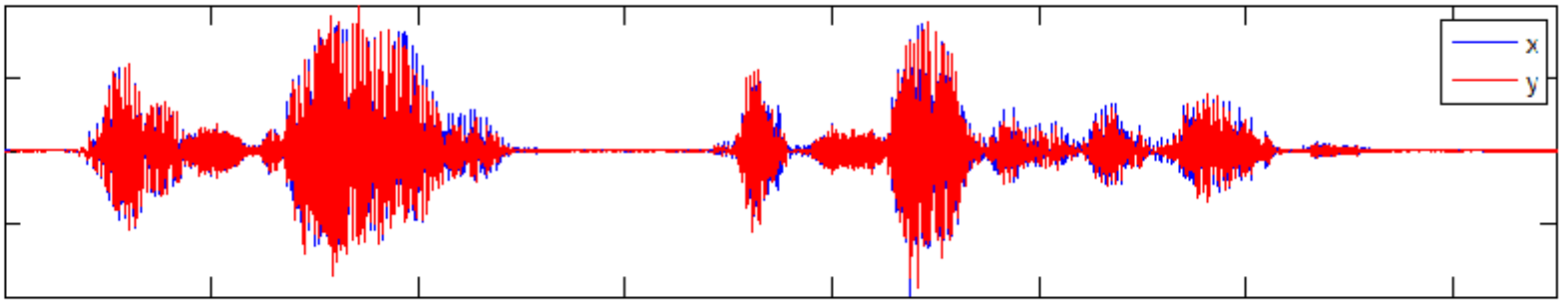


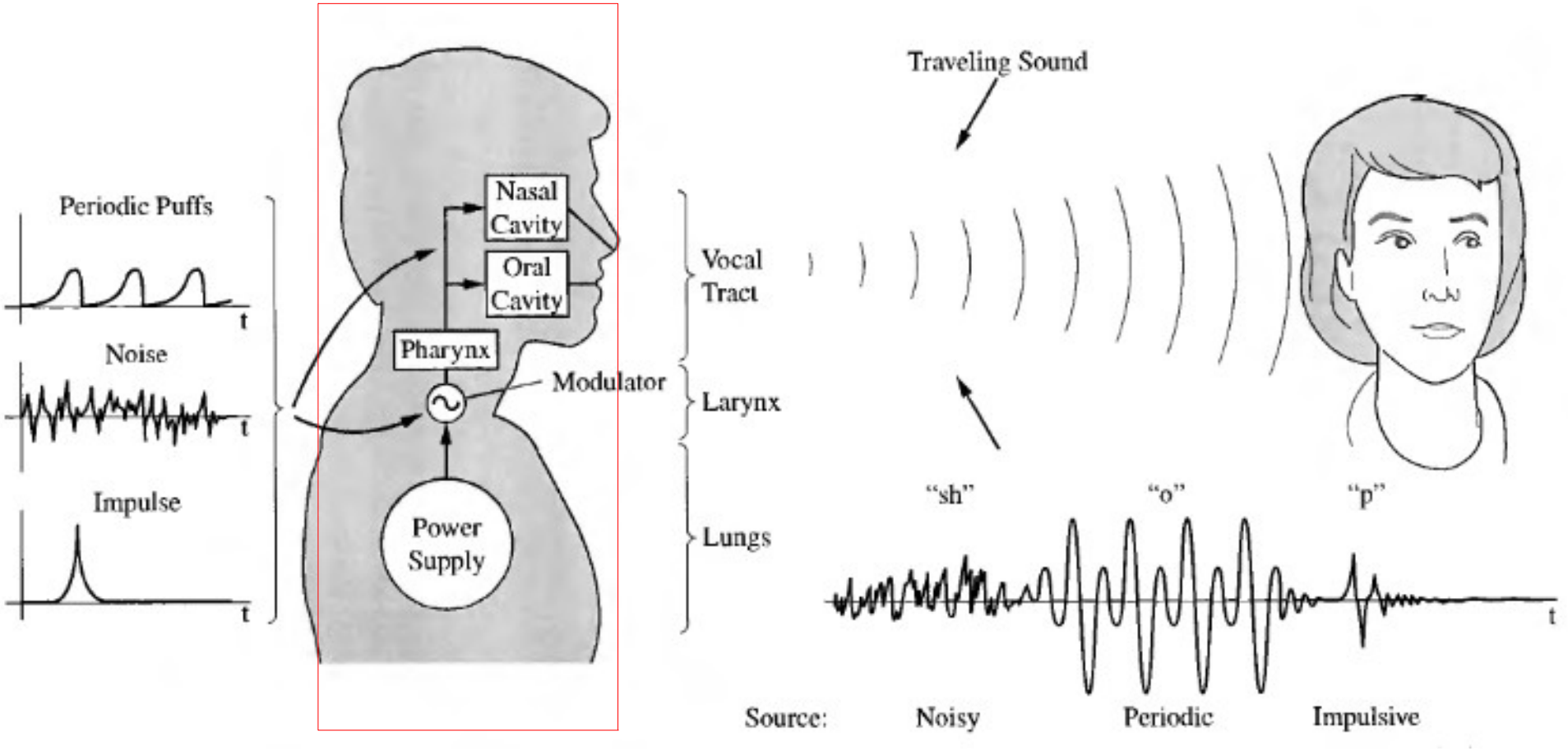
ELO-313: Procesamiento Digital de Señales con Aplicaciones.
Primer Semestre 2013 - UTFSM

Profesor: Dr. Matías Zañartu.

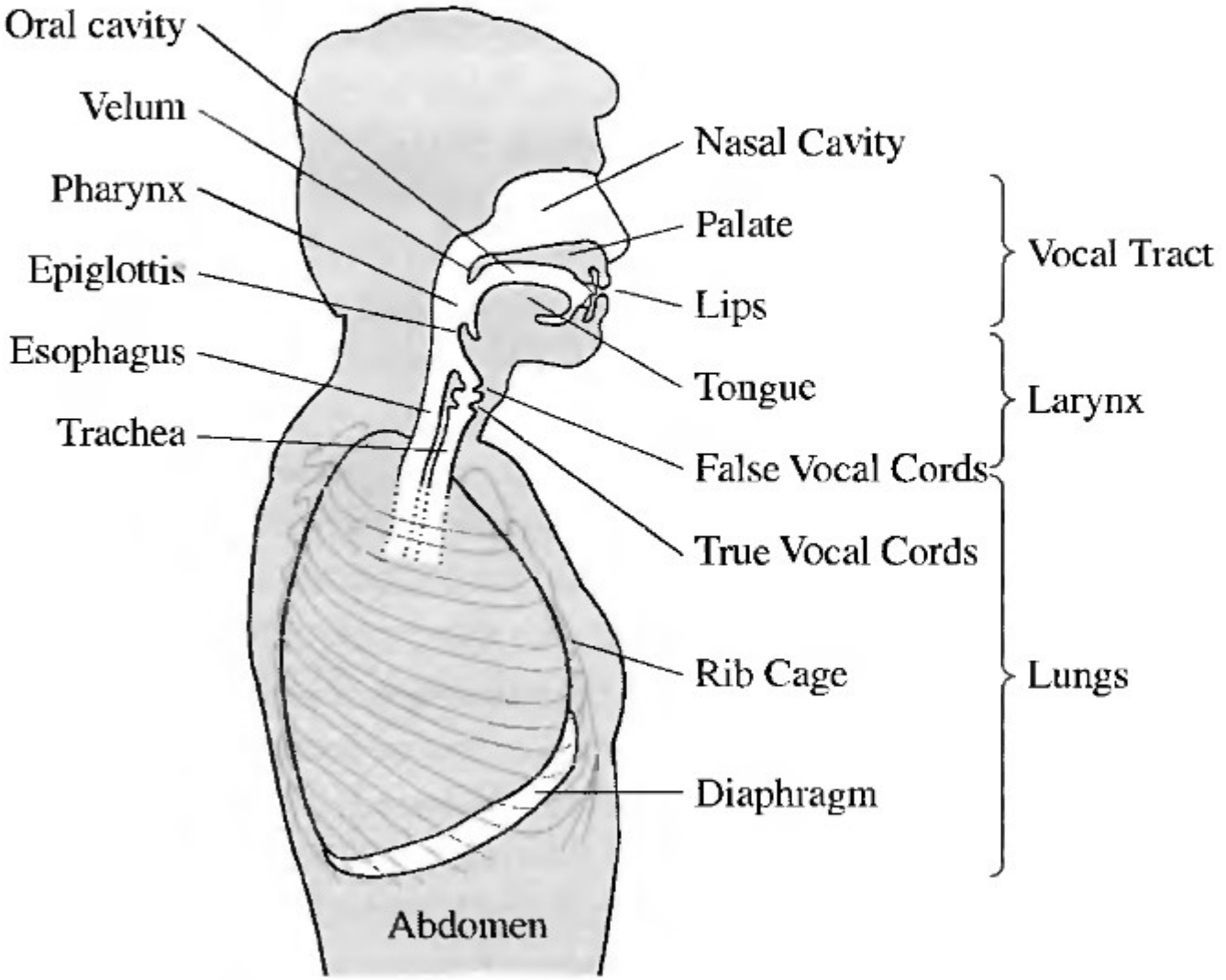
Introducción a la Voz Humana

Presentado por: Víctor Espinoza.

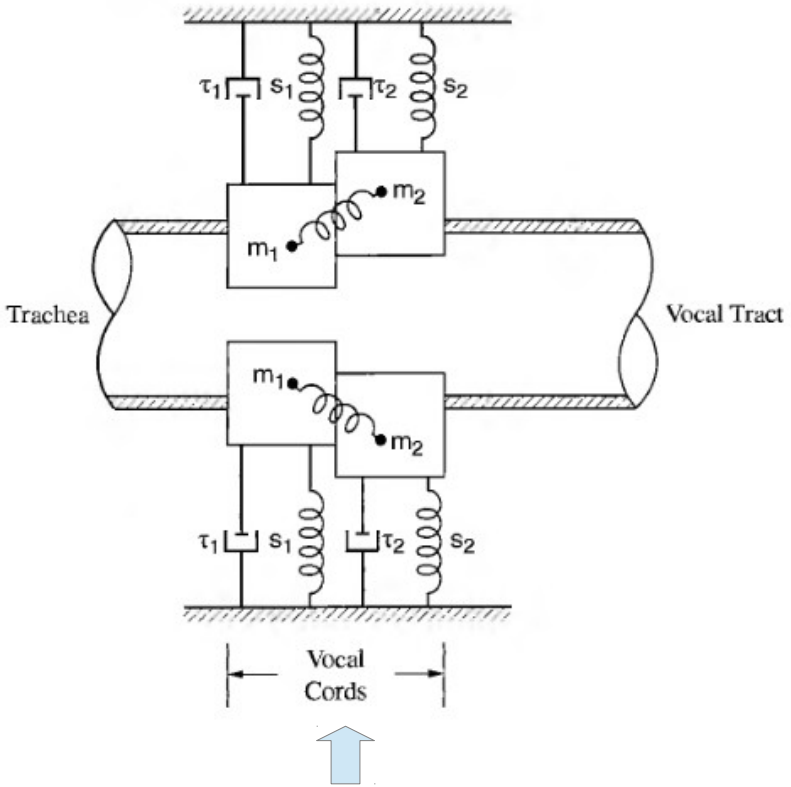
Generación Voz



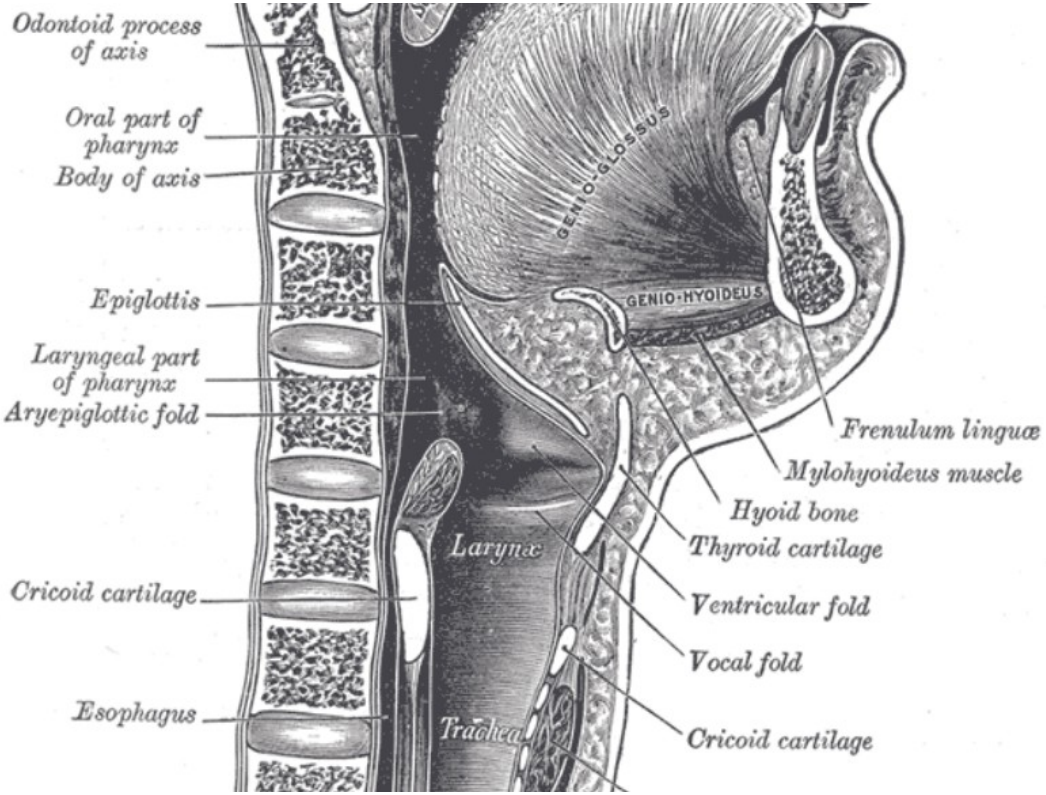
Anatomía


Oral cavity
Velum
Pharynx
Epiglottis
Esophagus
Trachea
Nasal Cavity
Palate
Lips
Tongue
False Vocal Cords
True Vocal Cords
Rib Cage
Diaphragm
Abdomen
Vocal Tract
Larynx
Lungs

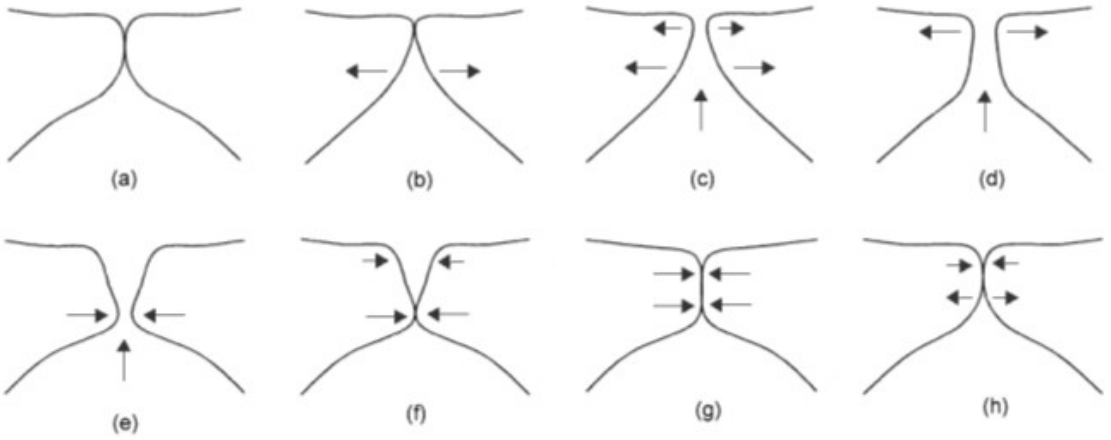
Fuente – Cuerdas Vocales



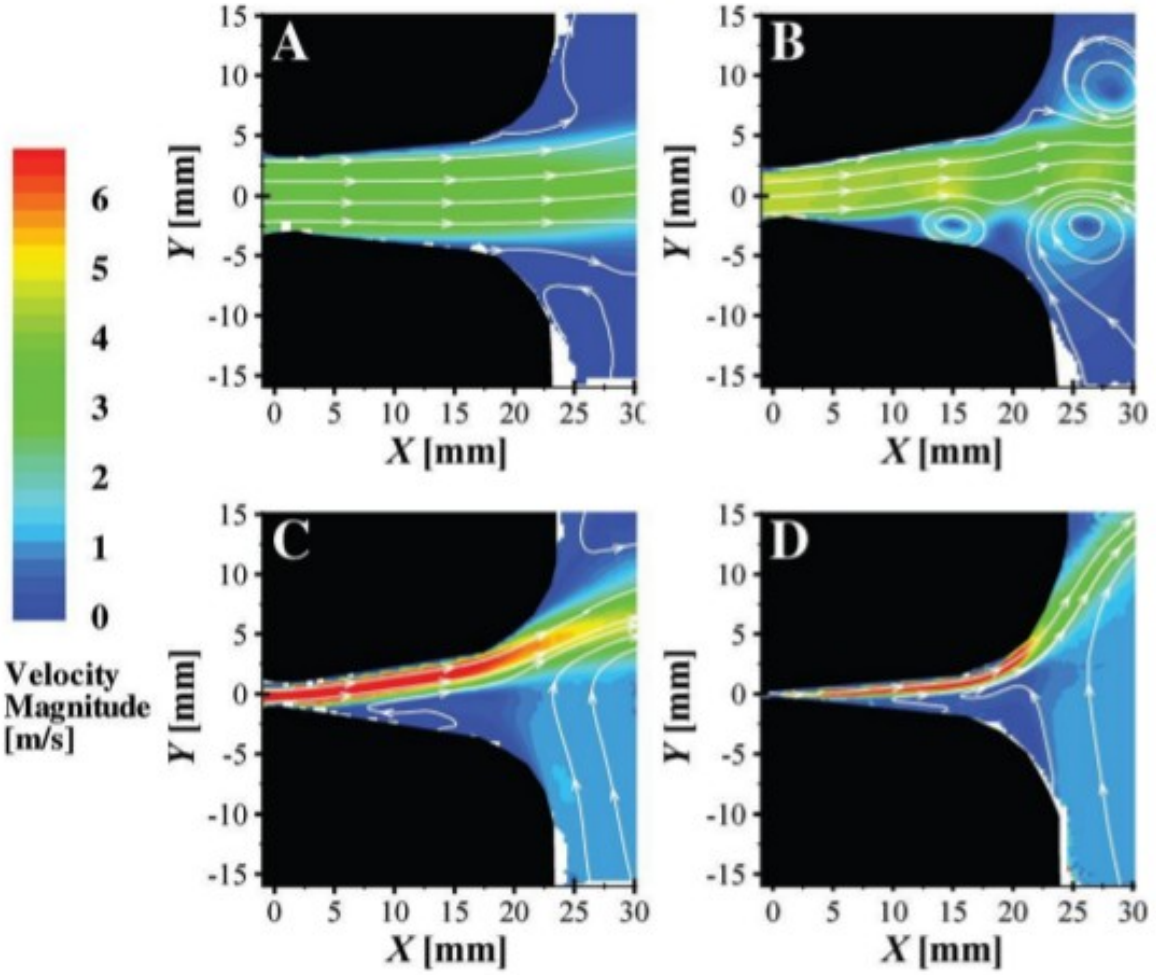
Modelo Mecánico

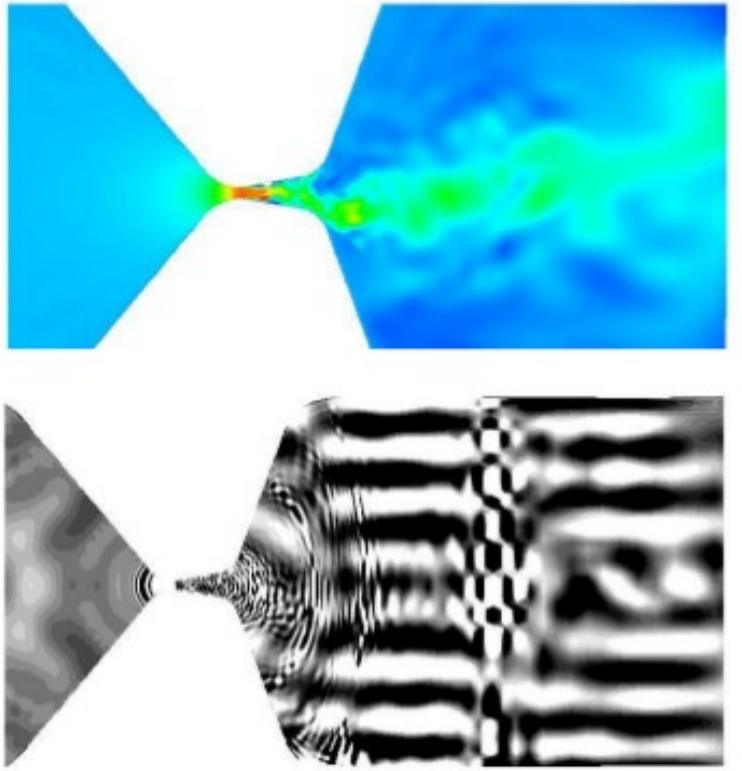


Ciclo Cuerdas Vocales



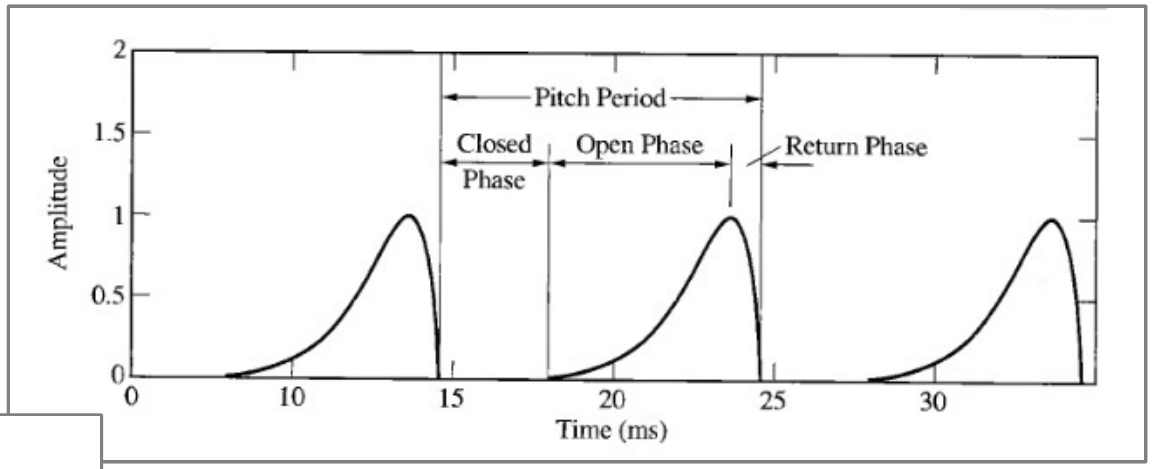
Fuente – Cuerdas Vocales

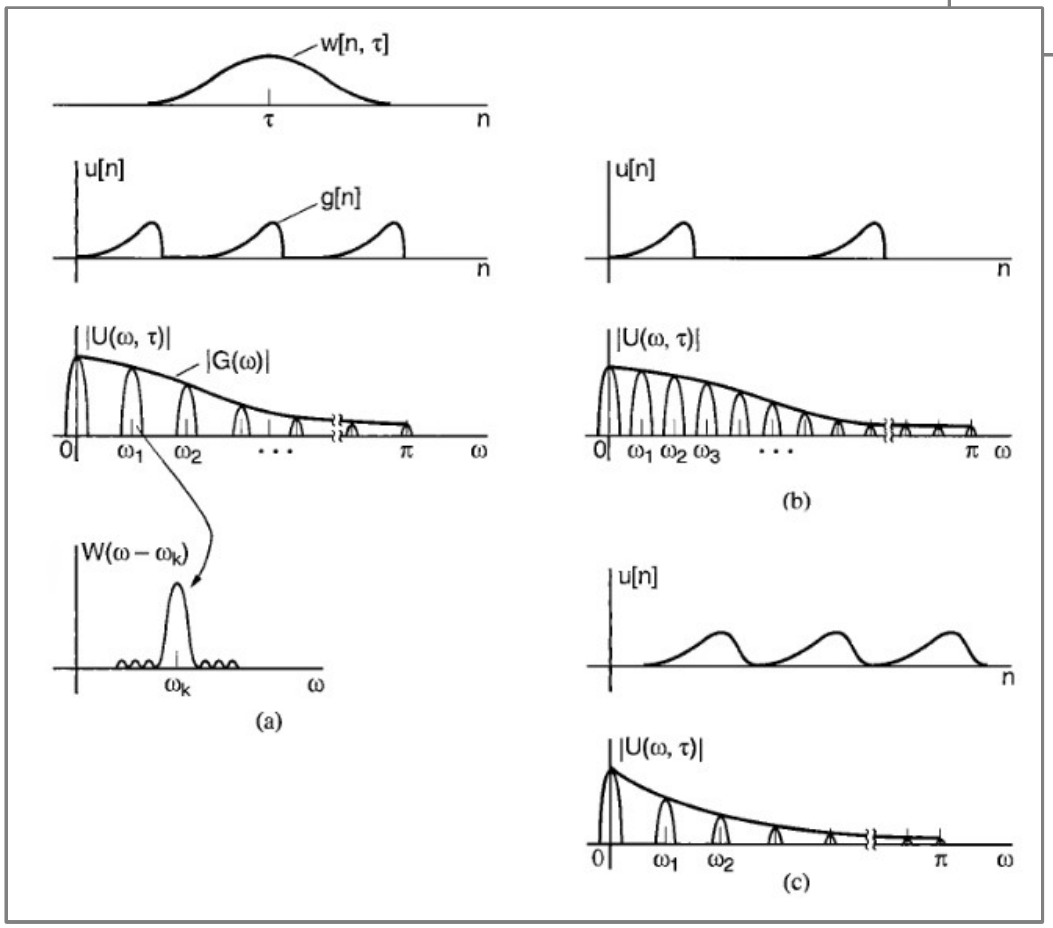




Fuente – Cuerdas Vocales

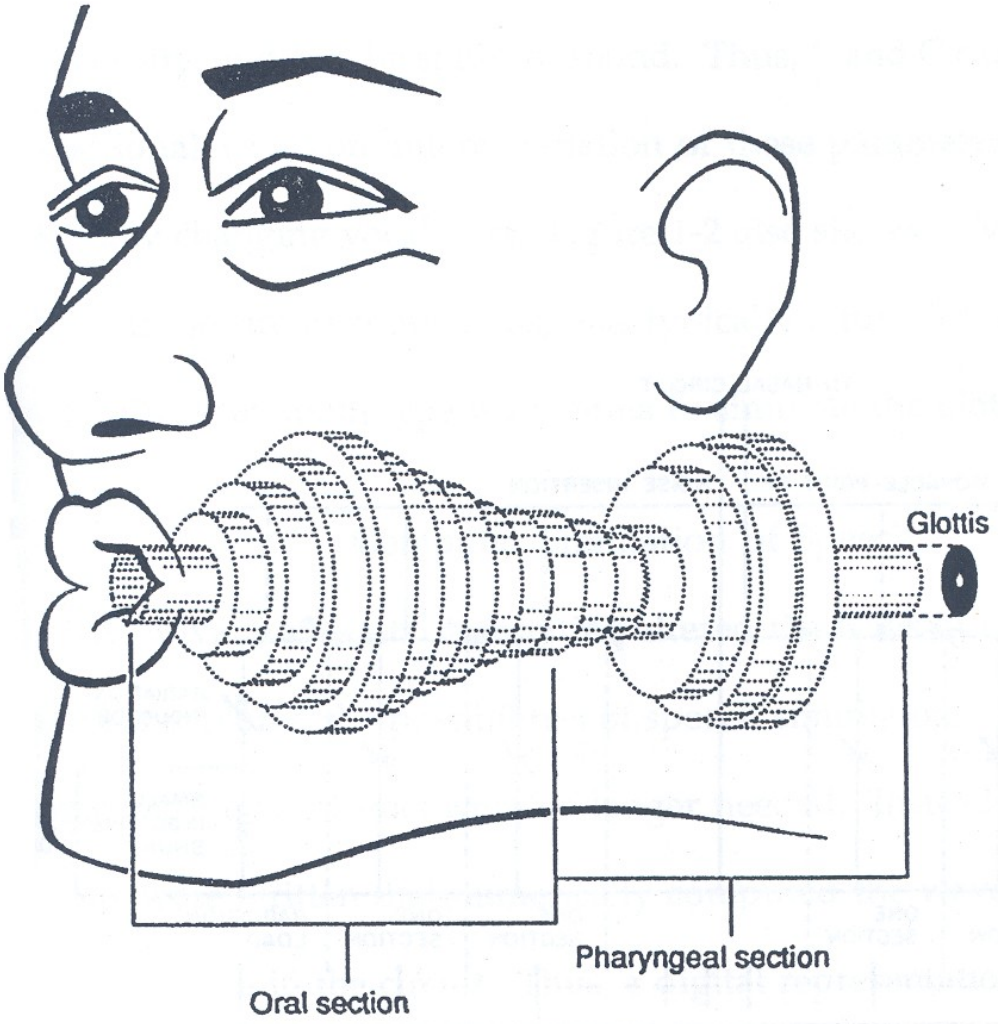
Impulso Glotal (tiempo)





Impulso Glotal (frecuencia)

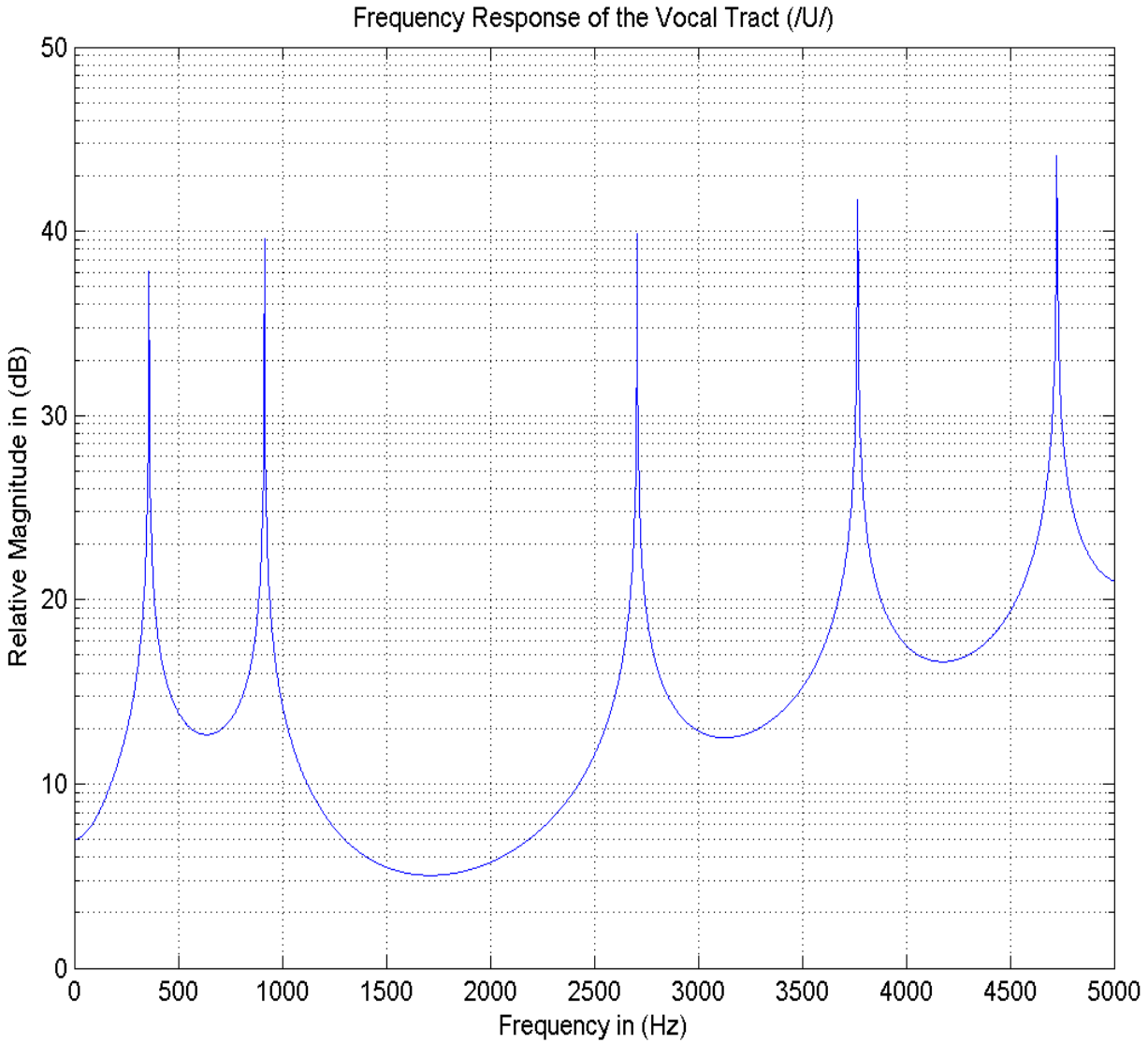
Sistema – Tracto Vocal



Discretization of the vocal tract with cylindrical sections (a conceptualization from Titze, 1994).

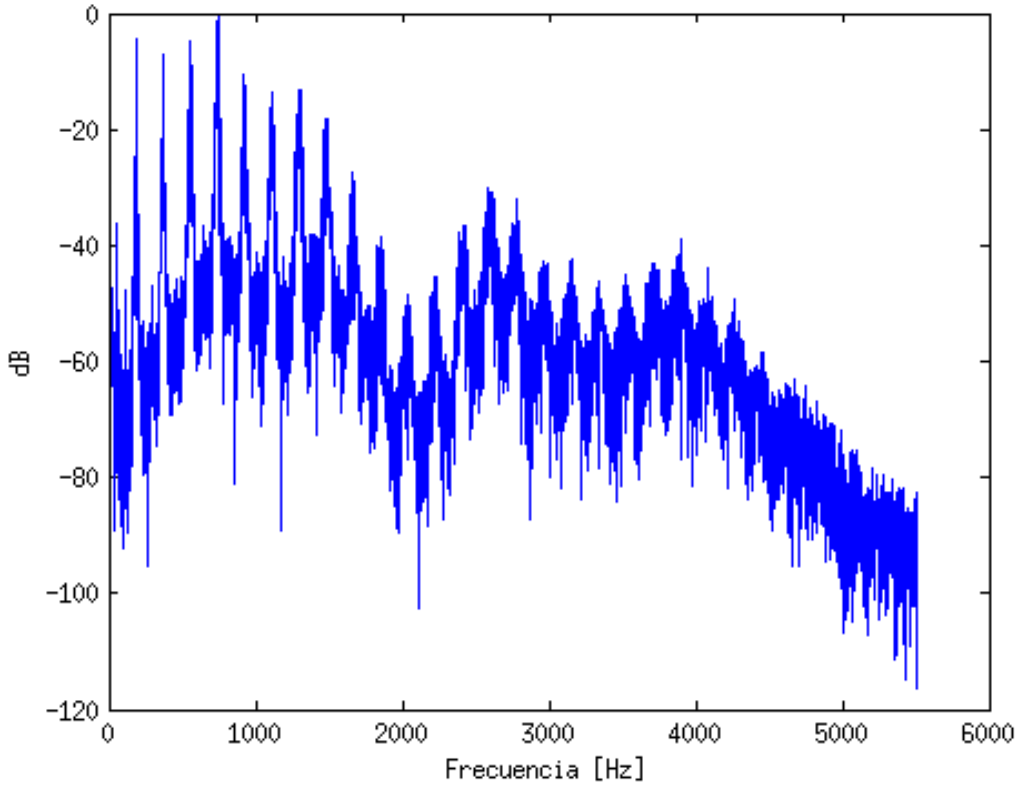
- Modelo en base a tubos.
- Articulación determina diámetro sección en modelo. Ej: /a/ <> /e/.
- Articulación depende apertura bucal y posición lengua.
- No considera cavidad nasal.

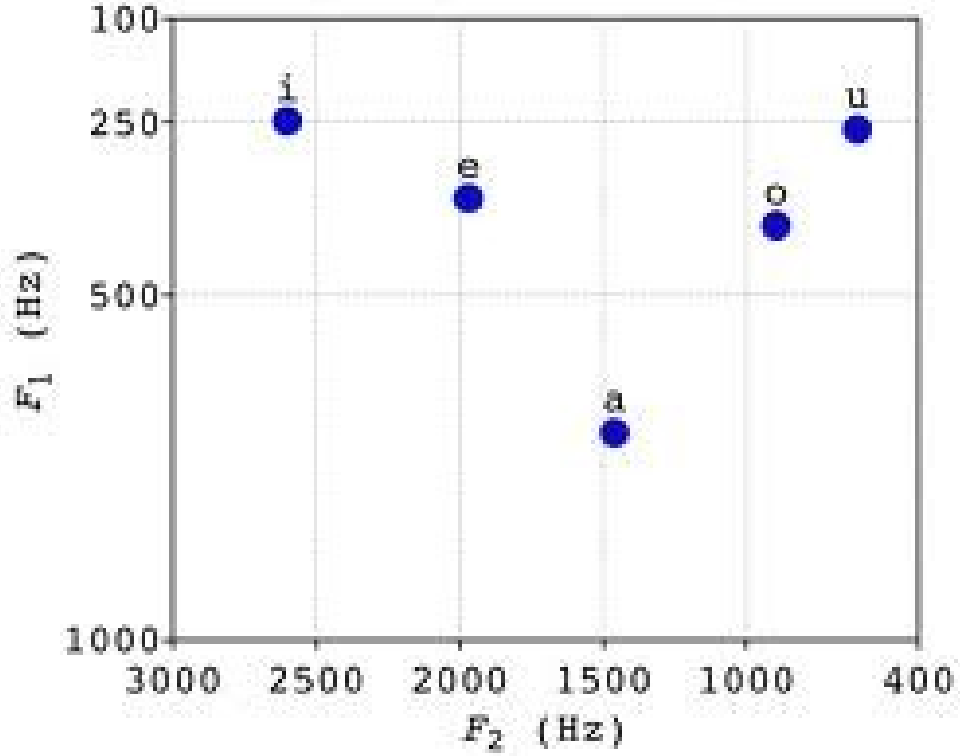
Sistema – Tracto Vocal



- Tubos → Infinitas resonancias.
- Resonancias → Polos
- “Polos” en Voz → Formantes.
- Formantes → Muy importante para la codificación del lenguaje!!!
- Frecuencias de resonancia dependen de la articulación.

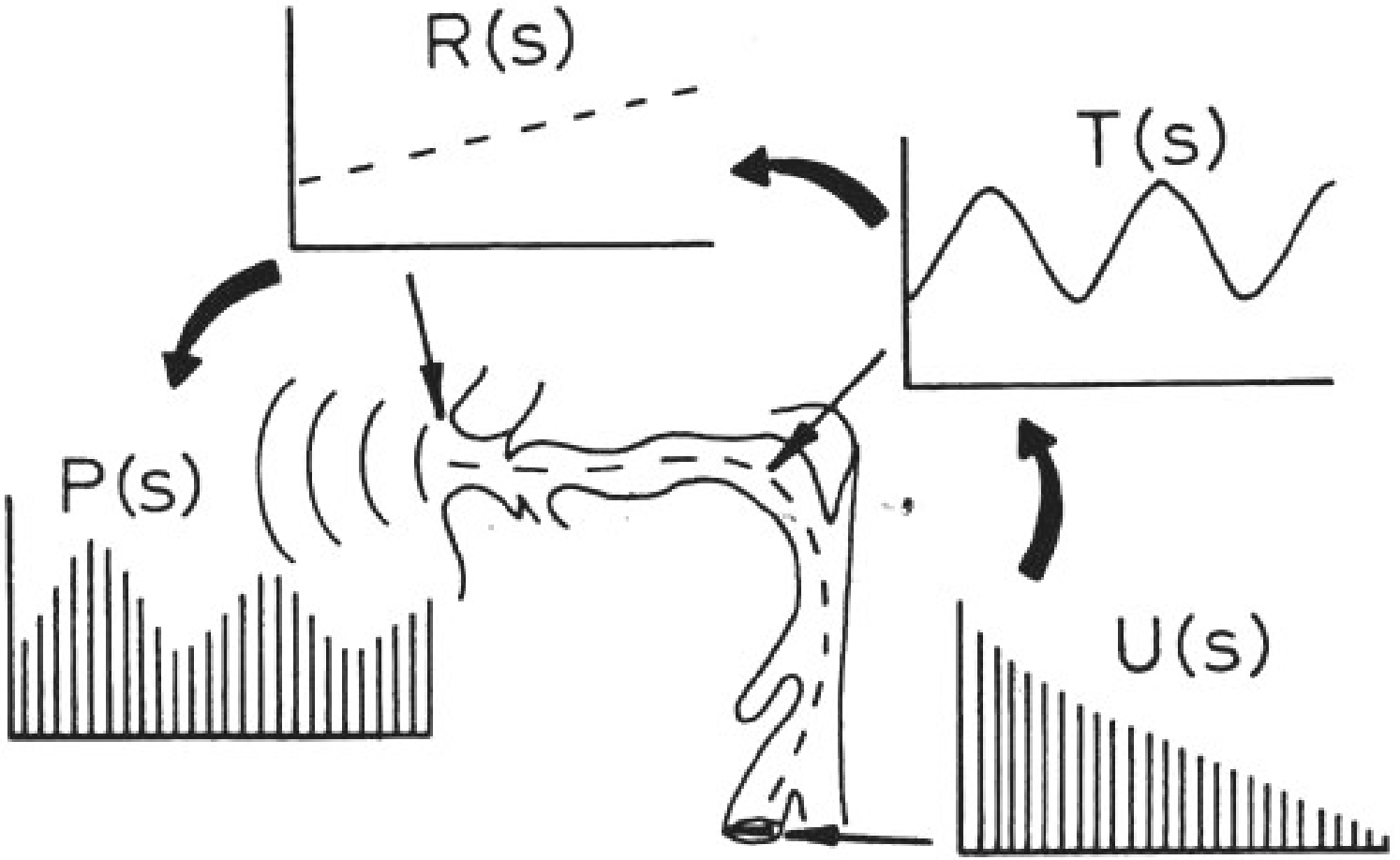
Sistema – Tracto Vocal





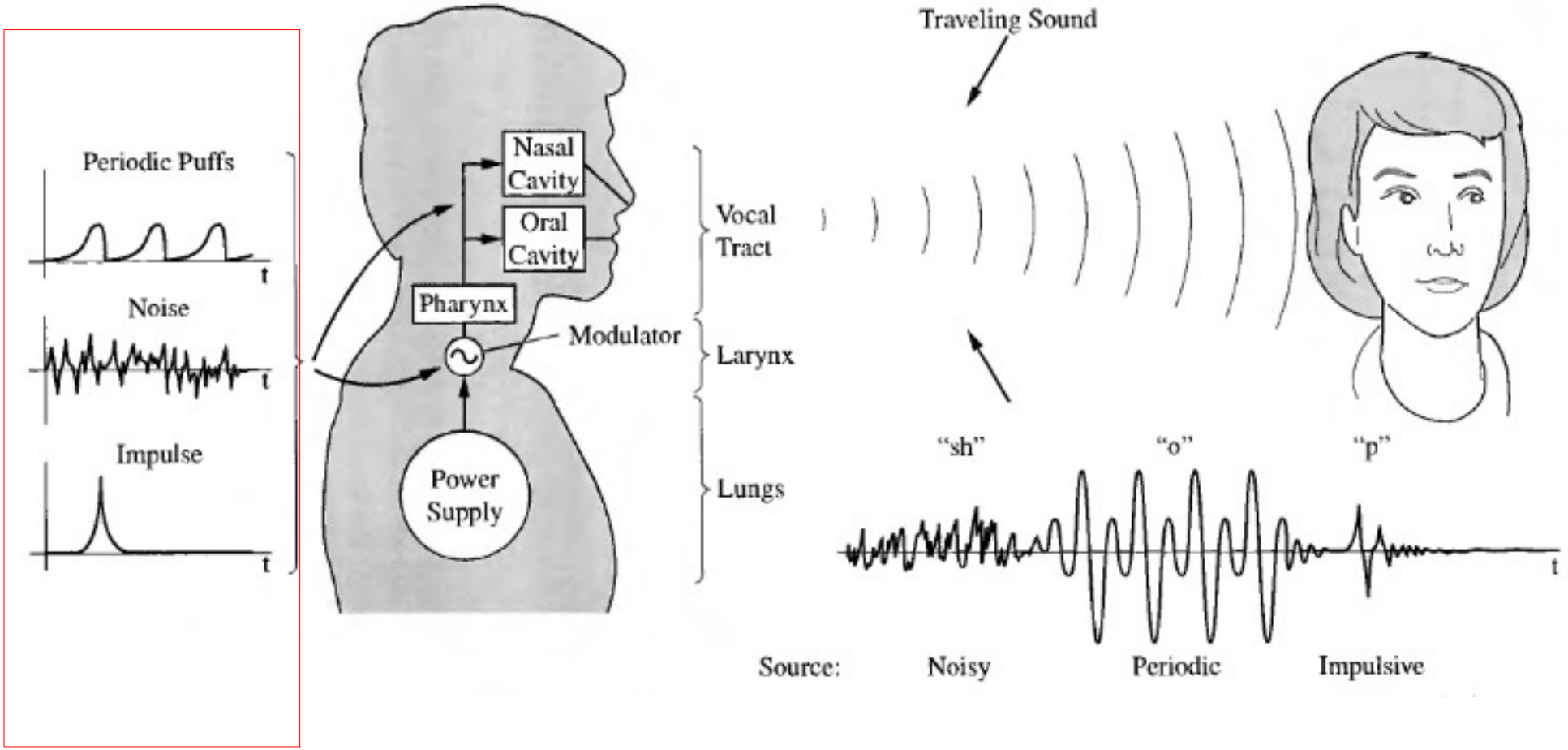
Carta de Formantes Vocales

Sistema Completo Fonación

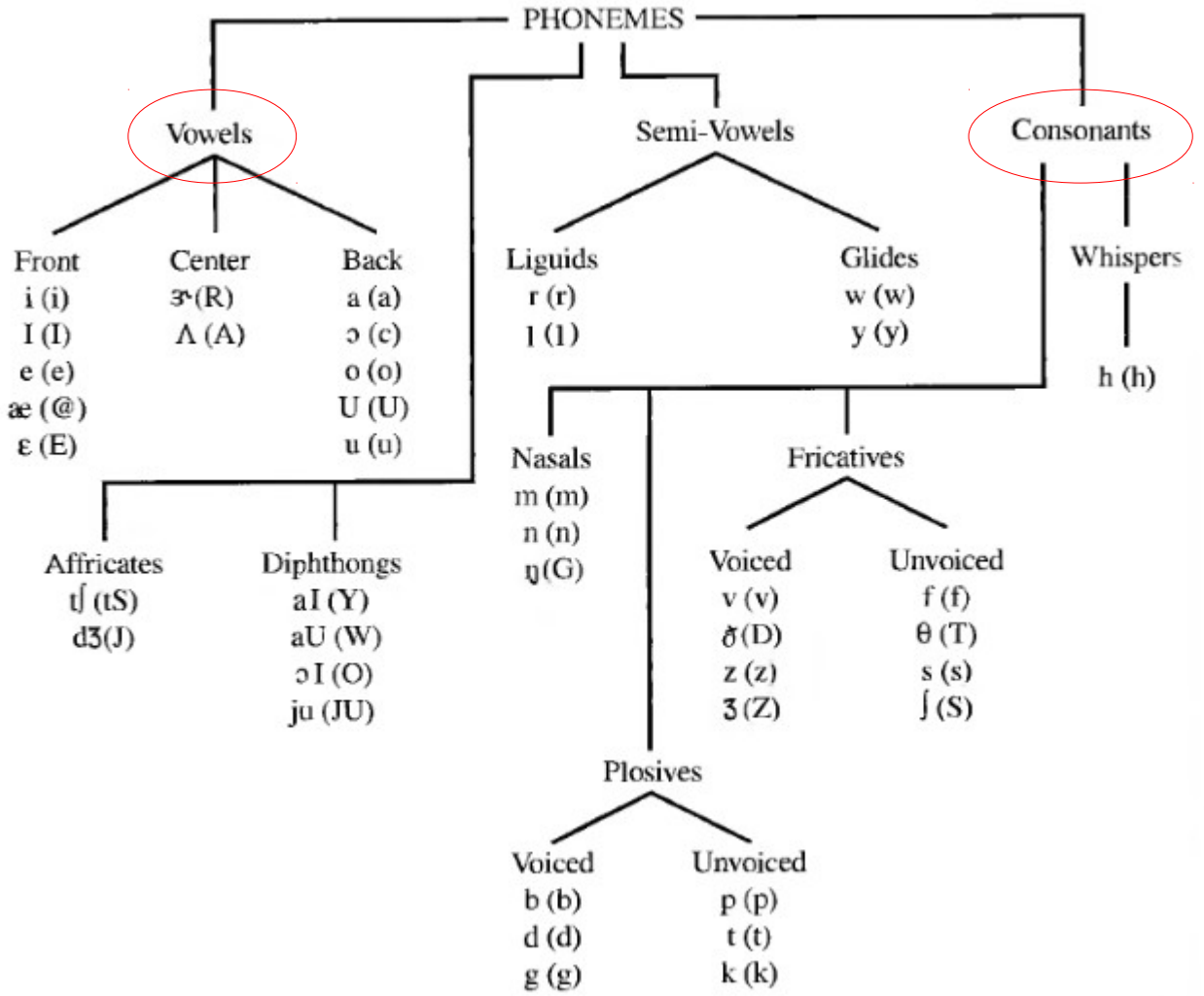


$$P(s) = U(s)\ T(s)\ R(s)$$

Sonidos Fuente



Fonética

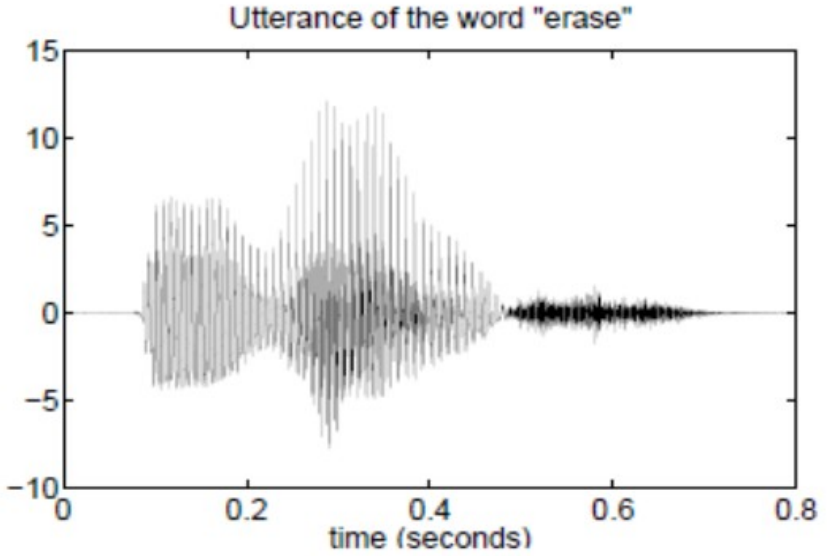

PHONEMES
Vowels
Semi-Vowels
Consonants
Front
i (i)
I (I)
e (e)
æ (@)
ɛ (E)
Center
ɝ (R)
ʌ (A)
Back
a (a)
ɔ (c)
o (o)
U (U)
u (u)
Liguids
r (r)
l (l)
Glides
w (w)
y (y)
Whispers
h (h)
Affricates
tʃ (tS)
dʒ (J)
Diphthongs
aI (Y)
aU (W)
ɔI (O)
ju (JU)
Nasals
m (m)
n (n)
ŋ (G)
Fricatives
Voiced
v (v)
ð (D)
z (z)
ʒ (Z)
Unvoiced
f (f)
θ (T)
s (s)
ʃ (S)
Plosives
Voiced
b (b)
d (d)
g (g)
Unvoiced
p (p)
t (t)
k (k)

DSP en Voz

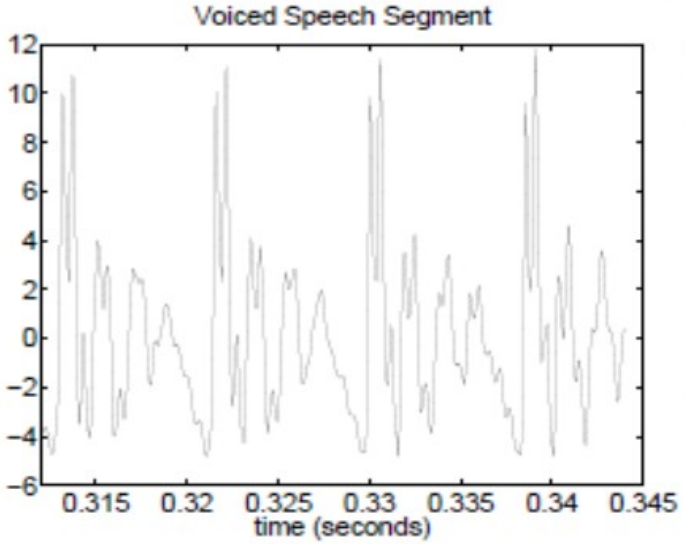
- Herramientas de Análisis:

  - Time Signal – Waveform
  - FFT – Fast Fourier Transform
  - Spectrogram – Short Time Fourier Transform
  - LPC – Linear Predictive Code
  - Cepstrum – Homomorphic Signal Processing

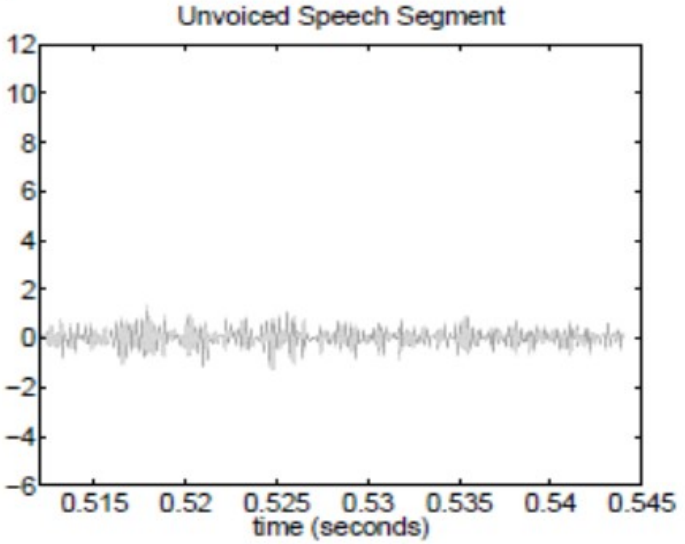
DSP en Voz



(a)



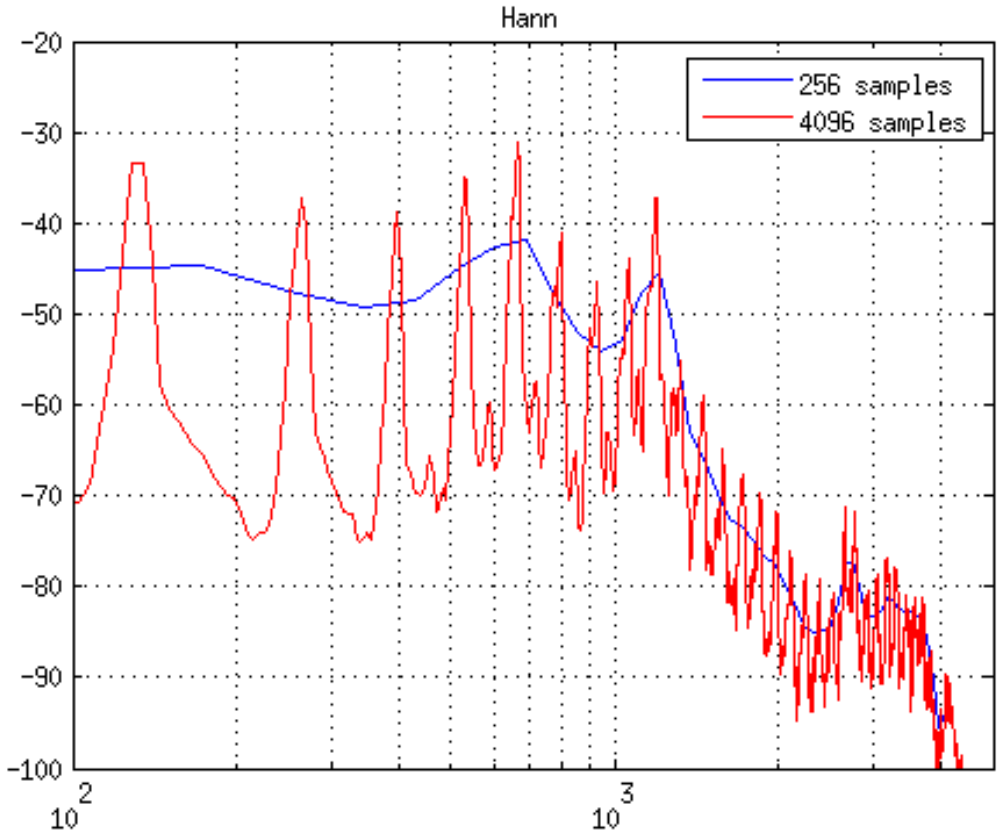
(b)

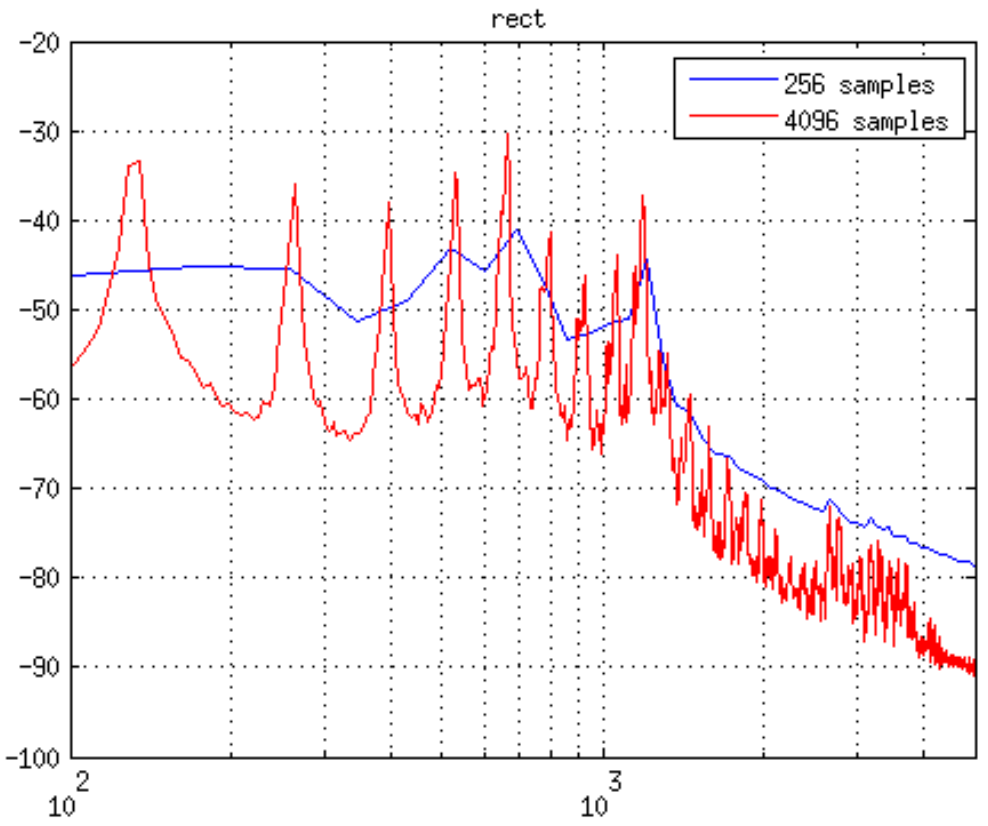


(c)

DSP en Voz

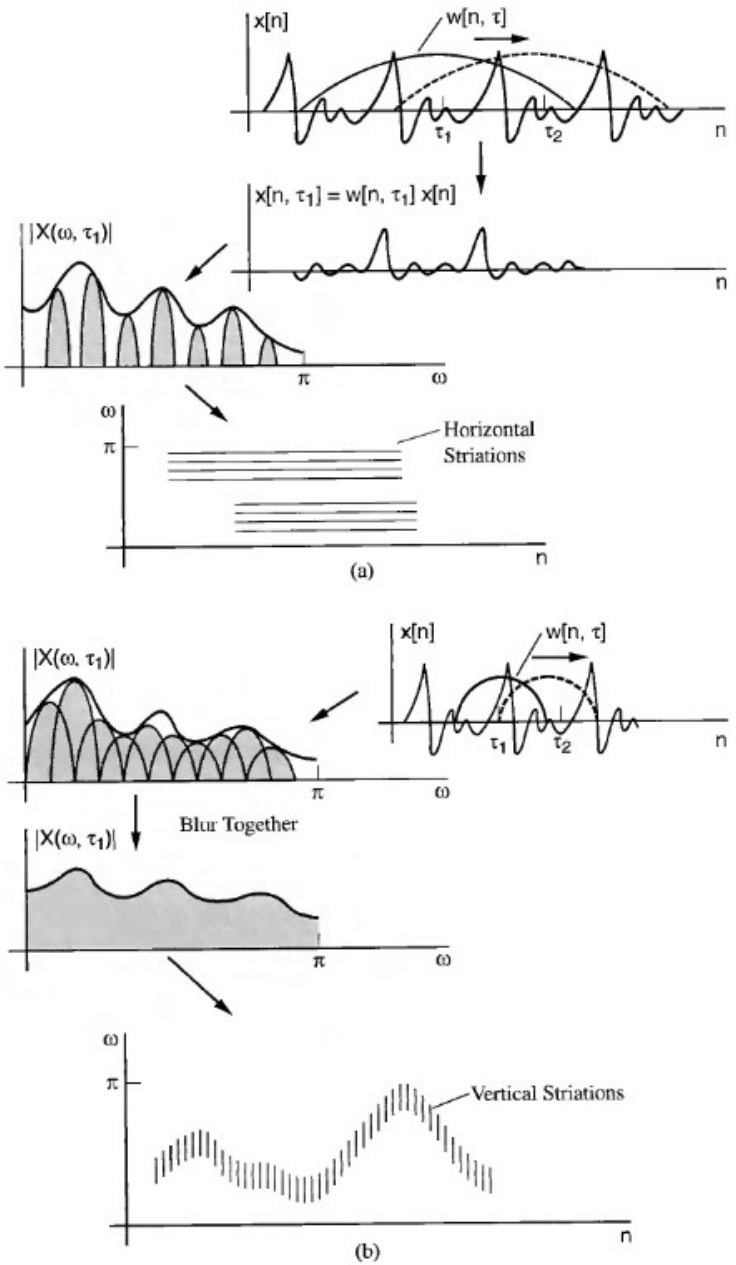
- FFT
  - Parameters:
    - Smoothing Window (Hann, Hamming, etc)
    - Buffer size (128, 256, etc)
    - Decimate(*)
    - Overlap(*)
    - Padding(*)



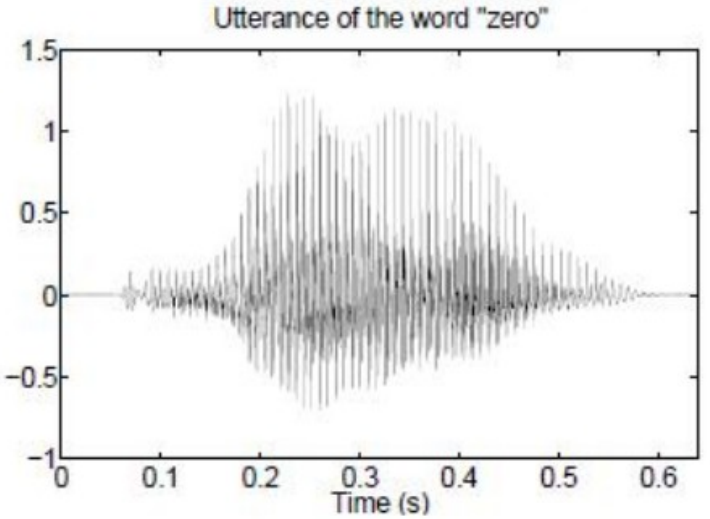


DSP en Voz

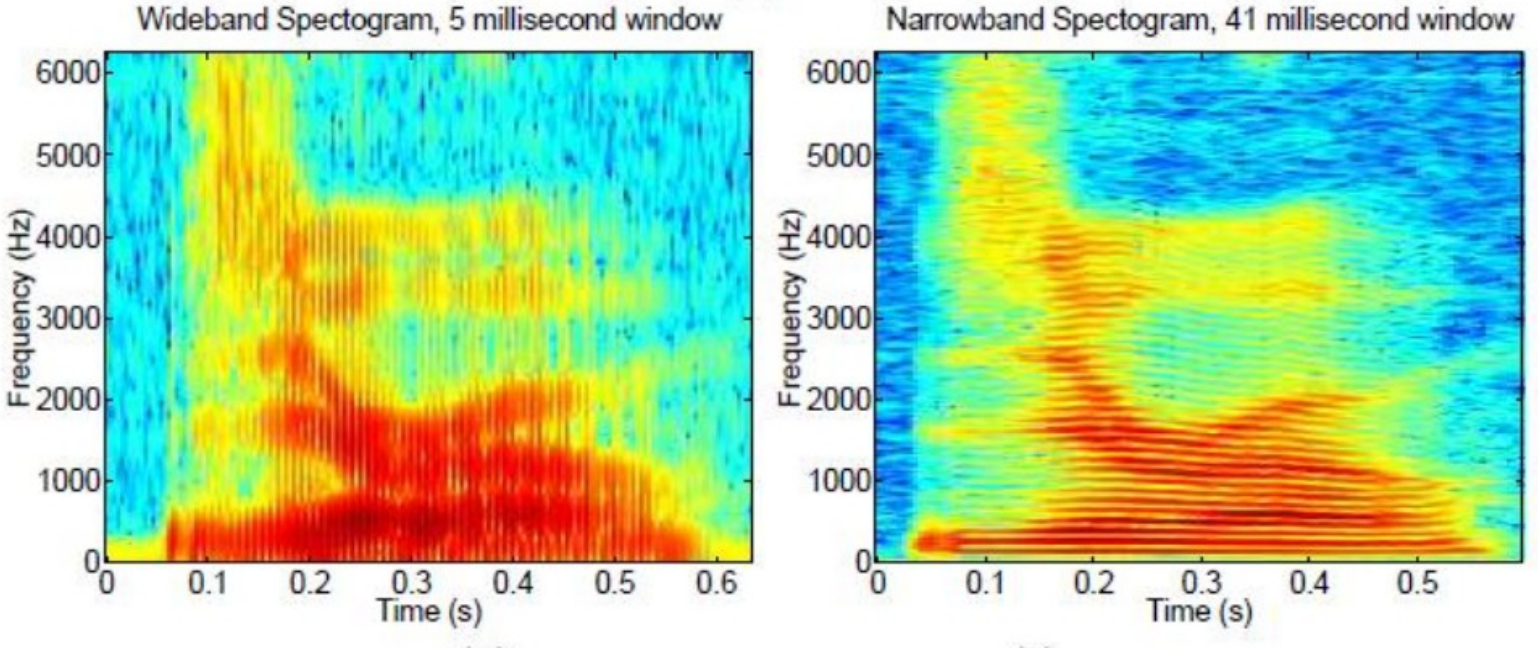
- Spectrogram:
  - Parameters:
    - Same as FFT



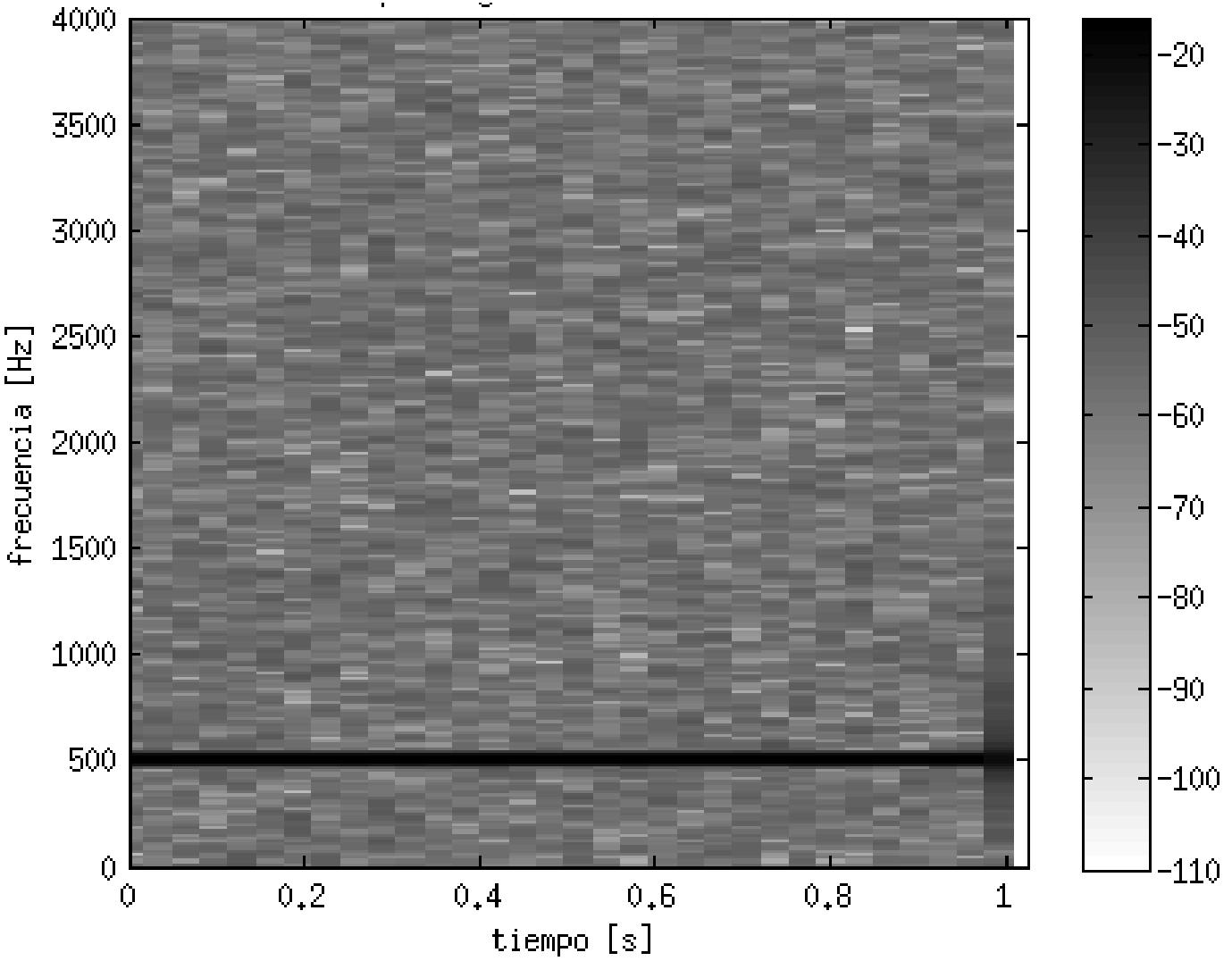
DSP en Voz


Utterance of the word "zero"
1.5
1
0.5
0
−0.5
−1
0
0.1
0.2
0.3
0.4
0.5
0.6
Time (s)

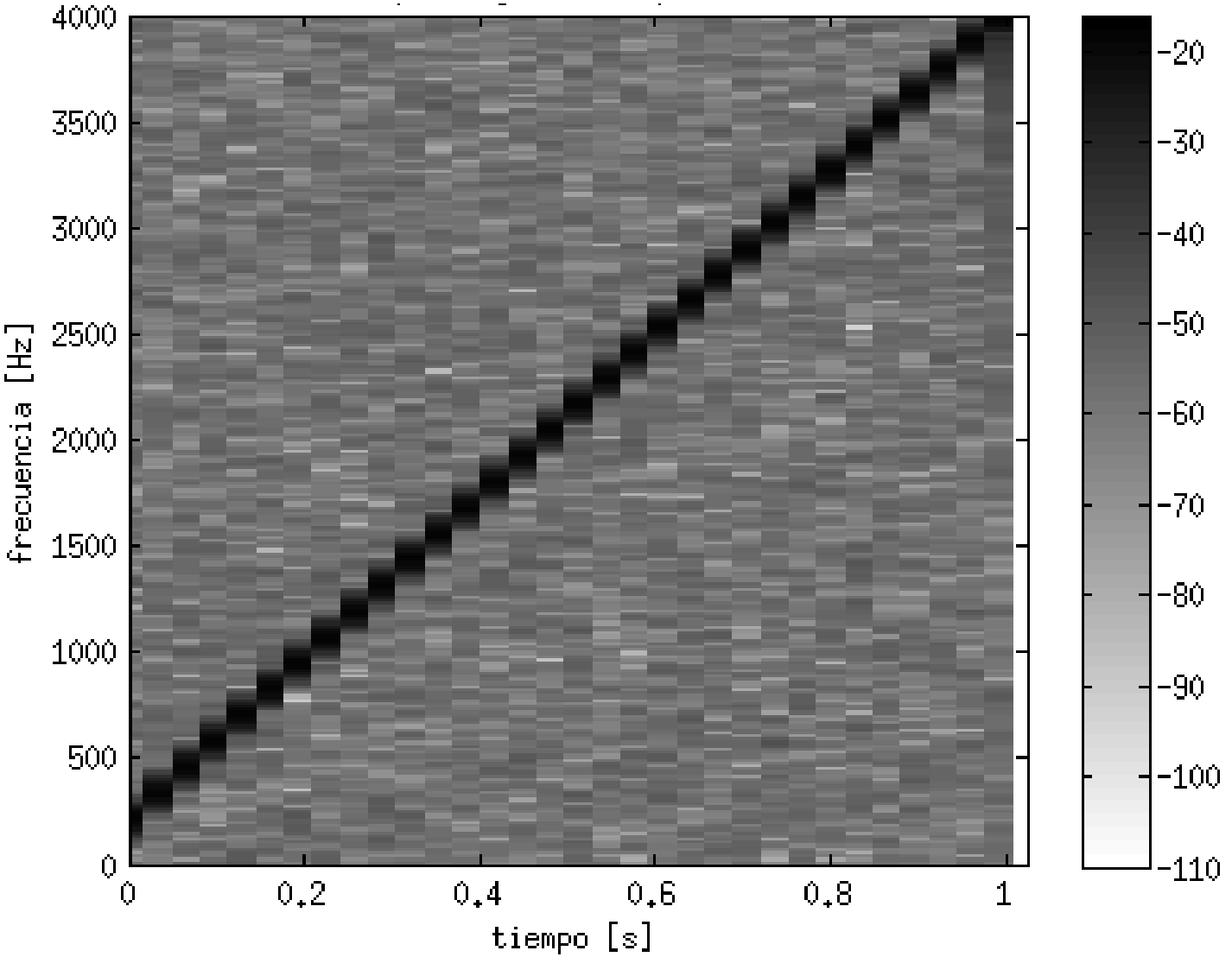
(a)


Wideband Spectogram, 5 millisecond window
Narrowband Spectogram, 41 millisecond window
Frequency (Hz)
6000
5000
4000
3000
2000
1000
0
0.1
0.2
0.3
0.4
0.5
0.6
Time (s)

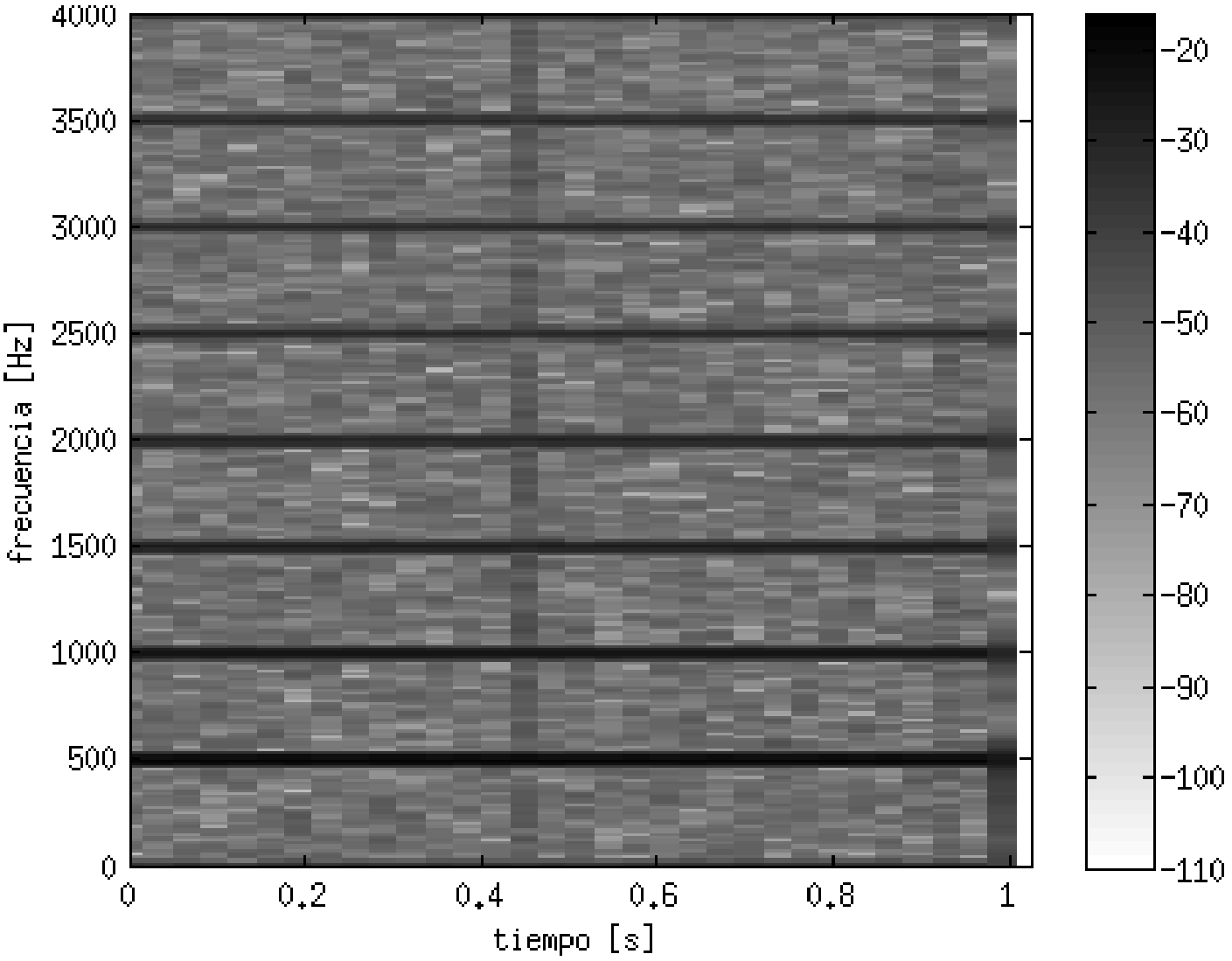
DSP en Voz


4000
3500
3000
2500
2000
1500
1000
500
0
frecuencia [Hz]
0
0.2
0.4
0.6
0.8
1
tiempo [s]
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110

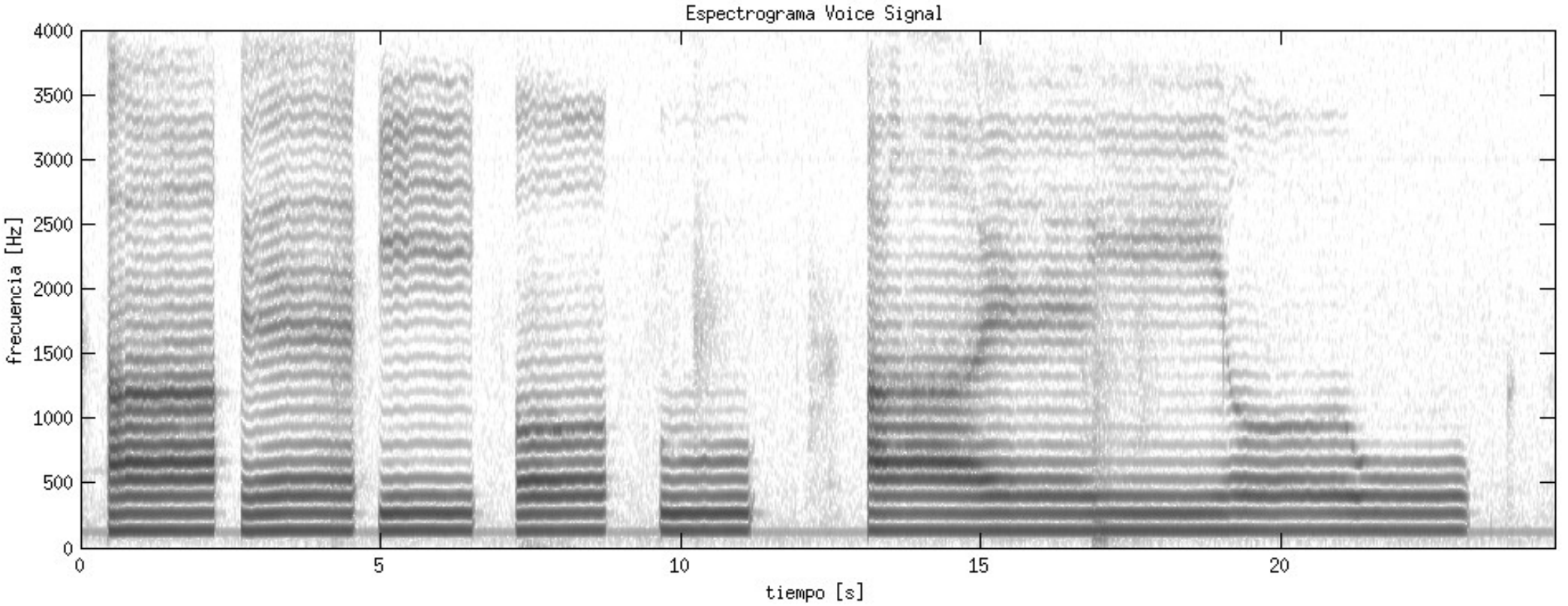
DSP en Voz


4000
3500
3000
2500
2000
1500
1000
500
0
frecuencia [Hz]
0
0.2
0.4
0.6
0.8
1
tiempo [s]
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110

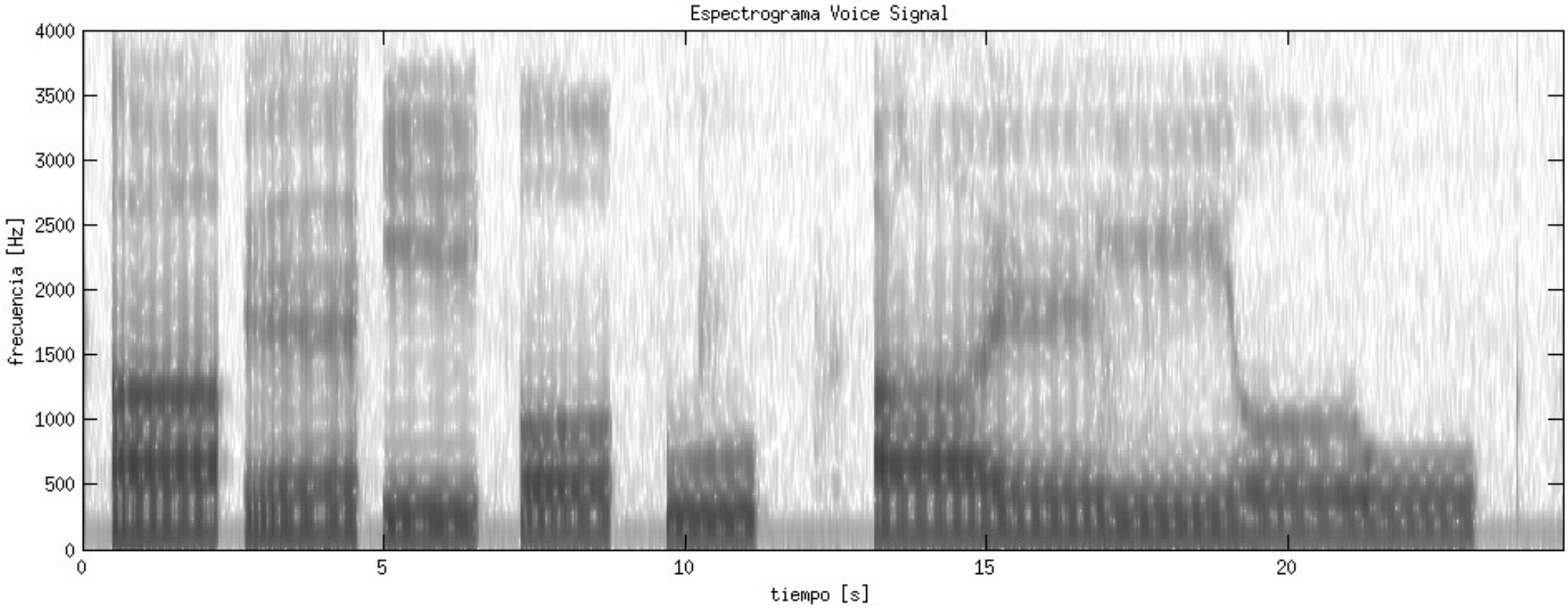
DSP en Voz


4000
3500
3000
2500
2000
1500
1000
500
0
frecuencia [Hz]
0
0.2
0.4
0.6
0.8
1
tiempo [s]
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110

DSP en Voz


Espectrograma Voice Signal
frecuencia [Hz]
4000
3500
3000
2500
2000
1500
1000
500
0
0
5
10
15
20
tiempo [s]

DSP en Voz


Espectrograma Voice Signal
frecuencia [Hz]
4000
3500
3000
2500
2000
1500
1000
500
0
0
5
10
15
20
tiempo [s]

Aplicaciones STFT

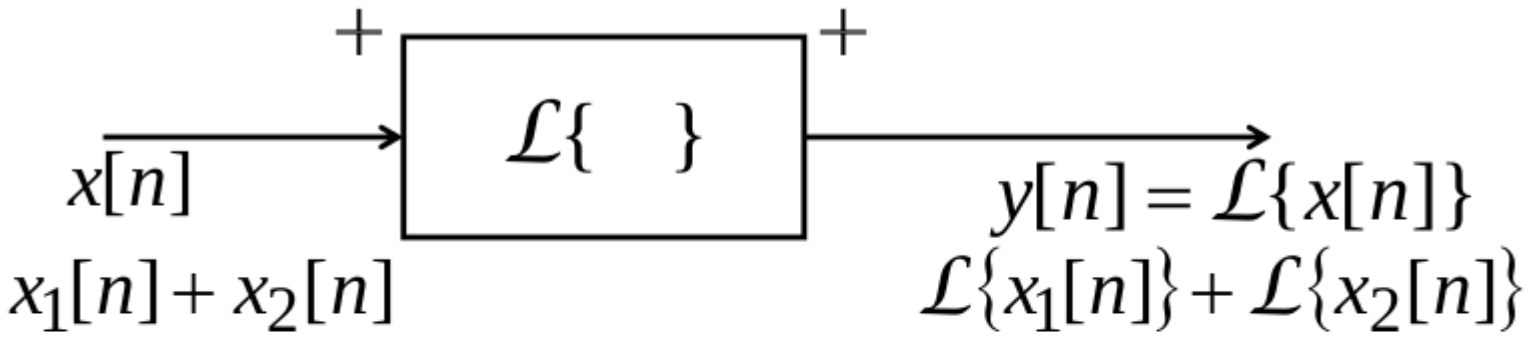
- Audio Digital
  - MPEG 2 – layer III (MP3)
  - AAC – Advanced Audio Coding (iTunes)
  - Dolby Digital AC-3
    - Audio 5.1 (Audio Cinema, DVD, Bluray, Netflix)
  - DTS (Digital Theater Sound)
    - Audio 5.1 (Audio Cinema, DVD, Bluray).

Cepstrum

- Superposición
- Sistema Homomórfico por Convolución
- Cepstrum (Definición basada en DTFT)
- Separación fuente-sistema – Ejemplo 1
- Remoción de “Eco” - Ejemplo 02
- Codificación – Ejemplo 03

Cepstrum

- Superposición

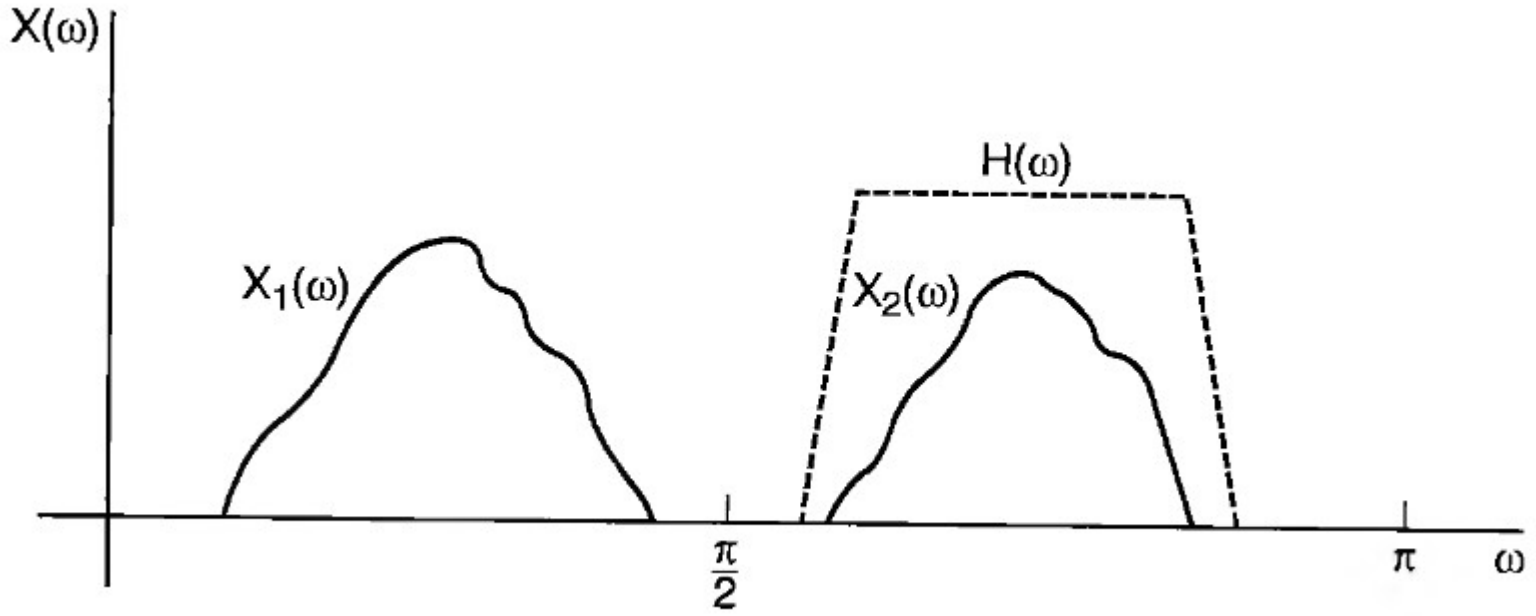


$$x[n] = ax_1[n] + bx_2[n]$$

$$y[n] = \mathcal{L}\{x[n]\} = a\mathcal{L}\{x_1[n]\} + b\mathcal{L}\{x_2[n]\}$$

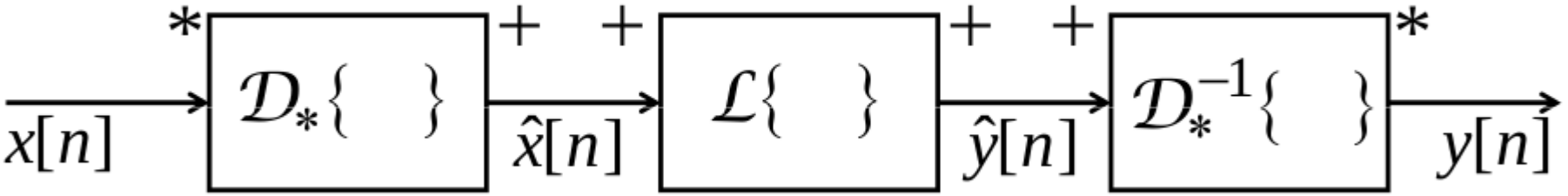
Cepstrum

- Separación



Cepstrum

- Sistema Homomórfico



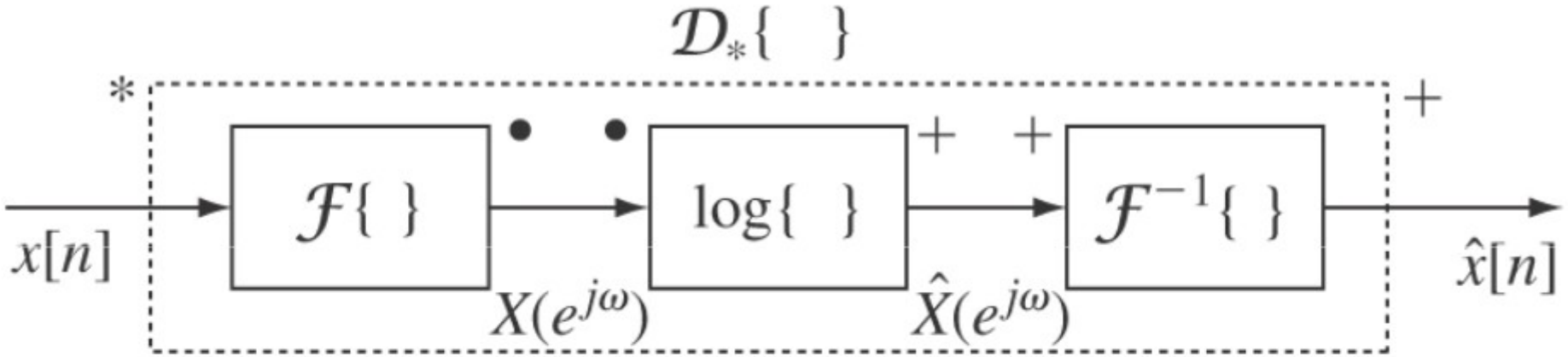
$\mathcal{D}_*\{\quad\}$: Characteristic system for convolution

Cepstrum

$$\begin{aligned}\hat{x}[n] &= \mathcal{D}_*\{x[n]\} = \mathcal{D}_*\{x_1[n] * x_2[n]\} \\ &= \mathcal{D}_*\{x_1[n]\} + \mathcal{D}_*\{x_2[n]\} \\ &= \hat{x}_1[n] + \hat{x}_2[n]\end{aligned}$$

$$\begin{aligned}\mathcal{D}_*^{-1}\{\hat{y}[n]\} &= \mathcal{D}_*^{-1}\{\hat{y}_1[n] + \hat{y}_2[n]\} \\ &= \mathcal{D}_*^{-1}\{\hat{y}_1[n]\} * \mathcal{D}_*^{-1}\{\hat{y}_2[n]\} \\ &= y_1[n] * y_2[n] = y[n]\end{aligned}$$

Cepstrum (DTFT)

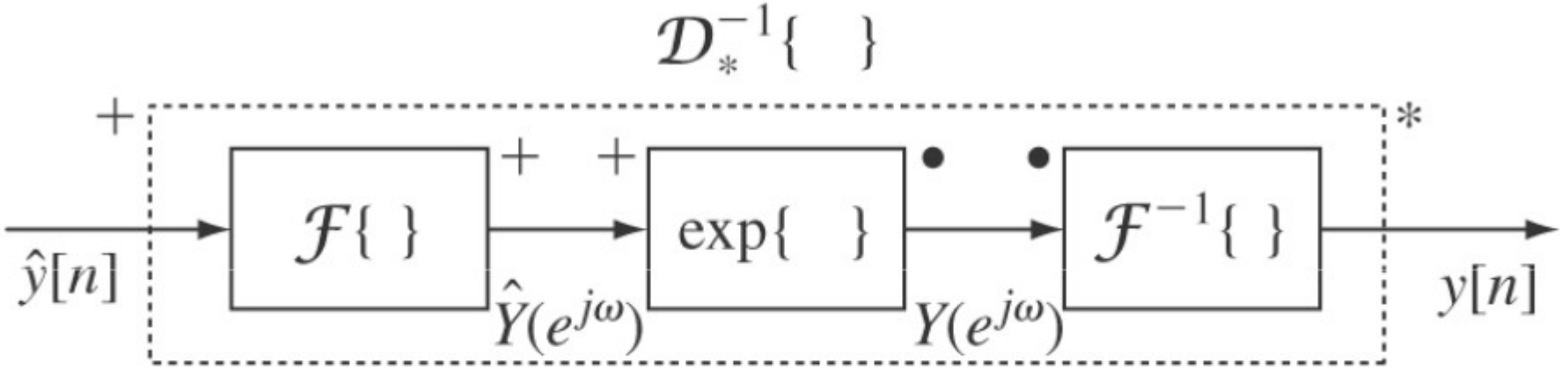


$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n}$$

$$\hat{X}(e^{j\omega}) = \log\left[X(e^{j\omega})\right] = \log\left|X(e^{j\omega})\right| + j\arg\left[X(e^{j\omega})\right]$$

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

Cepstrum (DTFT)

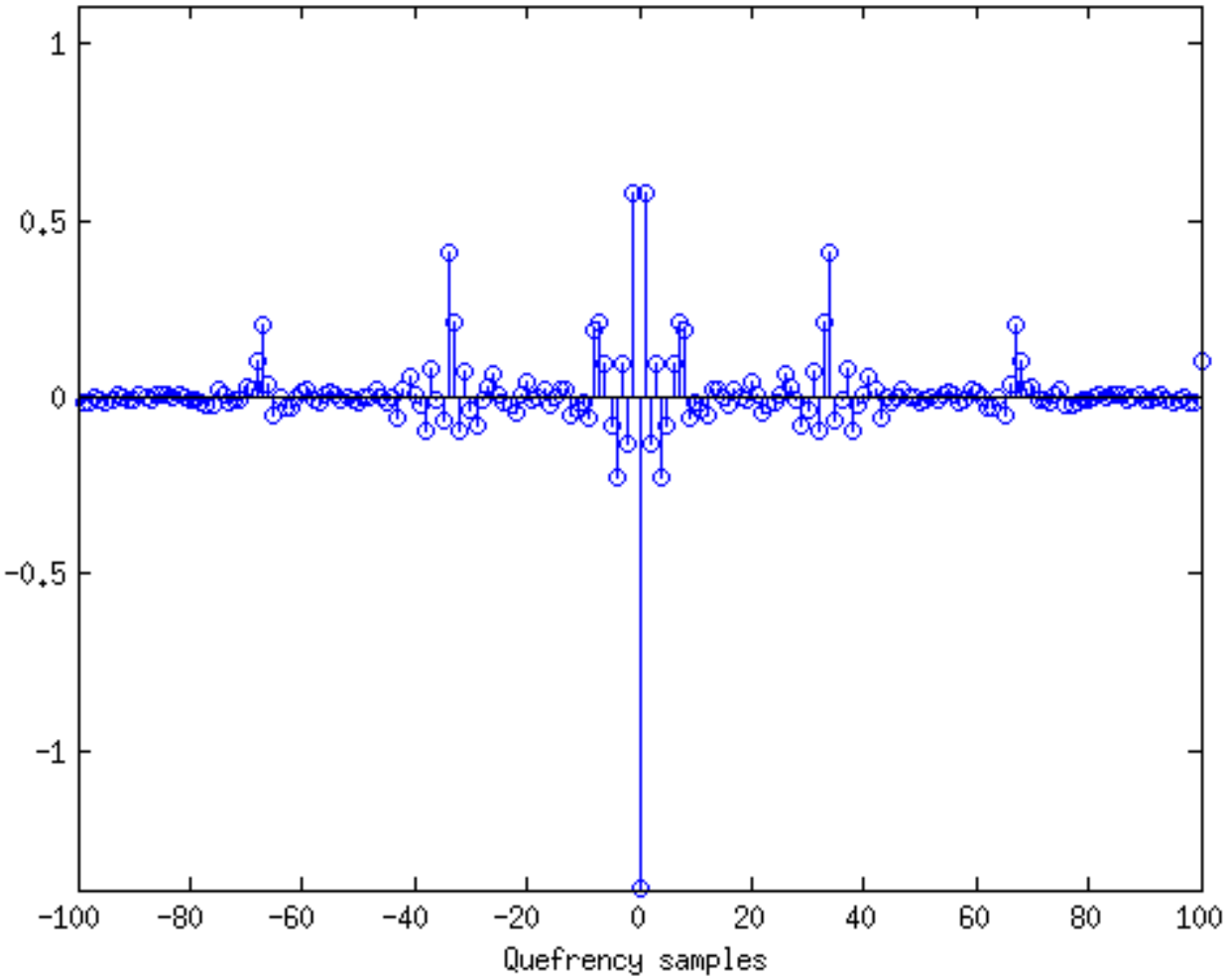


$$\hat{Y}(e^{j\omega}) = \sum_{n=-\infty}^{\infty} \hat{y}[n] e^{-j\omega n}$$

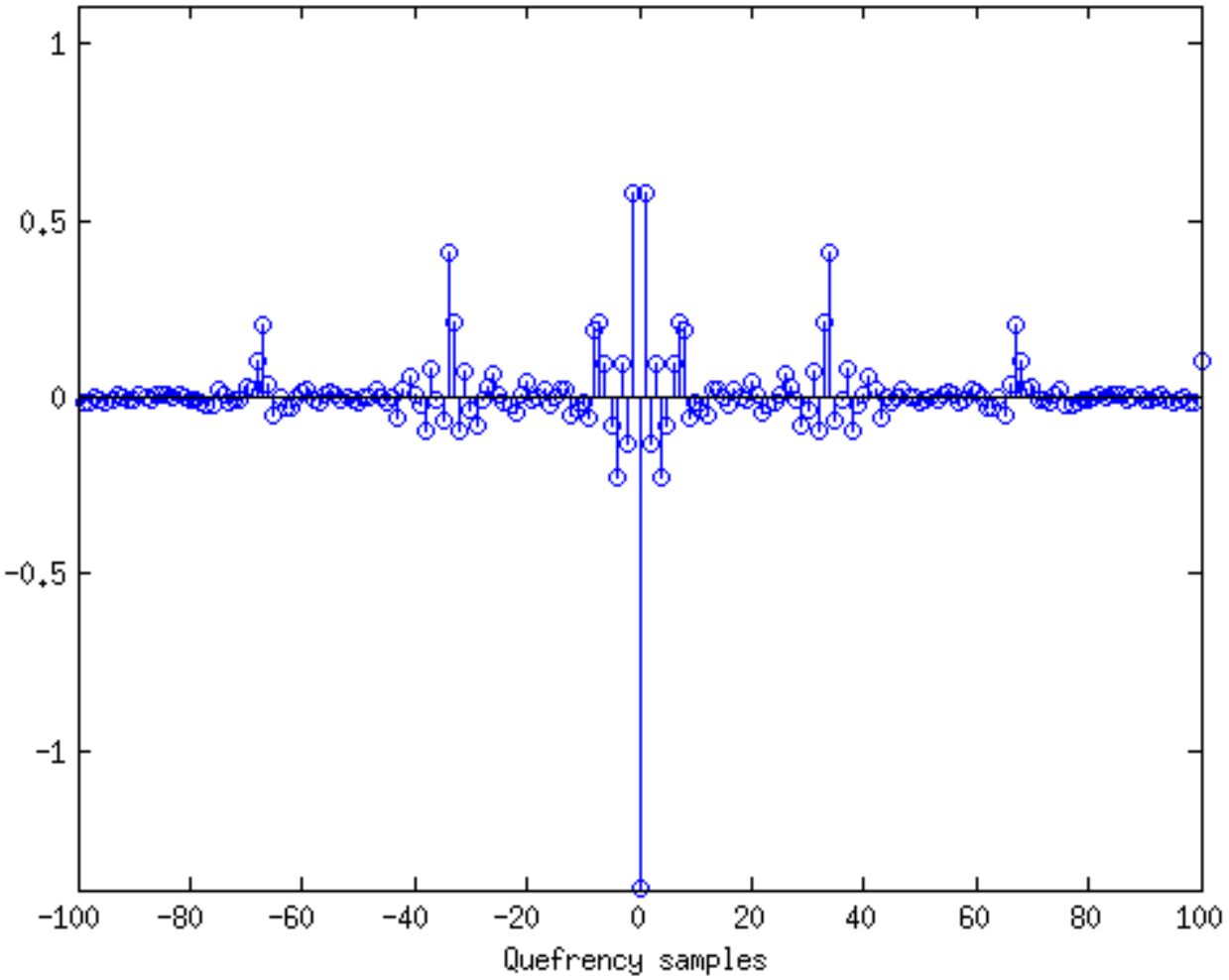
$$Y(e^{j\omega}) = \exp\left[\hat{Y}(e^{j\omega})\right]$$

$$y[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} Y(e^{j\omega}) e^{j\omega n} d\omega$$

Cepstrum

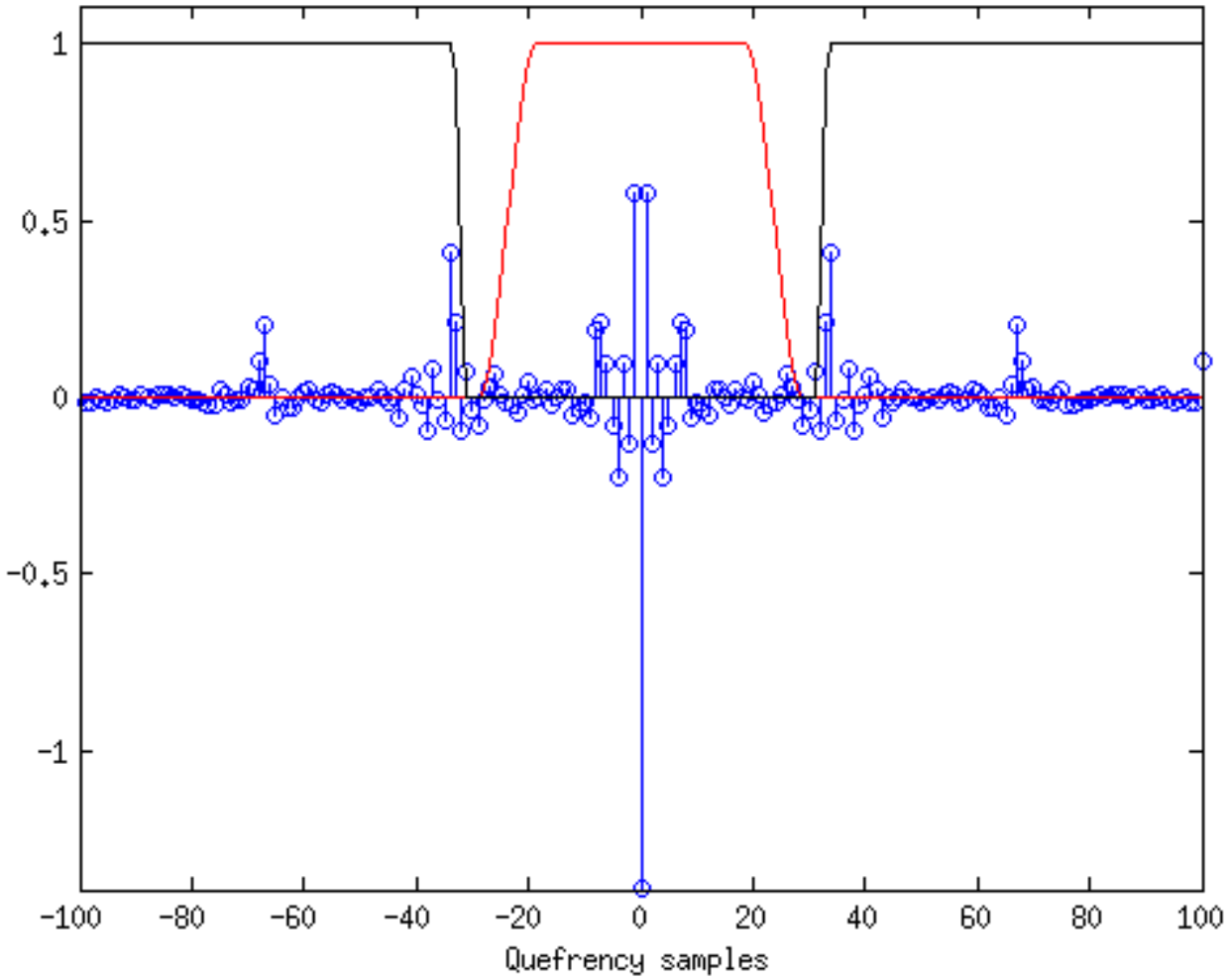


Cepstrum



Cepstrum

- Liftering (Cepstrum real)



Cepstrum

Terminología

Spectrum → Cepstrum

Analysis → Alanysis

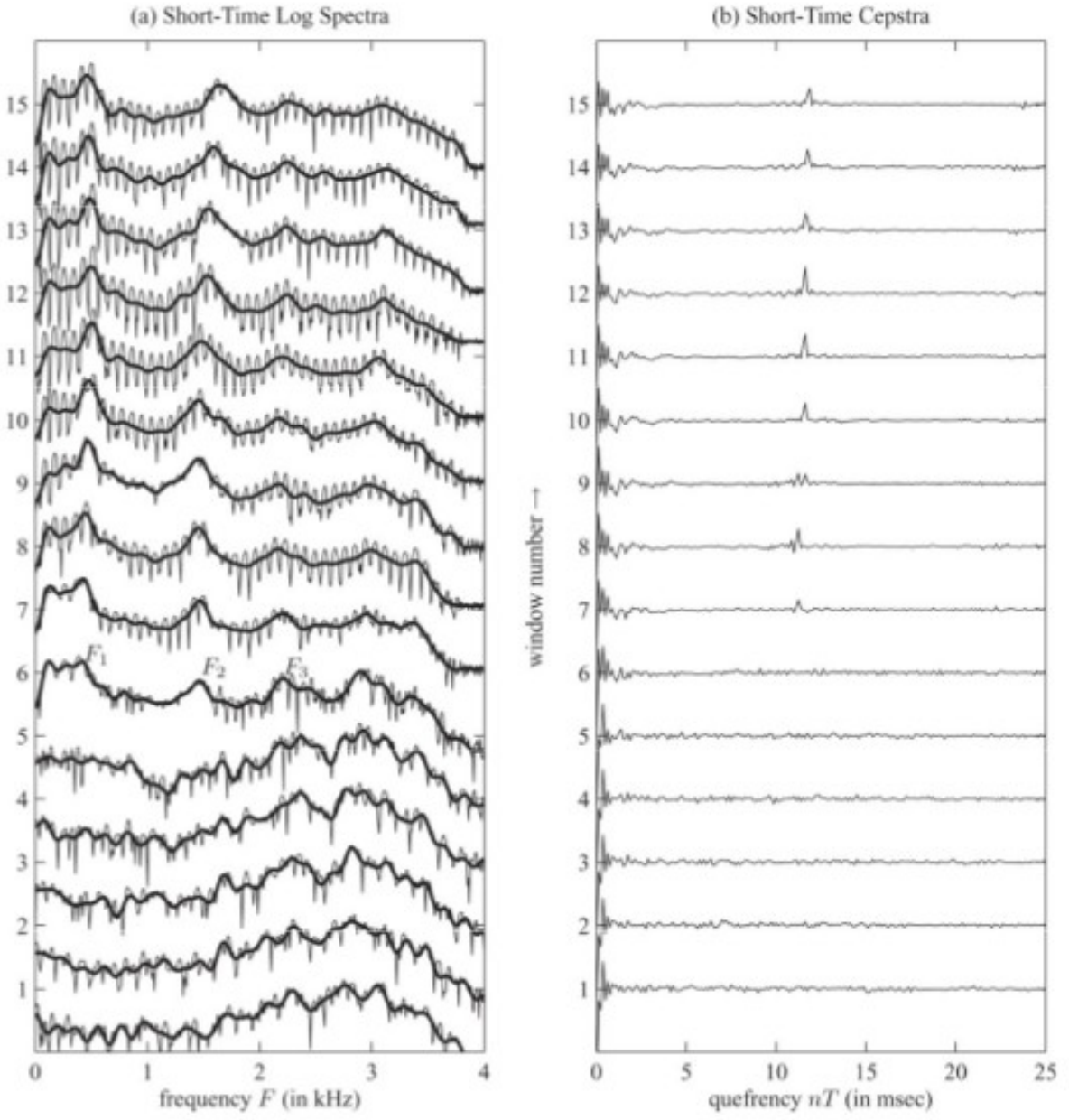
Filtering → Liftering

Frequency → Quefrency

Harmonic → Rahmonic

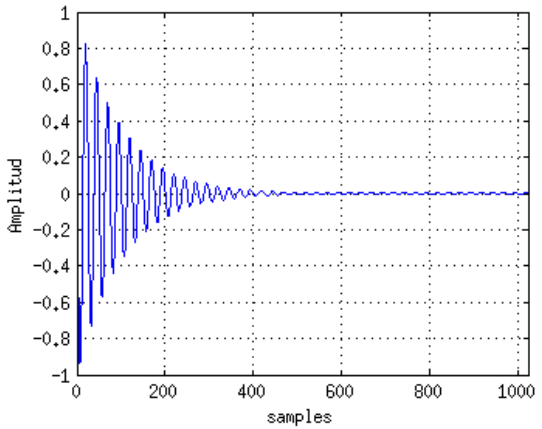
Cepstrum: Ejemplos

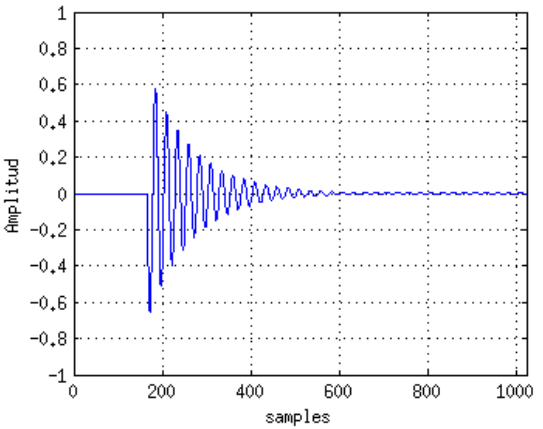
- Pitch
- System

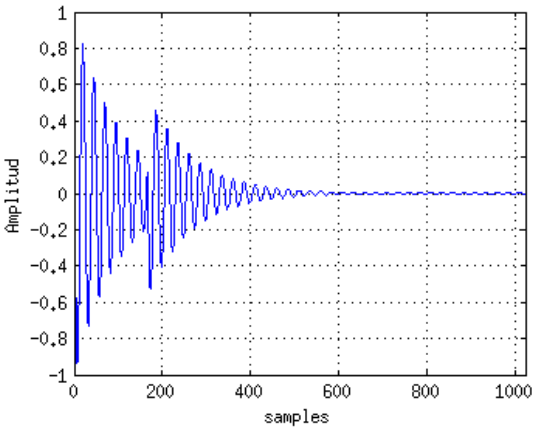


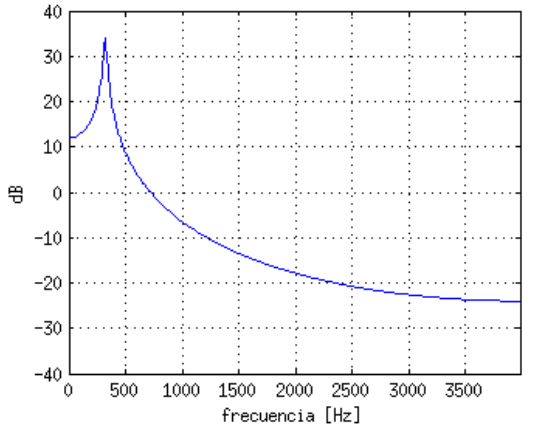
Cepstrum: Ejemplos

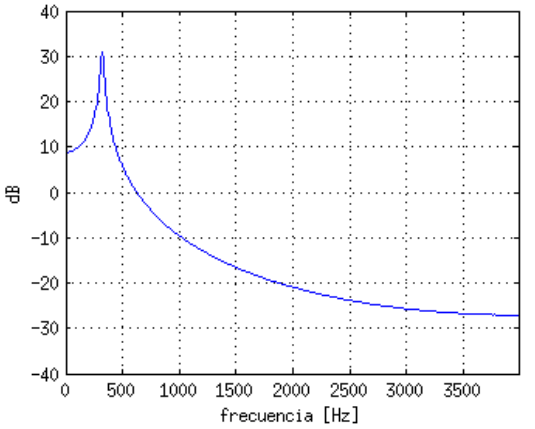
- Remoción de Eco: Ver Matlab...

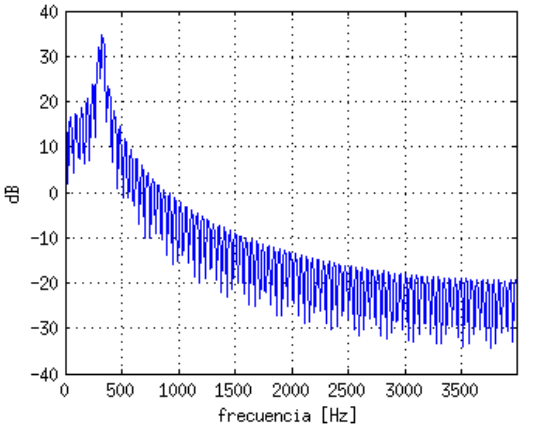






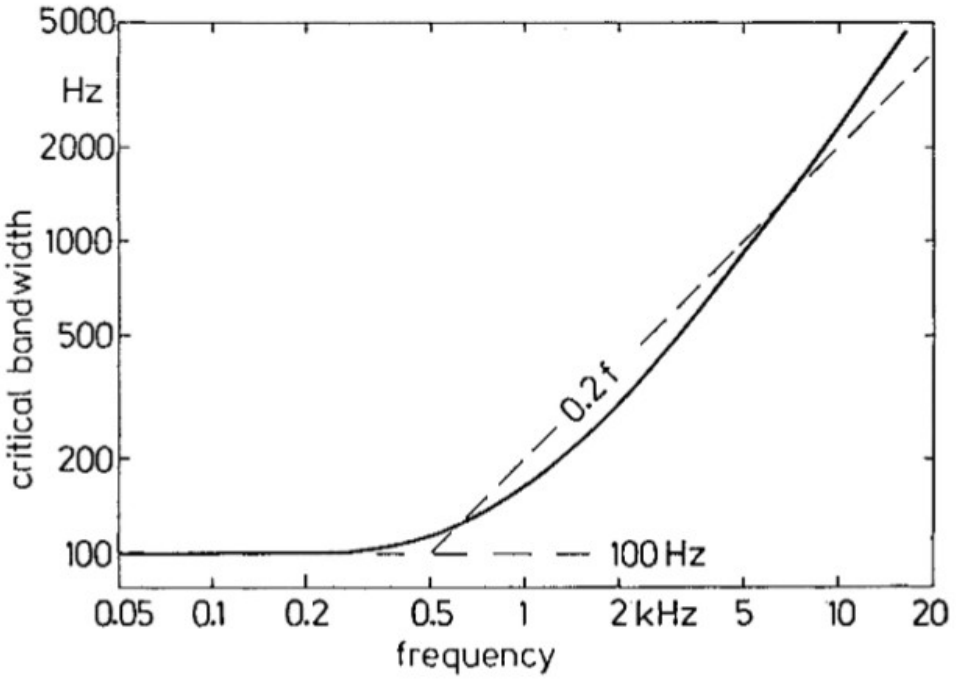




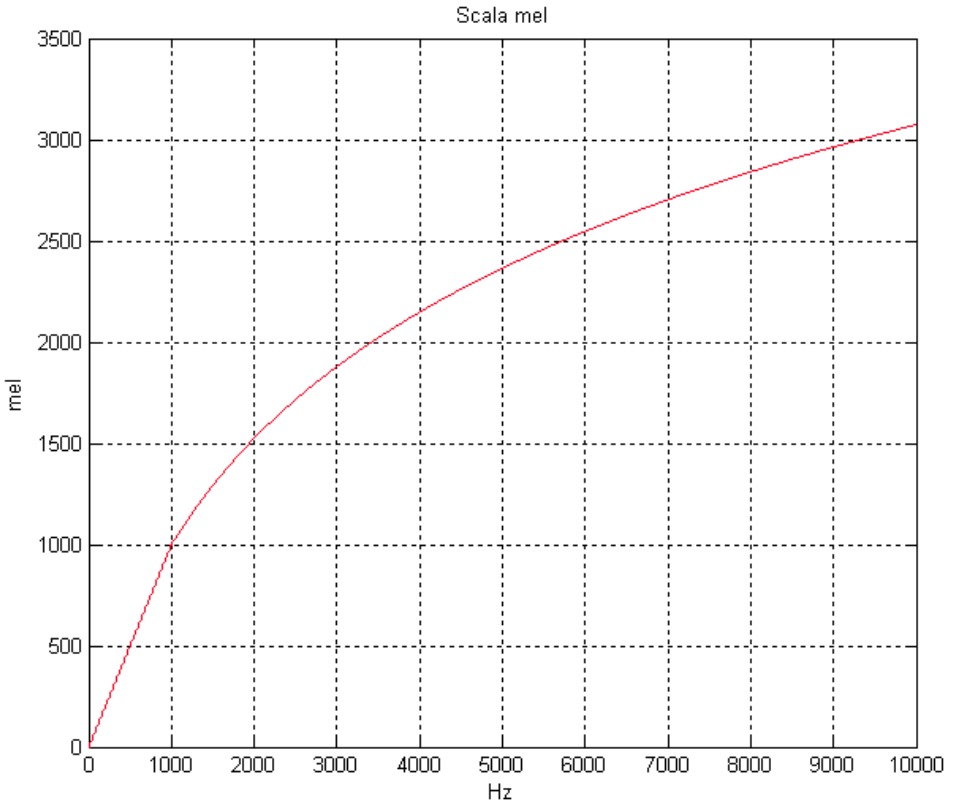


Cepstrum: Ejemplos

- Perceptual Coding



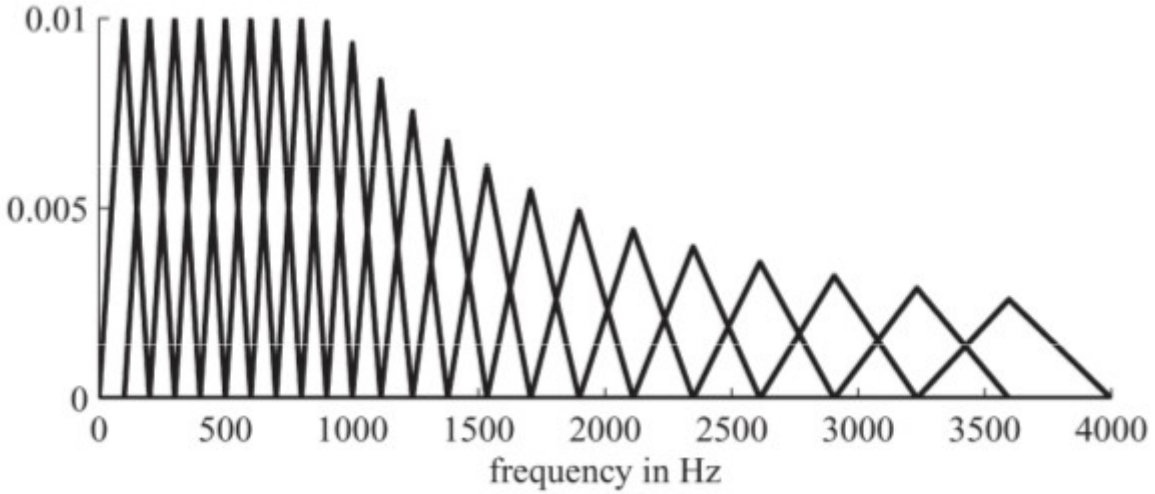
Frecuencia v/s Variación BW



Frecuencia v/s Escala Mel

Cepstrum: Ejemplos

- MFCC: Mel Frequency Cepstral Coefficients



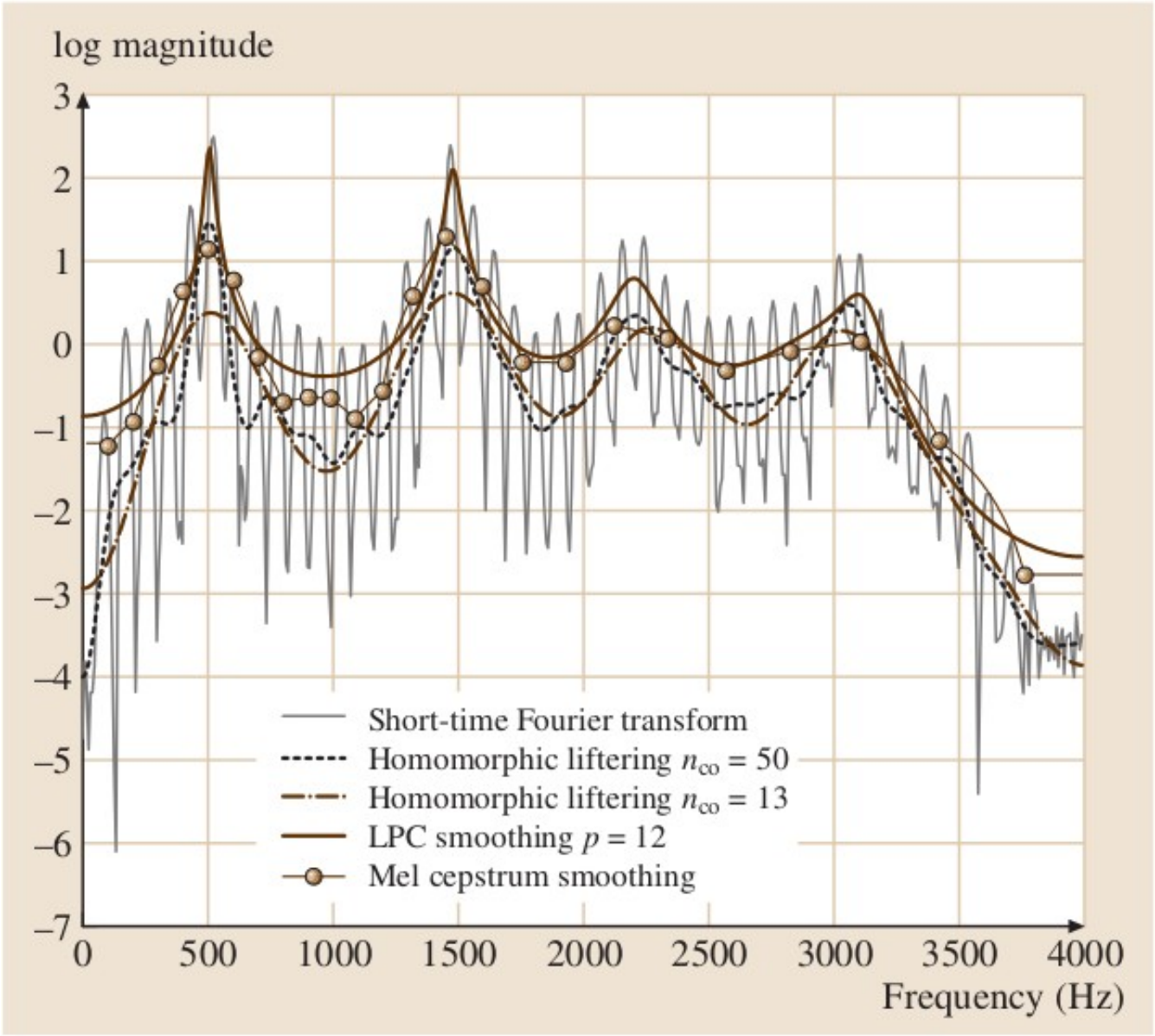
$$\mathrm{MF}_m[r] = \frac{1}{A_r} \sum_{k=L_r}^{U_r} |V_r[k] X_m[k]|^2$$

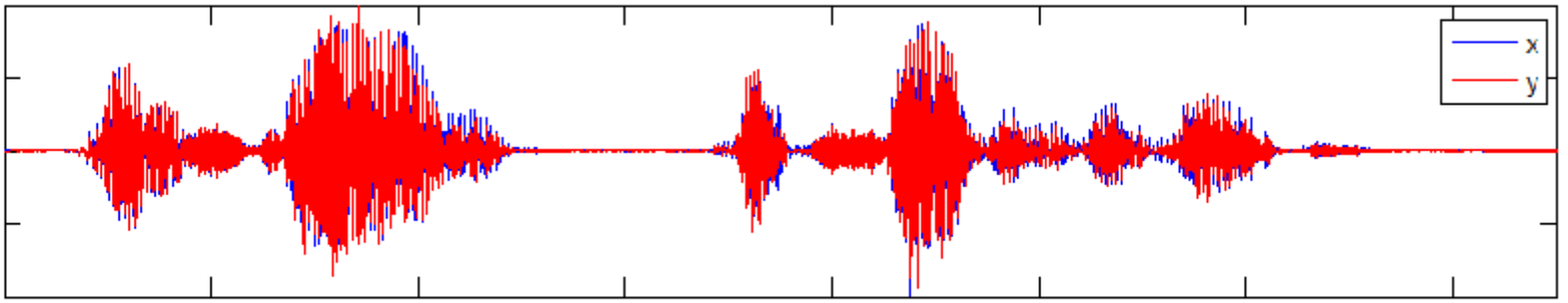
$$A_r = \sum_{k=L_r}^{U_r} |V_r[k]|^2$$

$$\mathrm{mfcc}_m[n] = \frac{1}{R} \sum_{r=1}^{R} \log(\mathrm{MF}_m[r]) \cos\left[\frac{2\pi}{R} \left(r + \frac{1}{2} \right) n \right], \quad n = 1, 2, \ldots, N_{\mathrm{mfcc}}$$

Típicamente, mfcc con DCT, Nmfcc=13 y R= 24 bandas para fs=8000 Hz.
Gran reducción (compresión) de datos!!!

Cepstrum: Ejemplos





ELO-313: Procesamiento Digital de Señales con Aplicaciones.
Primer Semestre 2013 - UTFSM

Profesor: Dr. Matías Zañartu.

Introducción a la Voz Humana

Presentado por: Víctor Espinoza.