentice Hall, 1993.
  Secciones: 4.1, 4.2.1, 4.2.2, 6.1 y 6.2.
  → Error en la figura 6.3 (c), pp 361.
- H.L. Rufiner, "Análisis y modelado digital de la voz: Técnicas recientes y aplicaciones", Editorial UNL, 2009. (Capítulo 3).
- J. Makhoul, "Linear Prediction: A Tuturial Review," Proc. IEEE, vol 63, no. 4, páginas 561-580, 1975.

Bibliografía básica

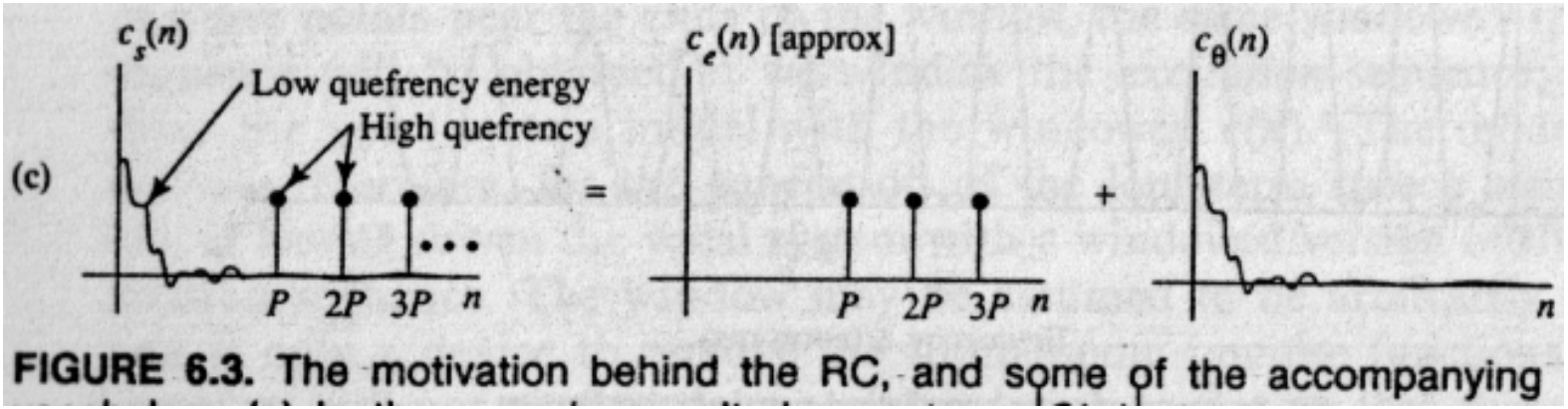
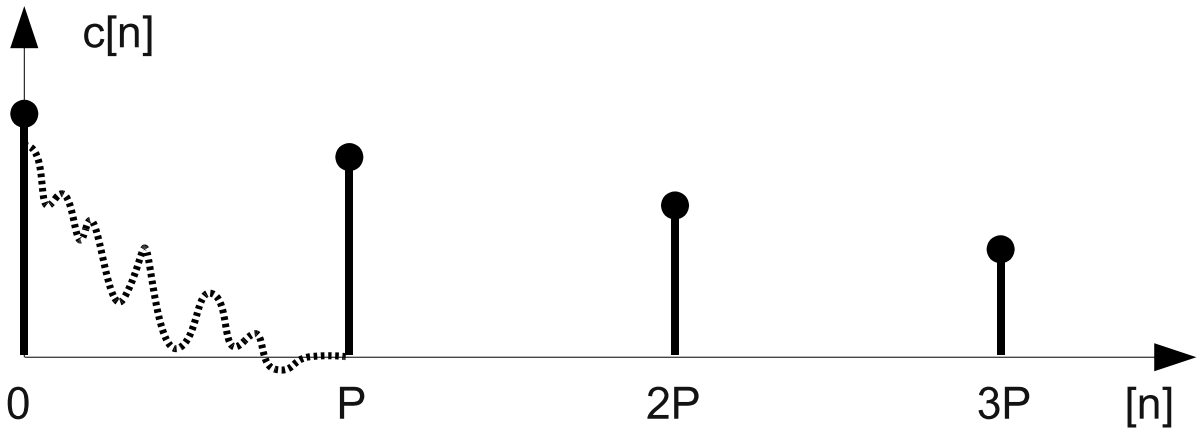


FIGURE 6.3. The motivation behind the RC, and some of the accompanying



Bibliografía básica

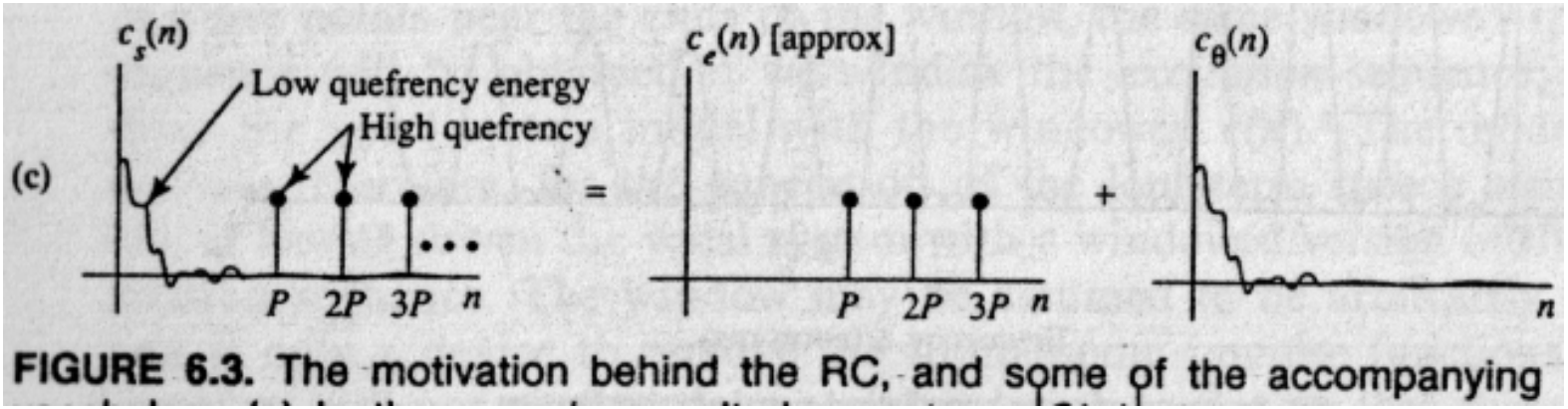


FIGURE 6.3. The motivation behind the RC, and some of the accompanying