UNIVERSIDAD DE GRANADA

PLAN DE ESTUDIOS: DIPLOMADO EN LOGOPEDIA

PROCESAMIENTO DE VOZ

Ángel de la Torre Vega
Dpto. Teoría de la Señal, Telemática y Comunicaciones

ORGANIZACIÓN DE LA ASIGNATURA

ORGANIZACIÓN DE LA ASIGNATURA

- Asignatura: **PROCESAMIENTO DE VOZ**
- Titulación: Diplomado en Logopedia (3er curso, 2o cuatrimestre)
- Tipo: Optativa
- Profesores: Ángel de la Torre (Dpto. TSTC, ETSIIT, 2a planta, desp. 22)
  Diego Pablo Ruiz (Dpto. FA, Facultad de Ciencias)
  Artur Schmitt (Dpto. FA, Facultad de Ciencias)
- Dpto: Teoría de la Señal, Telemática y Comunicaciones (50%)
  Física Aplicada (50%)
- Horario:
  - Teoría: X y J de 19:00 a 20:00 en A.03
  - Prácticas: J de 20:00 a 21:00 en A.03

ORGANIZACIÓN DE LA ASIGNATURA

- Créditos: Total: 4.5 créditos
  - Teoría: 3 créditos (30 horas)
  - Prácticas: 1.5 créditos (15 horas)

- Prácticas:
  - Prácticas en aulas de ordenadores y laboratorio Fac. Ciencias

- Evaluación:
  - Examen final de teoría y cuestiones (67%)
  - Prácticas: trabajo en aula de ordenadores/laboratorio y memoria de prácticas (33%)

CONTENIDOS (1a parte: 15 horas)

1. INTRODUCCIÓN
2. FUNDAMENTOS DE PROCESAMIENTO DE SEÑAL
3. MODELO DIGITAL DE PRODUCCIÓN DE VOZ
4. REPRESENTACIÓN DE LA SEÑAL DE VOZ
5. ANÁLISIS DE SEÑALES DE VOZ
6. SÍNTESIS Y CODIFICACIÓN DE VOZ
7. RECONOCIMIENTO DE VOZ Y RECONOCIMIENTO DE LOCUTORES
8. PROCESAMIENTO DE LA SEÑAL DE VOZ EN SISTEMAS DE AYUDA A LA AUDICIÓN
9. APLICACIONES DE LAS TECNOLOGÍAS DEL HABLA EN EL DIAGNÓSTICO, TRATAMIENTO Y SEGUIMIENTO LOGOPÉDICO

Tema 1: INTRODUCCIÓN

1.1.- La señal de voz.

1.2.- El procesamiento de voz en logopedia.

1.3.- Objetivos de la asignatura.

Tema 2: FUNDAMENTOS DE PROCESAMIENTO DE SEÑAL

2.1.- Introducción.

2.2.- Señales.

2.3.- La transformada de Fourier.

2.4.- Propiedades de la transformada de Fourier.

2.5.- El teorema de muestreo.

2.6.- Señales analógicas y señales digitales.

2.7.- La transformada discreta de Fourier.

2.8.- Procesamiento digital de señales.

Tema 3: MODELO DIGITAL DE PRODUCCIÓN DE VOZ

3.1.- Introducción.

3.2.- Producción de la voz.

3.3.- Resonancias del tracto vocal.

3.4.- Modelo excitación – filtrado.

3.5.- Evolución temporal de los parámetros del modelo.

3.6.- Caracterización de los sonidos de voz:

- Tono, timbre, intensidad y duración.
- Frecuencia fundamental, formantes, evolución temporal.
- Representación espectral de tiempo corto.

Tema 4: REPRESENTACIÓN DE LA SEÑAL DE VOZ

4.1.- Introducción.

4.2.- Características de la señal de voz.

4.3.- Representación de la forma de onda.

4.4.- Energía de tiempo corto.

4.5.- Tasa promedio de cruces por cero.

4.6.- Función de autocorrelación de tiempo corto.

4.7.- Estimación del tono fundamental.

4.8.- Análisis de Fourier de tiempo corto. Espectrograma (WB y NB).

4.9.- Linear Prediction Coding: Análisis LPC.

4.10.- Análisis basado en banco de filtros.

4.11.- Procesamiento homomórfico. Cepstrum (FFT, LPC y MFCC).

Tema 5: ANALISIS DE SEÑALES DE VOZ

5.1.- Introducción.

5.2.- La forma de onda.

5.3.- Las vocales.

5.4.- Consonantes estacionarias sonoras y sordas.

5.5.- Consonantes no estacionarias.

5.6.- Coarticulación.

5.7.- Variabilidad.

5.8.- La señal de voz en presencia de ruido:

- Ruido blanco y ruido coloreado
- Ruido no estacionario
- Detección de actividad de voz

Tema 6: SÍNTESIS Y CODIFICACIÓN DE VOZ

6.1.- Introducción.

6.2.- Codificación y decodificación de voz.

6.3.- Síntesis de voz.

6.4.- Manipulación de la señal de voz.

Tema 7: RECONOCIMIENTO DE VOZ Y RECONOCIMIENTO DE LOCUTORES

7.1.- Introducción.

7.2.- Reconocimiento automático de voz:

- Problemas asociados al reconocimiento de voz.
- Representación de la voz: el front-end.
- Modelado acústico: GMMs, HMMs y ANNs.
- Modelado del lenguaje: vocabulario y gramática.
- El sistema de diálogo.

7.3.- Reconocimiento automático de locutor:

- Problemas asociados al reconocimiento de locutores
- Reconocimiento, identificación y verificación de locutor. Distintos enfoques del problema.
- Representación de la voz y modelado del locutor.

Tema 8: PROCESAMIENTO DE LA SEÑAL DE VOZ EN SISTEMAS DE AYUDA A LA AUDICIÓN

8.1.- Introducción.

8.2.- Audífonos:

- Amplificación.
- Bancos de filtros.
- Compresión.
- Control automático de ganancia y reducción de ruido.
- Realimentación acústica.

8.3.- Implantes cocleares:

- Funcionamiento del implante coclear.
- Procesamiento de la señal en un implante coclear.
- Posibilidades y limitaciones de los implantes cocleares.
- Programación de los implantes cocleares.
- Percepción del sonido con implantes cocleares.

Tema 9: APLICACIONES DE LAS TECNOLOGÍAS DEL HABLA EN EL DIAGNÓSTICO, TRATAMIENTO Y SEGUIMIENTO LOGOPÉDICO

9.1.- Introducción.

9.2.- Herramientas de análisis de la voz.

9.3.- Herramientas basadas en reconocimiento de voz.

Bibliografía recomendada

- L.R. Rabiner y R.W. Schafer. “Digital Processing of Speech Signals”. Prentice Hall, 1978.
- S. Furui. “Advances in Speech Signal Processing”. Dekker, 1992.
- S.V. Vaseghi. “Advanced Digital Signal Processing and Noise Reduction”. John Wiley and Sons, 2000.
- J.L. Flanagan. “Speech Analysis, Synthesis and Perception”. Springer Verlag, 1972.
- A. Quilis, J.A. Fernandez. “Curso de fonética y fonología españolas”. CSIC, 1989.
- A. de la Torre, A.M. Peinado, A.J. Rubio. “Reconocimiento Automático de Voz en Condiciones de Ruido”. Universidad de Granada, 2001.
- Revistas: Speech Communication, IEEE Trans. Speech and Audio Processing, Computer Speech and Language.

TEMA 1

INTRODUCCIÓN

Tema 1: INTRODUCCIÓN

1.1.- La señal de voz.

1.2.- El procesamiento de voz en logopedia.

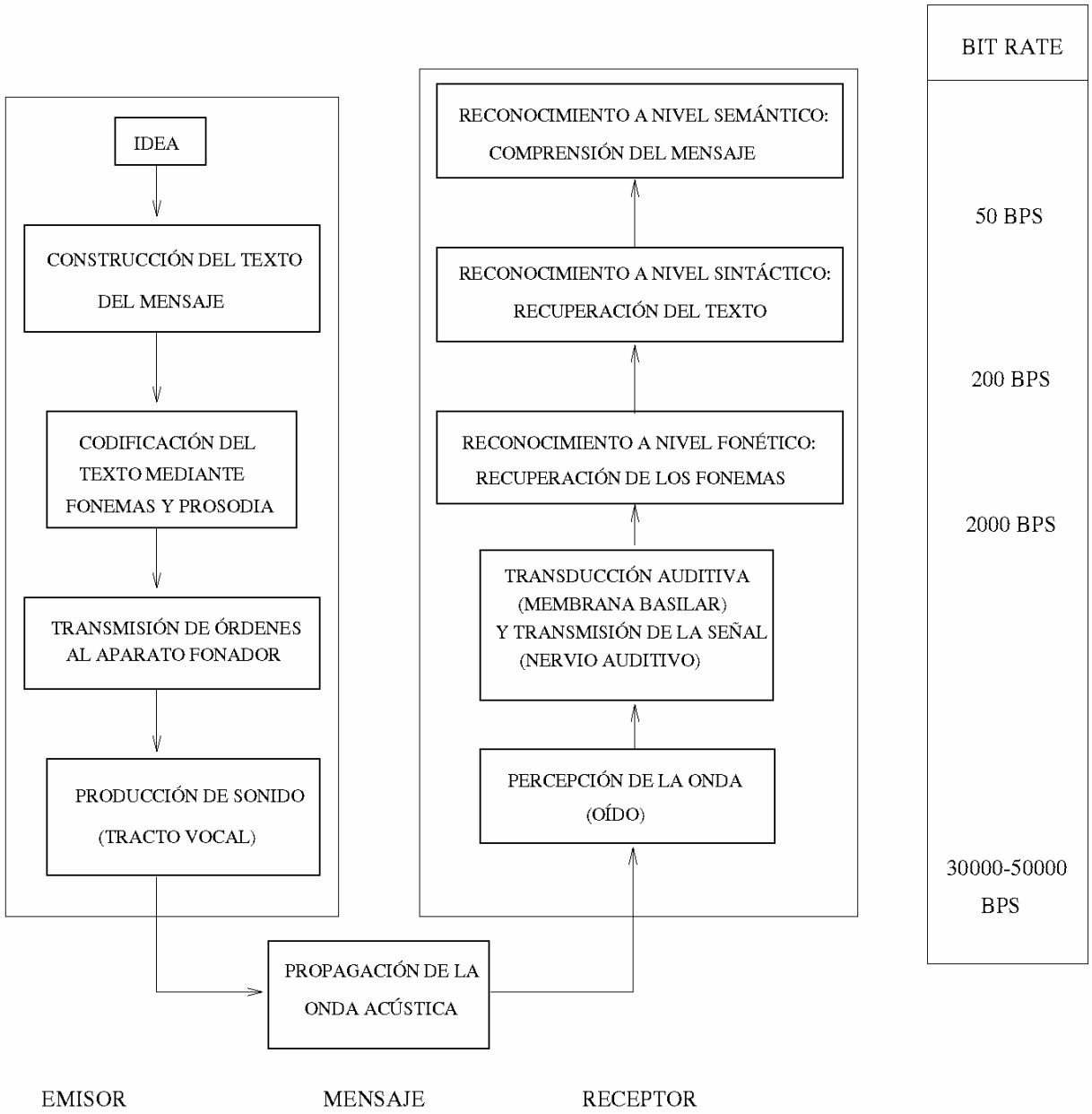
1.3.- Objetivos de la asignatura.

1.1.- La señal de voz

- **¿Qué es la voz?:**
  - La voz es una **onda de presión**: $P(x, y, z, t) = P(\mathbf{r}, t)$
  - Producción: aparato fonador
  - Propagación (dispersión, difracción…)
  - Audición: se percibe y analiza en el oído; se procesa en el cerebro
- **Rango de intensidad:** 50 – 70 dB SPL
- **Rango de frecuencia:** 60 Hz – 6 kHz (telefonía: 350 Hz – 3.5 kHz)
- **La señal de voz:**
  - Señal eléctrica recogida por un micrófono (forma de onda): $A(t)$ (mV)
  - Señal digital (muestreo y cuantización) 64 kbps; … 86 kBps;
    - Frecuencia de muestreo (8 kHz; 22 kHz; 44,1 kHz)
    - Número de bits por muestra (8 bits; 12 bits; 16 bits)

- **Asignatura organizada en 2 partes:**
  - La onda de voz (2ª parte): Física Acústica
  - La señal de voz (1ª parte): Procesamiento de señales
- **Señal:**
  - Transmisión de información
  - Ruido
- **Información contenida en la señal de voz:**
  - Fonemas, sílabas, palabras
  - Frases, mensaje
  - Características suprasegmentales
  - Locutor
  - Patologías, vicios, acentos
  - Entorno acústico (ruido)

Transmisión de información en la comunicación oral



Procesamiento de voz (principales líneas)

- Análisis de voz
- Codificación y compresión de voz
- Síntesis de voz (conversión texto a voz)
- Reconocimiento automático de voz
- Reconocimiento y verificación de locutores
- Comprensión de voz y sistemas de diálogo
- Percepción de la voz
- Sistemas de ayuda a la audición

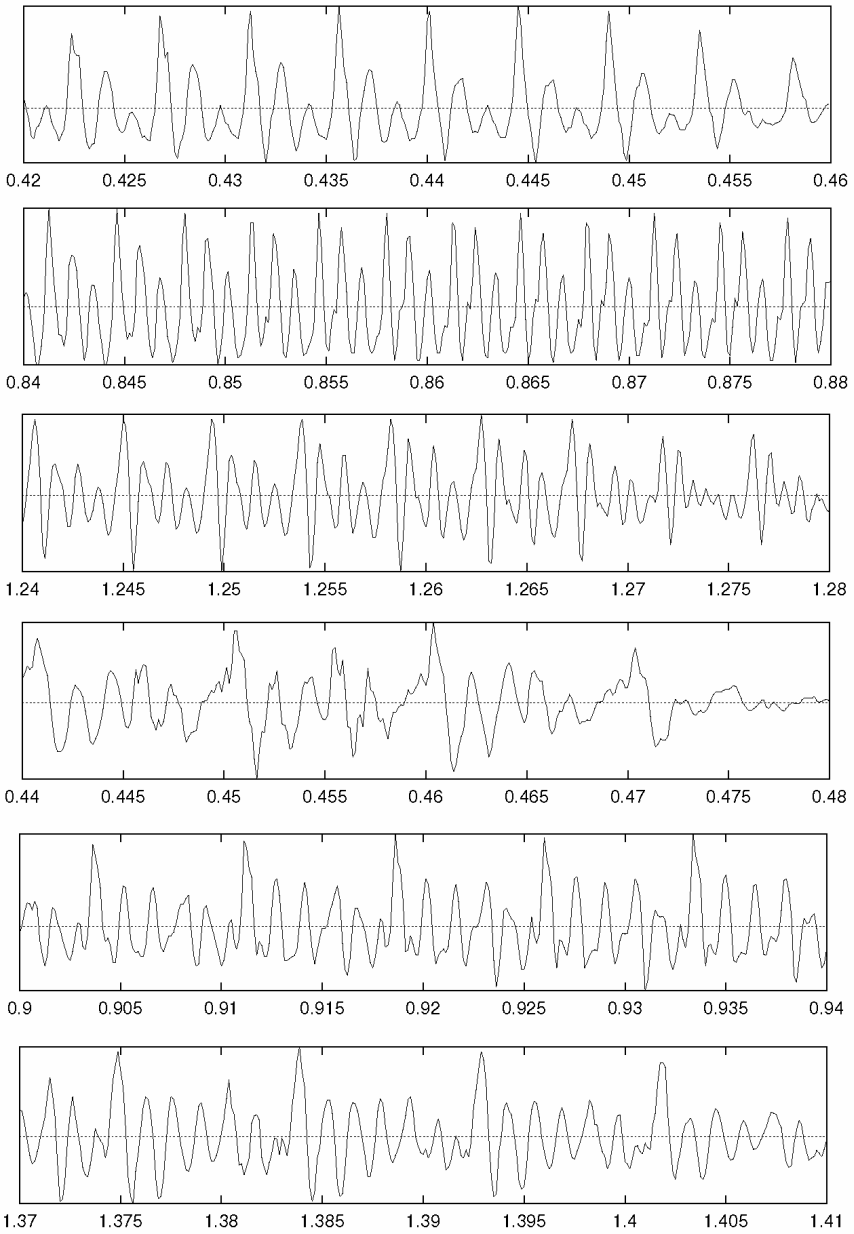
1.2.- El procesamiento de voz en logopedia

- **Tareas usuales en logopedia:**
  - (Re)habilitación en foniatría
  - (Re)habilitación en audición
  - Diseño y uso de material para evaluación
    - ¿Cómo evoluciona un determinado parámetro?
    - ¿Puede considerarse normal un determinado parámetro?
  - Diseño y uso de material para rehabilitación
- **El procesamiento de voz ayuda a:**
  - Entender la producción de la voz
  - Entender la percepción de la voz
  - Analizar y procesar señales de voz
  - Entender herramientas relacionadas con el procesamiento de voz

- **Herramientas relacionadas con el procesamiento de voz:**
  - Análisis de voz
  - Síntesis de voz
  - Reconocimiento de voz
  - Codificación de voz
- **Problemas del procesamiento de voz:**
  - Variabilidad:
    - Intra-locutor (estado de salud, de ánimo, velocidad, etc.).
    - Inter-locutor
    - Entorno de adquisición
  - Continuidad: concatenación y coarticulación
  - Información contenida en la voz muy redundante
  - Multi-interactividad entre niveles:
    - Nivel fonético
    - Características suprasegmentales
    - Nivel semántico; contexto; suplencia mental
  - Ruido: perturbación y efecto Lombard

Variabilidad de la señal de voz

- Arriba: Tres ejecuciones del fonema /a/ en la frase “voy a comprar pan” pronunciada por una mujer.
- Abajo: Tres ejecuciones del fonema /a/ en la frase “voy a comprar pan” pronunciada por un hombre.



1.3.- Objetivos de la asignatura

- **Objetivos globales:**
  - Conocer las características de la señal de voz
  - Conocer las operaciones de procesamiento de señal que se aplican a la señal de voz:
    - Principales técnicas de análisis
    - Aplicaciones de estas técnicas
  - Aplicaciones del procesamiento de voz en logopedia
- **Avances tecnológicos:**
  - Software de propósito general (MATLAB) para el análisis y procesamiento de señales de voz
  - Software de propósito específico (Dr. Speech) para el análisis y procesamiento de señales de voz
  - Existencia de otros paquetes relacionados con procesamiento de voz

- **El procesamiento de voz contribuye a resolver cuestiones como:**
  - ¿Qué relación hay entre la señal de voz y los fonemas?
  - ¿Qué relación hay entre la señal de voz y otras unidades o características?
  - ¿Qué herramientas de procesamiento de señal podemos (o debemos) usar para analizar la señal de voz?
  - ¿Cómo se manifiesta una patología de la voz en la señal?
  - ¿Cómo podemos ajustar un sistema de ayuda a la audición para optimizar la comprensión de la voz?
  - ¿Cómo podemos interpretar un error en la producción de un fonema?
  - ¿Cómo podemos interpretar un error en la detección o identificación de un fonema?

- **Organización del curso (parte de procesamiento de señales de voz):**
  - Tema 2: Fundamentos de procesamiento de señal (2 horas)
  - Tema 3: Modelo digital de producción de voz (1 hora)
  - Tema 4: Representación de la señal de voz (3 horas)
  - Tema 5: Análisis de señales de voz (3 horas)
  - Tema 6: Síntesis y codificación de voz (1 hora)
  - Tema 7: Reconocimiento de voz y reconocimiento de locutores (1 hora)
  - Tema 8: Procesamiento de la señal de voz en sistemas de ayuda a la audición (2 horas):
    - Audífonos
    - Implantes cocleares
  - Tema 9: Aplicaciones de las tecnologías del habla en el diagnóstico, tratamiento y seguimiento logopédico (1 hora).

TEMA 2

FUNDAMENTOS DE PROCESAMIENTO DE SEÑAL

Tema 2: FUNDAMENTOS DE PROCESAMIENTO DE SEÑAL

2.1.- Introducción.

2.2.- Señales.

2.3.- La transformada de Fourier.

2.4.- Propiedades de la transformada de Fourier.

2.5.- El teorema de muestreo.

2.6.- Señales analógicas y señales digitales.

2.7.- La transformada discreta de Fourier.

2.8.- Procesamiento digital de señales.

2.1.- Introducción

- **El principal propósito de la voz es la comunicación:**
  - La forma de onda contiene información
  - Teoría de la Información: Información contenida
  - Teoría de Señal: Cómo se transmite la información en la forma de onda

- **En la práctica, la representación de la voz está basada usualmente en la forma de onda:**
  - Modelos de producción
  - Procesamiento de señal

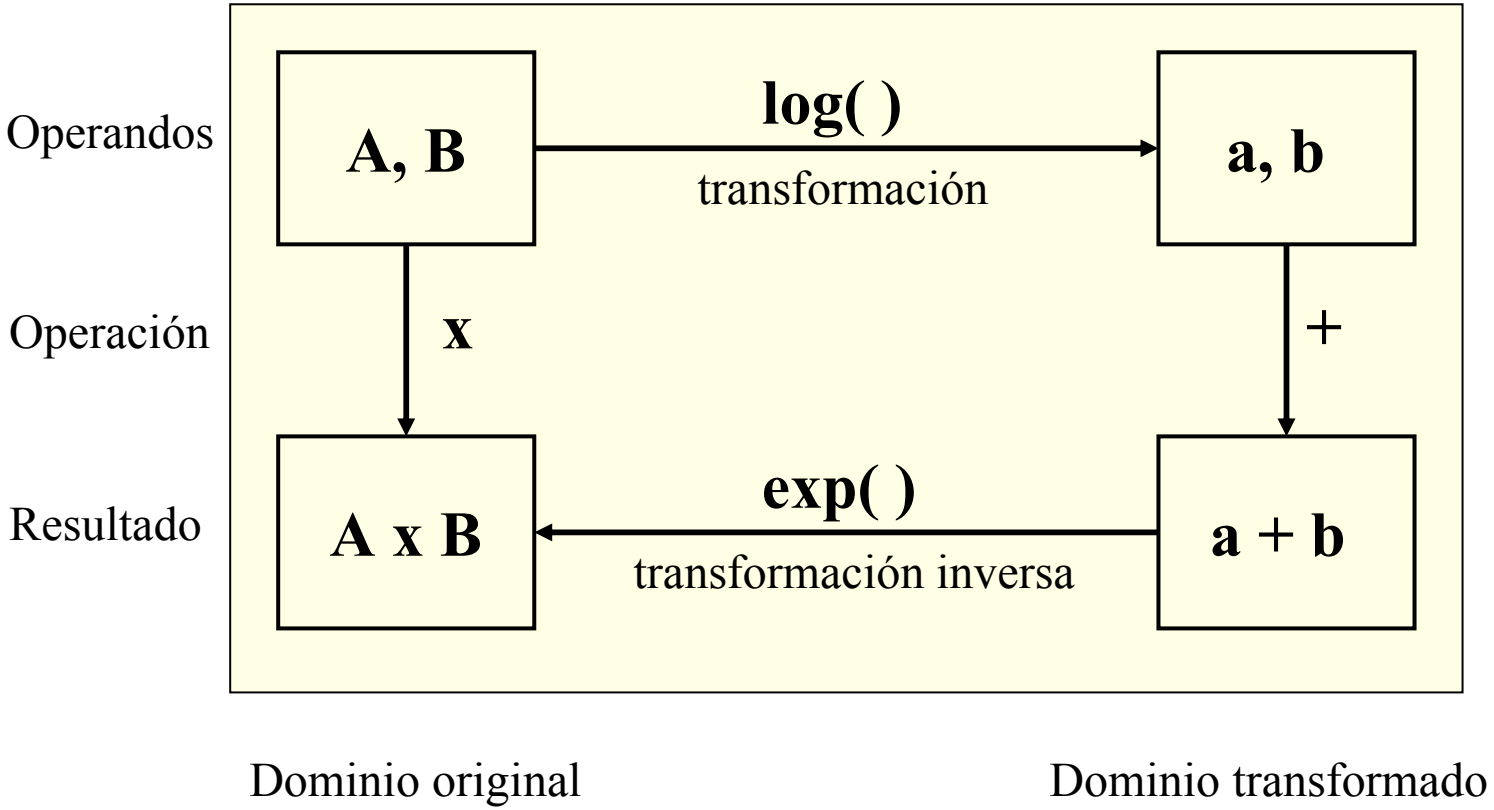
- **En este tema se revisan nociones básicas de procesamiento de señal:**
  - Concepto de señal
  - Representación de la señal en el dominio del tiempo y en el dominio de la frecuencia: la transformada de Fourier
  - Propiedades de la transformada de Fourier
  - Muestreo de señales: el teorema de muestreo
  - Representación digital de señales
  - La transformada discreta de Fourier: DFT y FFT
  - Procesamiento digital de señales

2.2.- Señales

- **CONCEPTO DE SEÑAL:**
  - UNA SEÑAL ES UNA VARIACIÓN DE UNA MAGNITUD QUE TRANSMITE UNA INFORMACIÓN
- **Tipos de señales:**
  - De una variable, de varias variables
  - Unidimensional, multidimensional
  - Discreta, continua, de variable discreta, de variable continua
- **Objetivo del procesamiento de señal: Comunicación eficiente:**
  - Codificación, transmisión, recepción, almacenamiento y representación de señales en sistemas de comunicación de forma eficiente y fiable
  - Extracción de información de señales ruidosas
- **Métodos de complejidad muy variada**: (no paramétricos, basados en modelos, bayesianos, etc.)

2.3.- La transformada de Fourier (FT)

- **Transformación de dominio:**
  - A veces una operación resulta más sencilla en un dominio transformado
  - Ejemplo: el producto resulta más sencillo en el dominio del logaritmo



- **La transformada de Fourier pasa del dominio del tiempo al dominio de la frecuencia:**
  - Cambio de representación: *g(t)* $\rightarrow$ *G(f)* $\quad (\omega = 2\pi f)$
  - Misma información en ambos dominios (es sólo un cambio de representación)
  - Existe la transformada inversa: *G(f)* $\rightarrow$ *g(t)*
  - Transforma una señal compleja *g(t)* en un espectro complejo *G(f)*

$$g(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} G(\omega)e^{j\omega t}d\omega \qquad G(\omega) = \int_{-\infty}^{\infty} g(t)e^{-j\omega t}dt$$

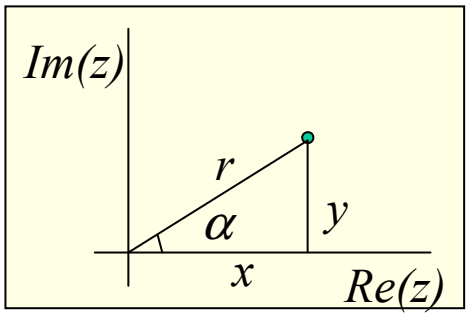
$$G(\omega) = \mathcal{F}[g(t)] \qquad g(t) = \mathcal{F}^{-1}[G(\omega)] \qquad g(t) \leftrightarrow G(\omega)$$

  - La transformada de Fourier descompone una señal en sus componentes senoidales

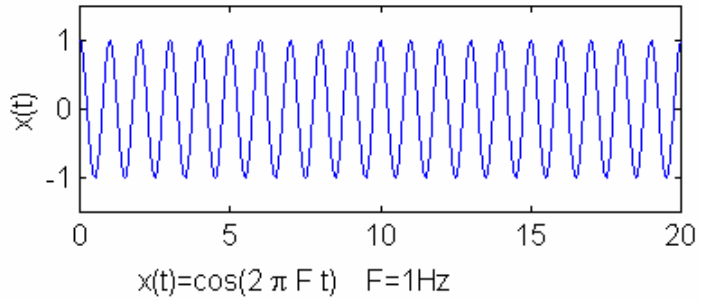
$$e^{j\alpha} = \cos(\alpha) + j\,\sin(\alpha)$$

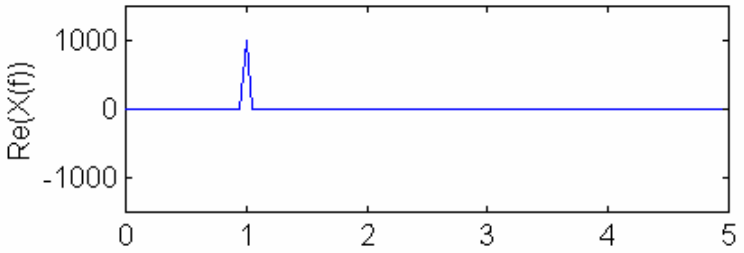
$z = x + j\,y$

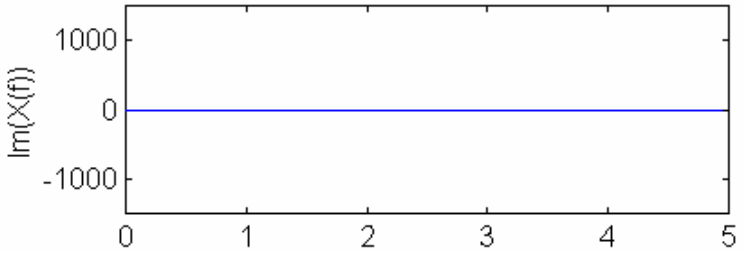
$j = (-1)^{1/2}$

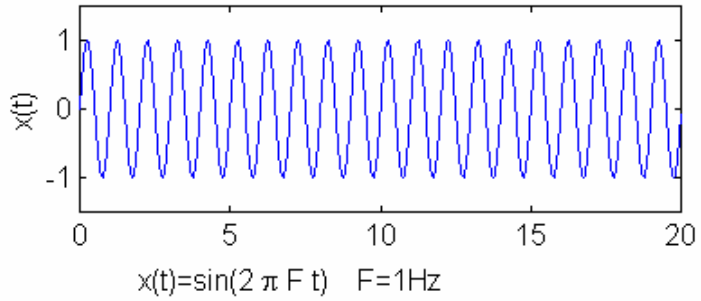


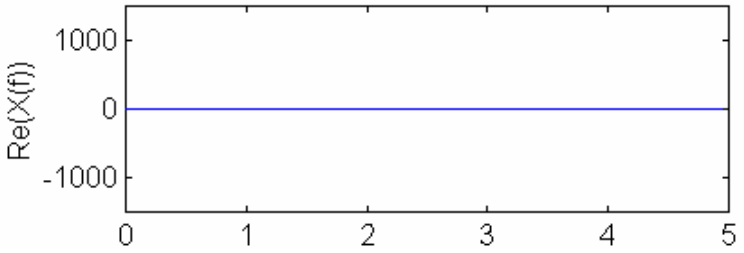
- **Transformada de una función coseno y una función seno:**

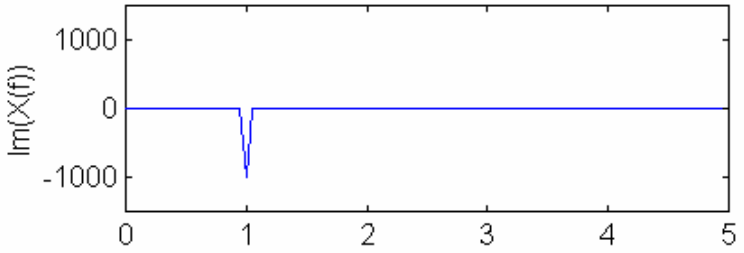




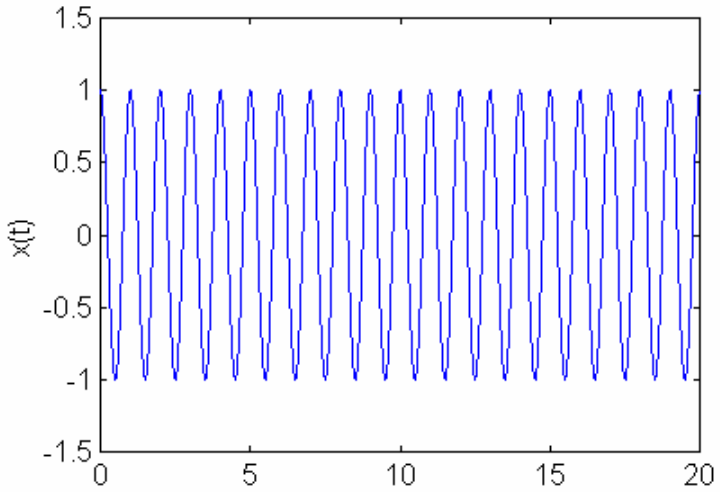


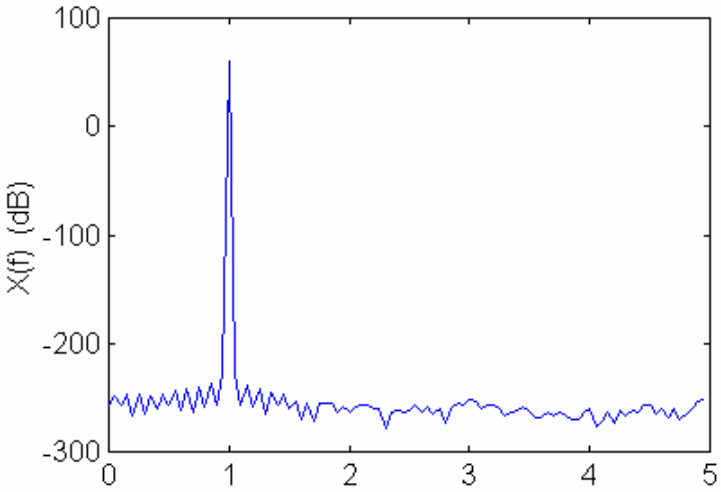


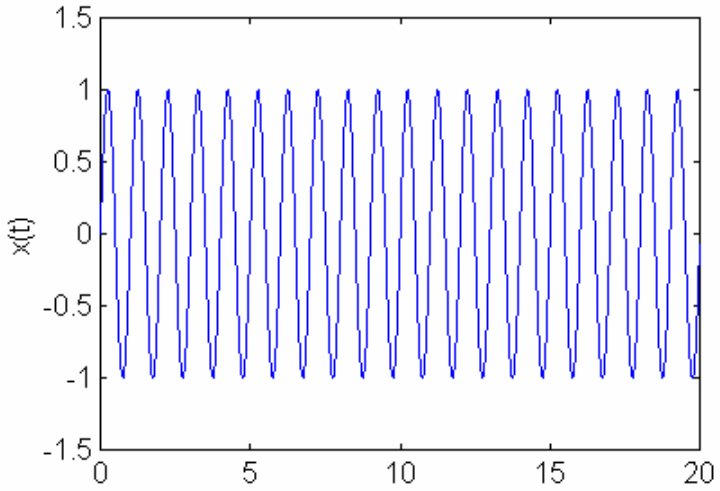


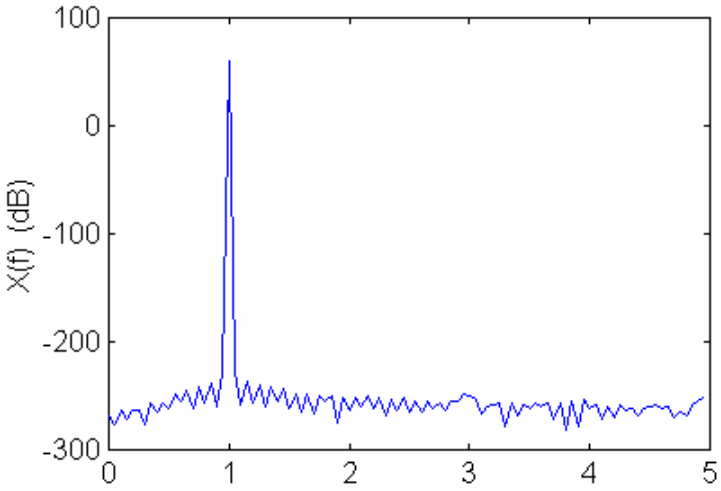


- **Espectro de potencia:**









2.4.- Propiedades de la transformada de Fourier

- **Utilidad de la transformada de Fourier:**
  - Descompone una señal en sus componentes de frecuencia
  - Propiedades:

1.- Linealidad: $g = k_1 g_1 + k_2 g_2 \Leftrightarrow G = k_1 G_1 + k_2 G_2$

2.- Dualidad o simetría: $G(t) \leftrightarrow 2\pi g(-\omega)$

3.- Escalado: $g(at) \leftrightarrow \frac{1}{|a|} G\left(\frac{\omega}{a}\right)$

4.- y 5.- Desplazamiento en t y en ω: $g(t - t_0) \leftrightarrow G(\omega) e^{-j\omega t_0}$ $\quad$ $g(t) e^{j\omega_0 t} \leftrightarrow G(\omega - \omega_0)$

6.- y 7.- Derivada en t y en ω: $\frac{dg(t)}{dt} \leftrightarrow j\omega G(\omega)$ $\quad$ $-jtg(t) \leftrightarrow \frac{dG(\omega)}{d\omega}$

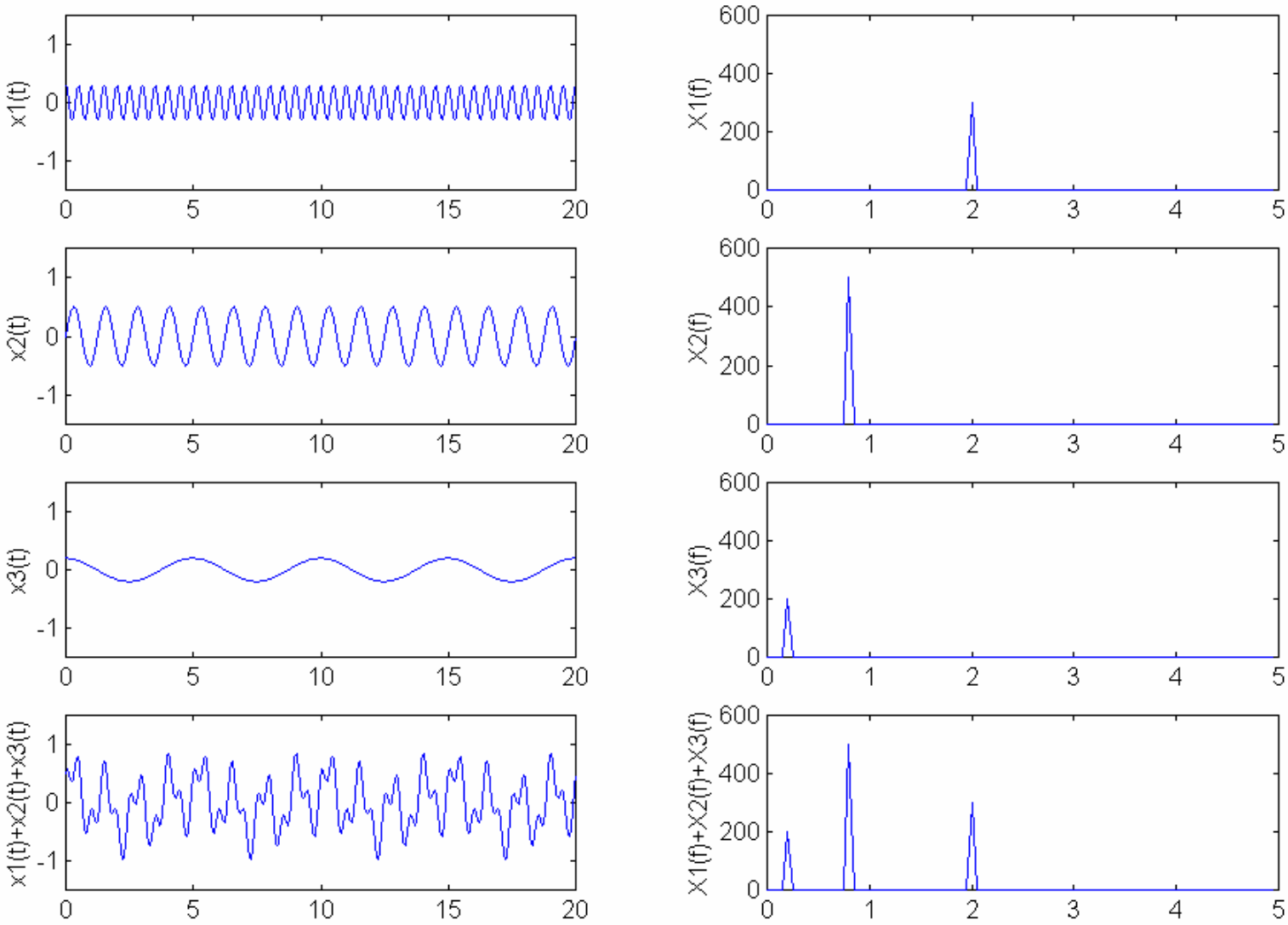
8.- Convolución: $g_1(t) * g_2(t) \leftrightarrow G_1(\omega) G_2(\omega)$ $\quad$ $g_1(t) g_2(t) \leftrightarrow \frac{1}{2\pi} G_1(\omega) * G_2(\omega)$

9.- Partes par/impar real/imag:

PAR-REAL $\leftrightarrow$ PAR-REAL
PAR-IMAG $\leftrightarrow$ PAR-IMAG
IMPAR-REAL $\leftrightarrow$ IMPAR-IMAG
IMPAR-IMAG $\leftrightarrow$ IMPAR-REAL

10.- Teorema de Parseval (señales reales): $E_g = \int_{-\infty}^{\infty} g^2(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |G(\omega)|^2 d\omega$

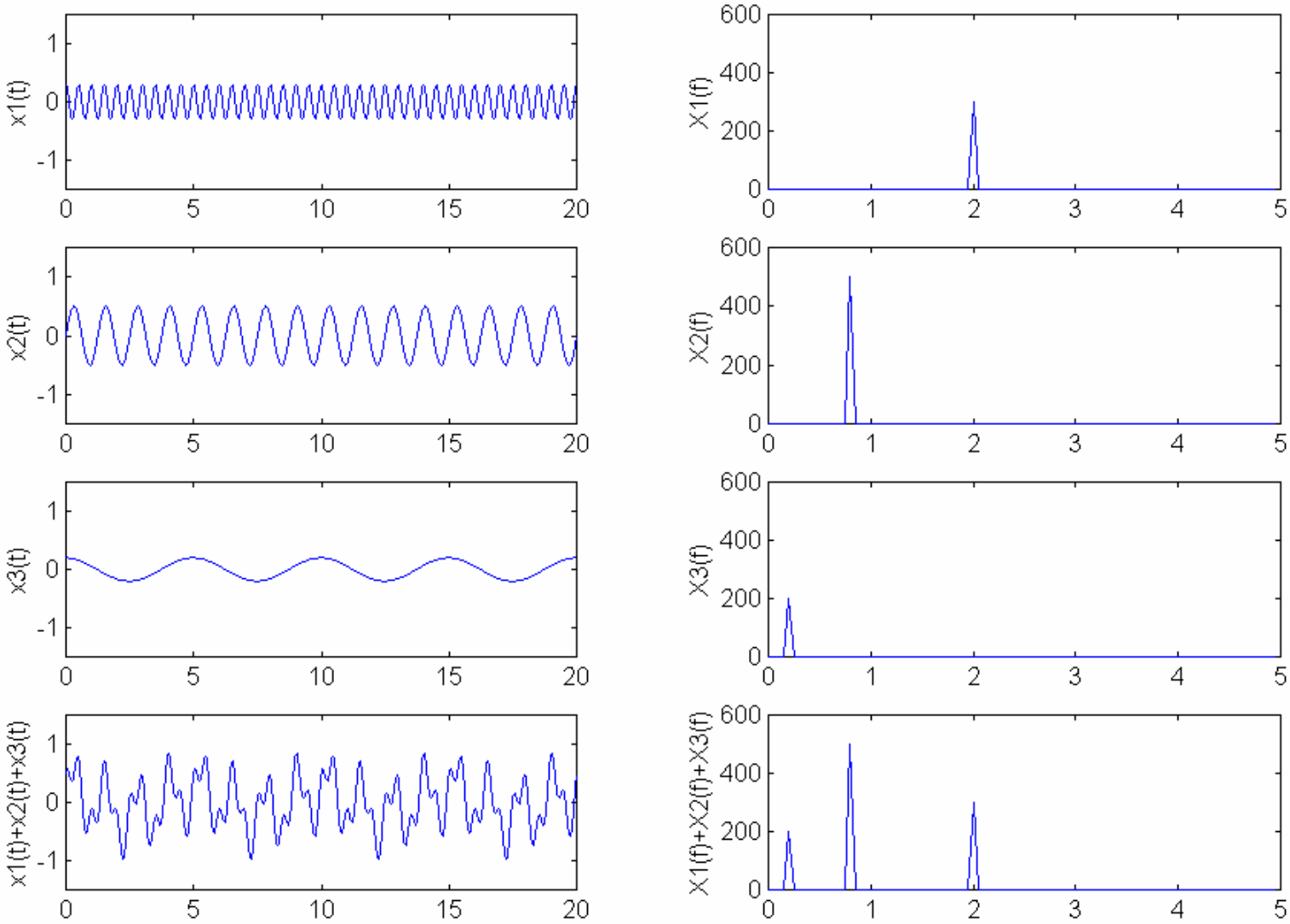
- **Descomposición en componentes de frecuencia:**



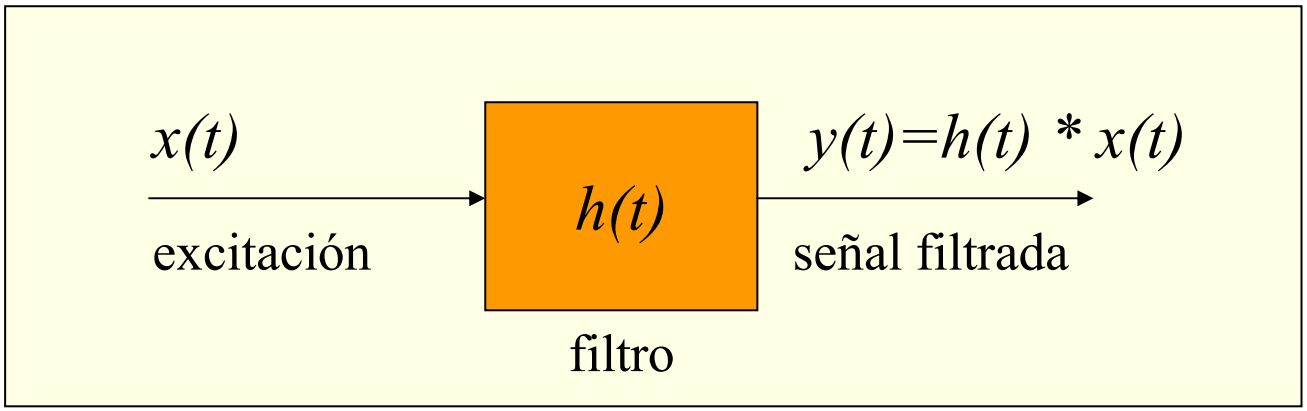
- **Linealidad de la transformada de Fourier:**

$$\begin{array}{ccc} \text{tiempo} & \rightarrow & \text{frecuencia} \\ x_1(t) & \rightarrow & X_1(f) \\ x_2(t) & \rightarrow & X_2(f) \\ a \cdot x_1(t) + b \cdot x_2(t) & \rightarrow & a \cdot X_1(f) + b \cdot X_2(f) \end{array}$$

- **Linealidad de la transformada de Fourier:**

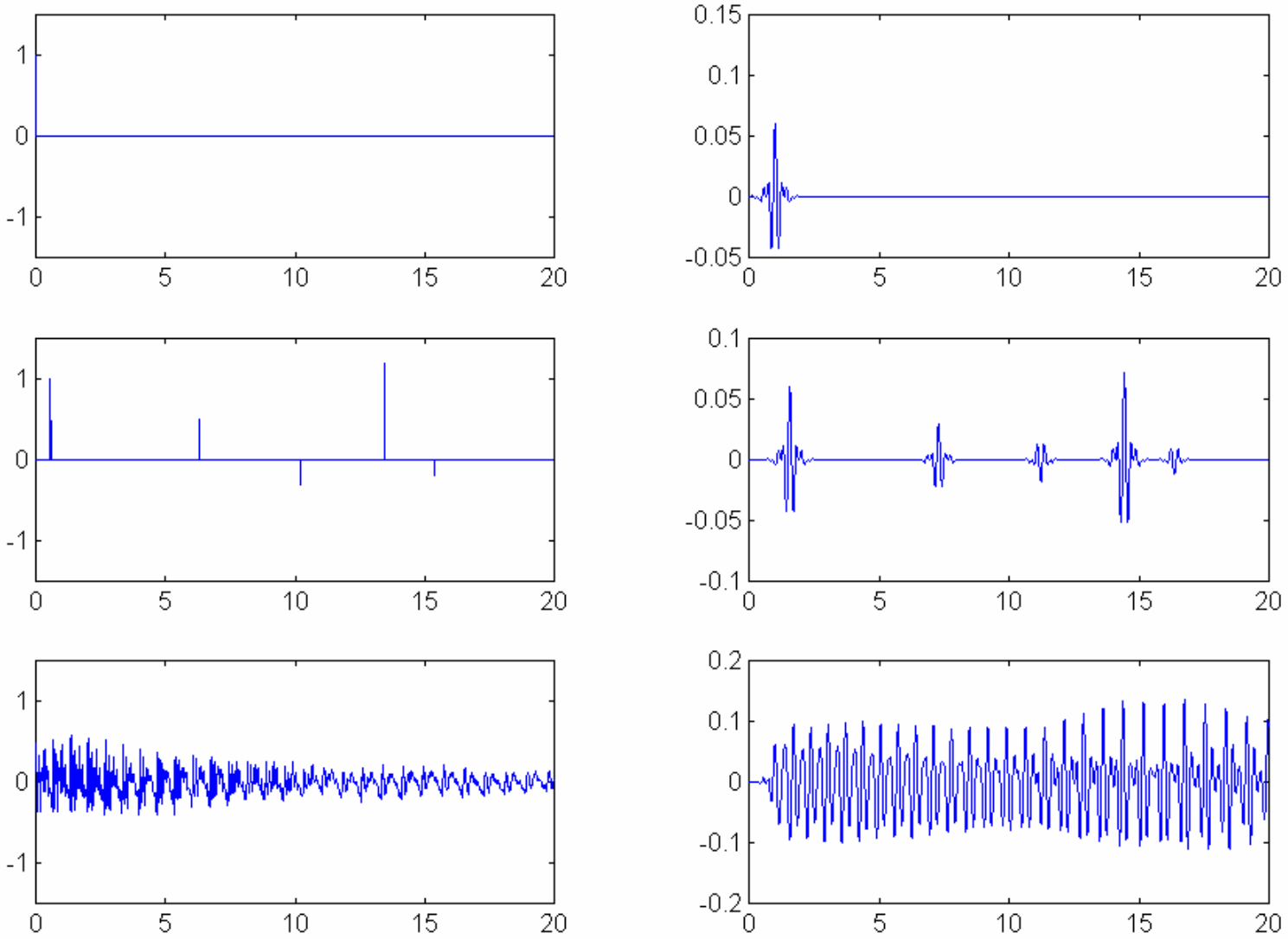


- **Filtrado:**

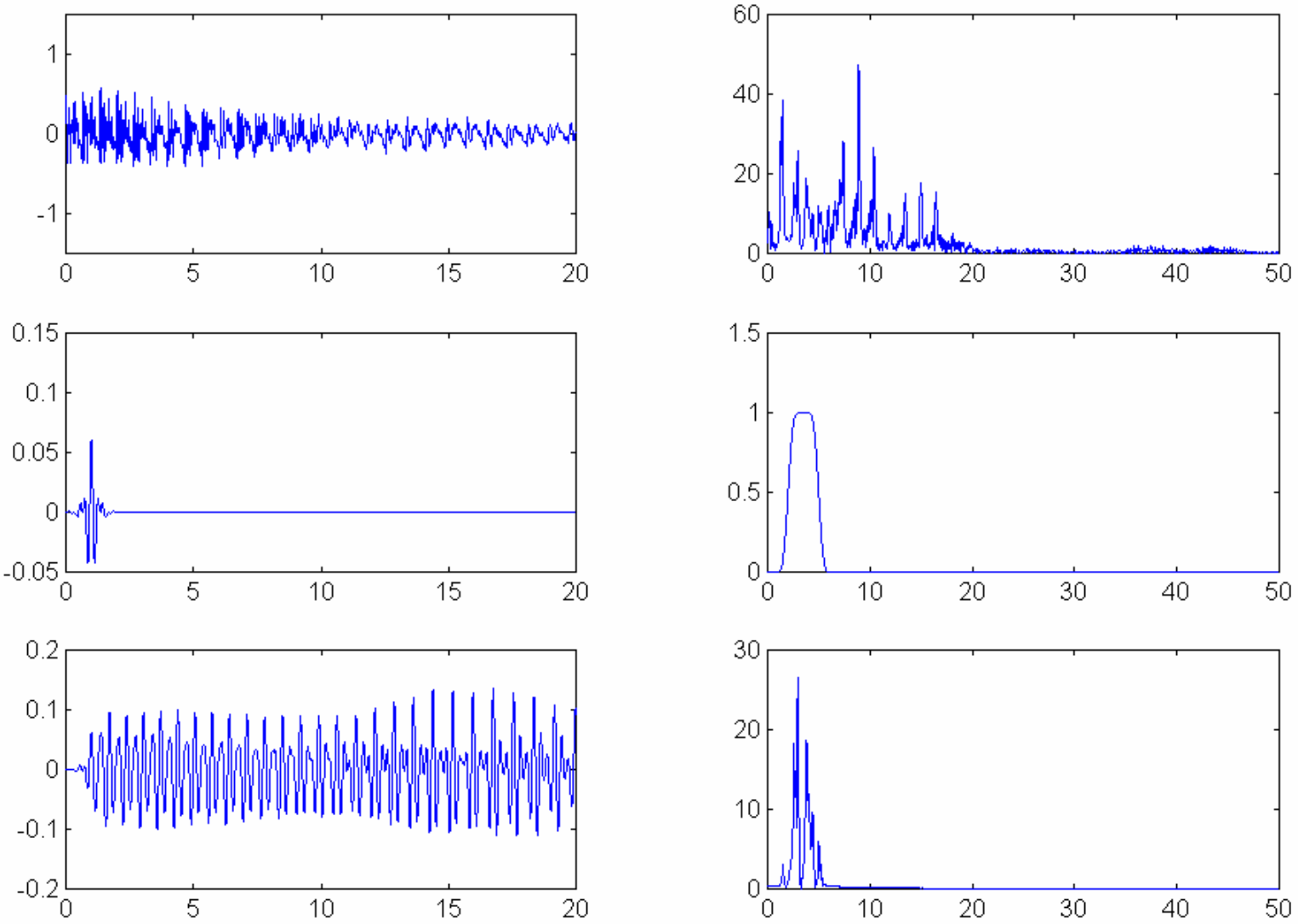


- **Caracterización del filtro:**
  - En el dominio del tiempo: por su respuesta impulsiva *h(t)*
  - En el dominio de la frecuencia: por su función de transferencia o respuesta en frecuencia *H(f)*

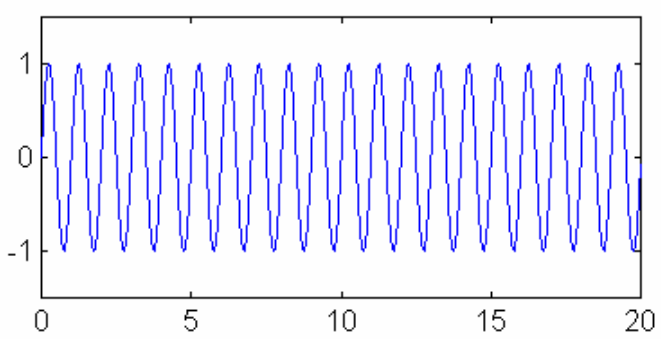
- **Filtrado en el dominio del tiempo: convolución**

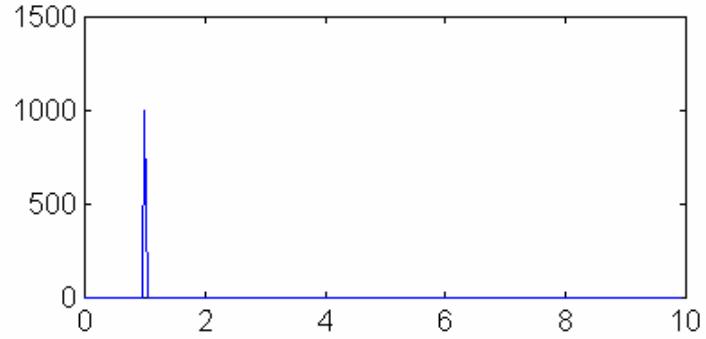


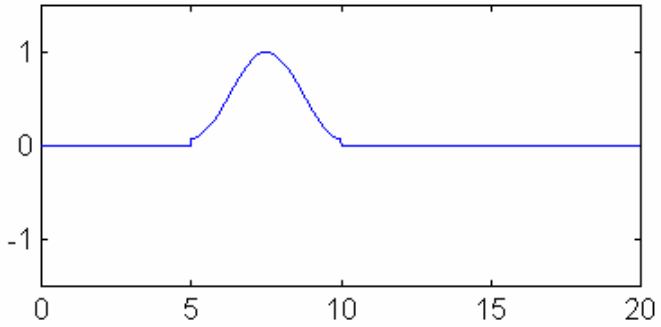
- **Filtrado en el dominio de la frecuencia: multiplicación**

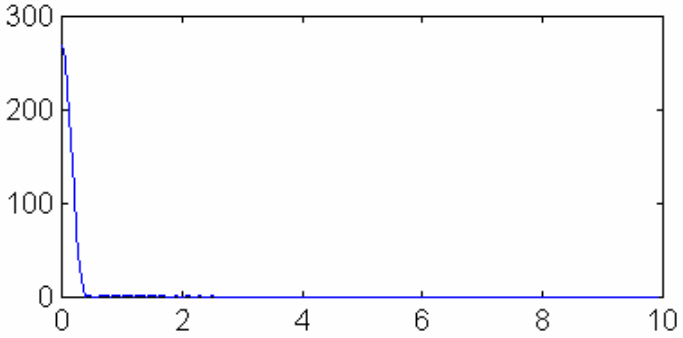


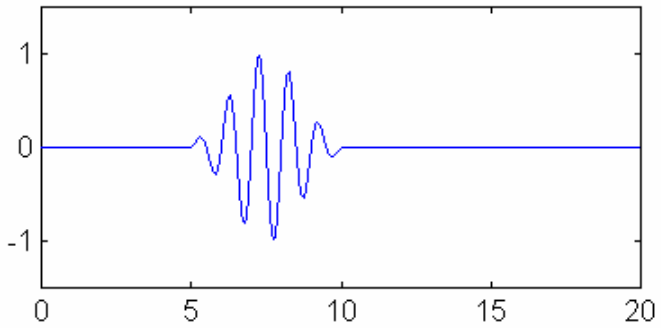
- **Ventanas (multiplicación en el tiempo; convolución en frecuencia)**

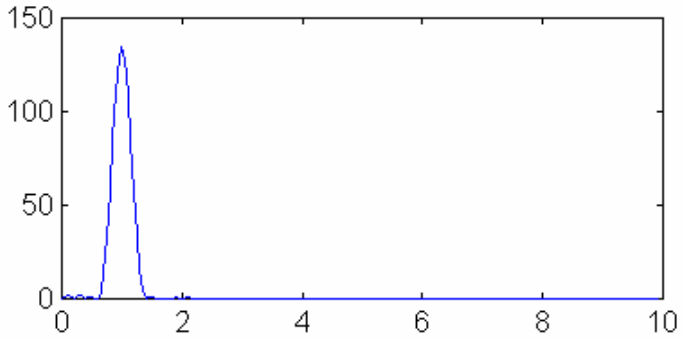




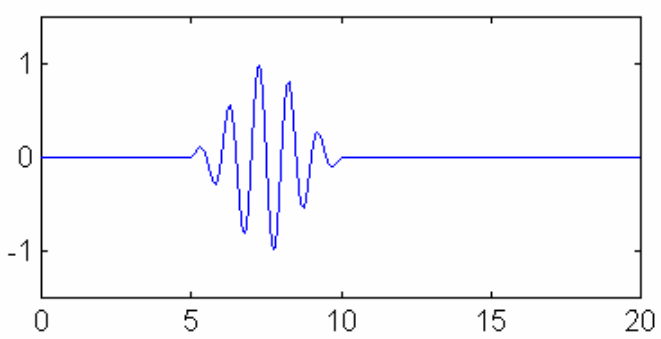


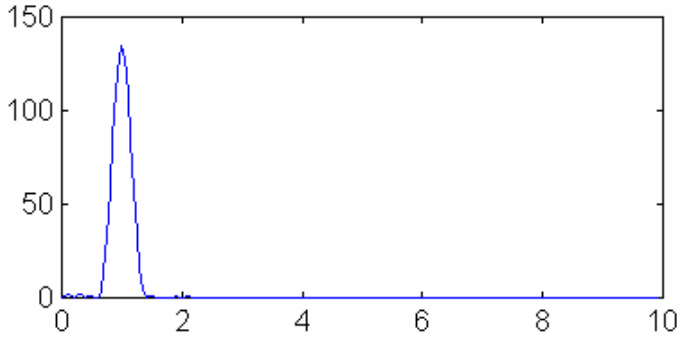


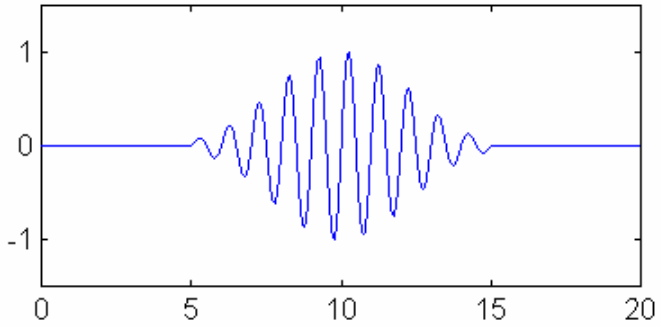


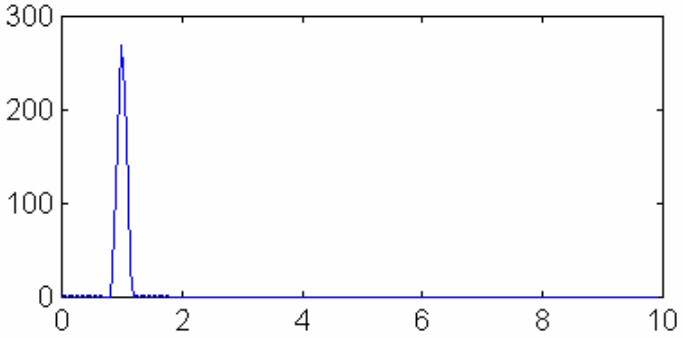


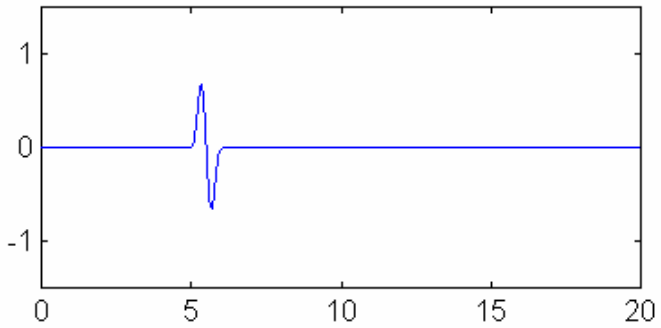
- **Tamaño de ventana y resolución espectral**

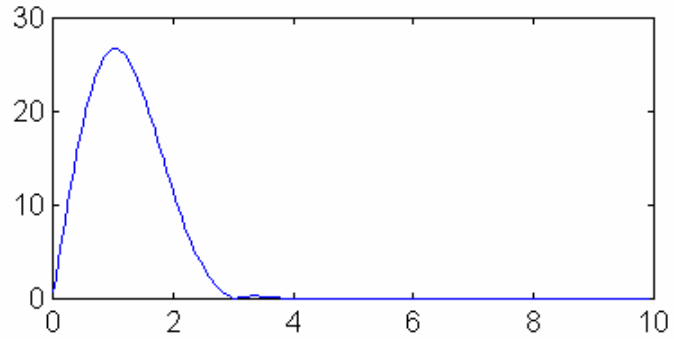




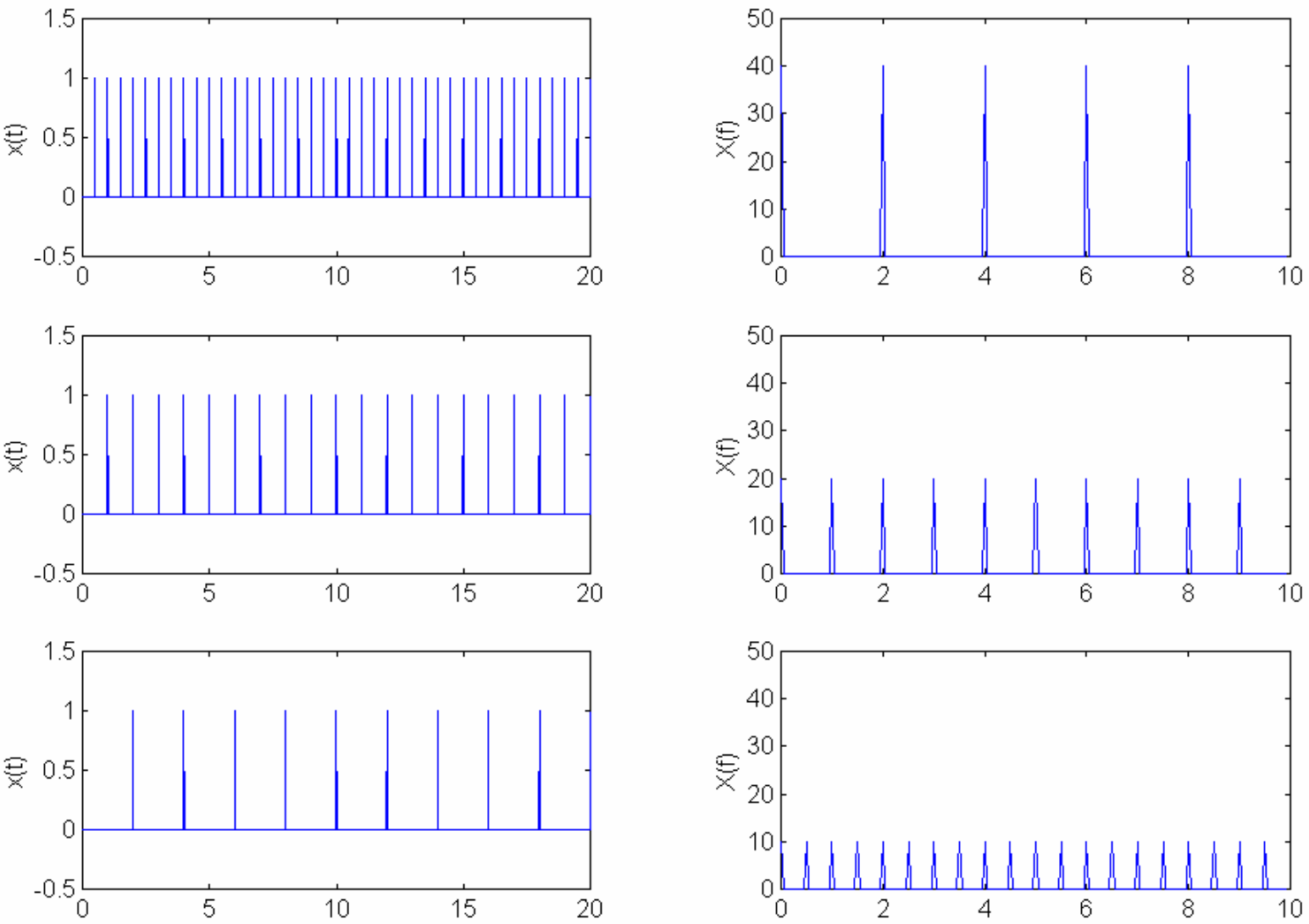




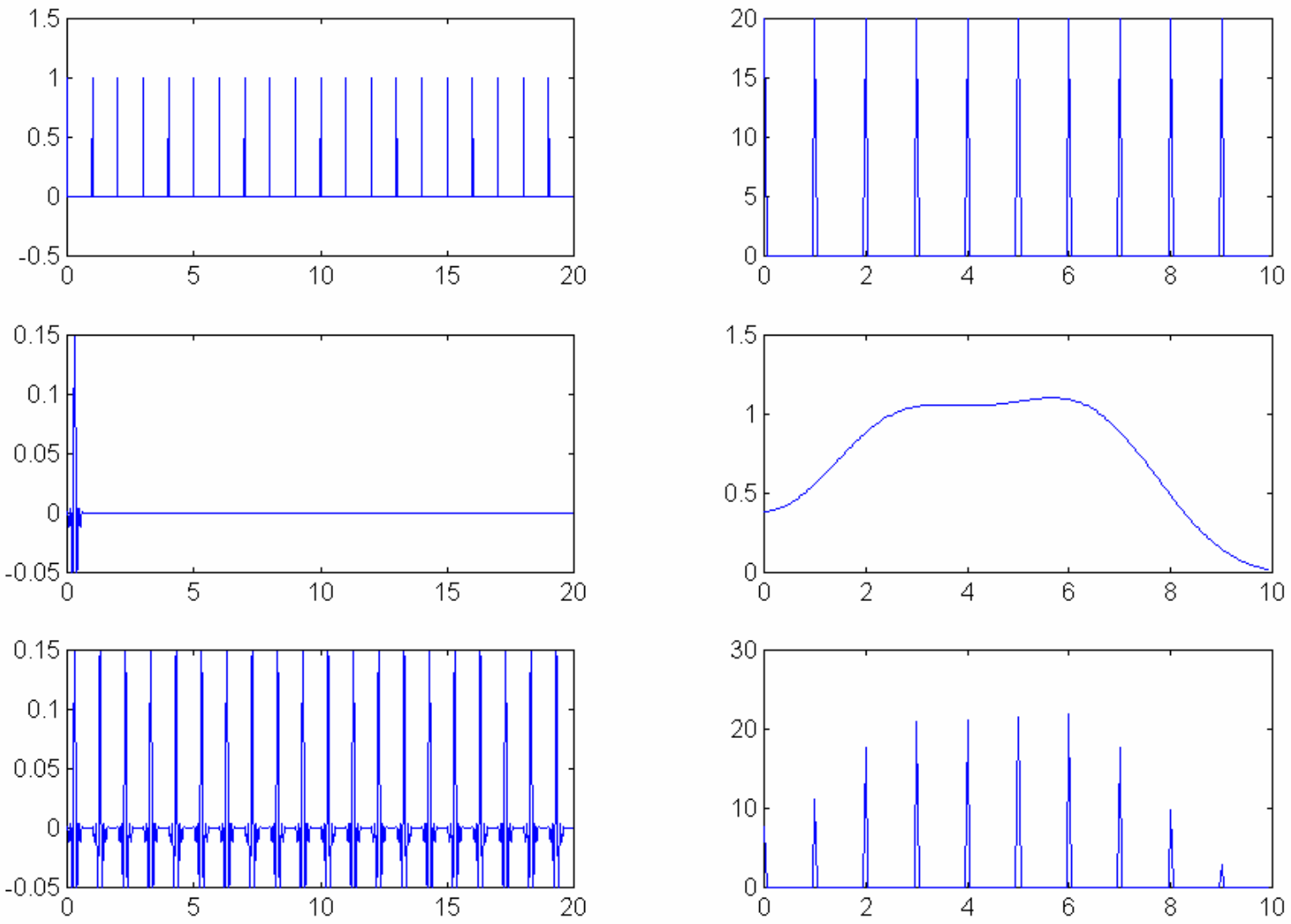




- **Transformada de un tren de pulsos**



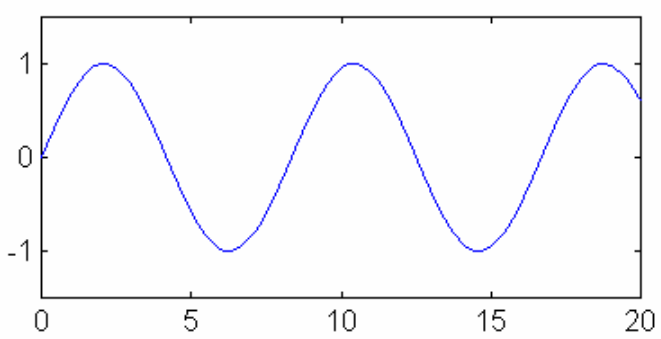
- **Transformada de una señal periódica**

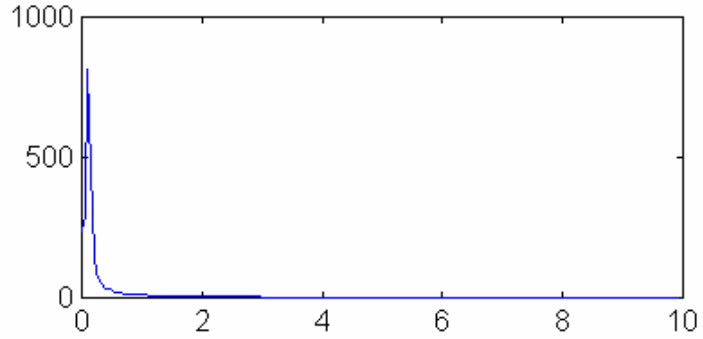


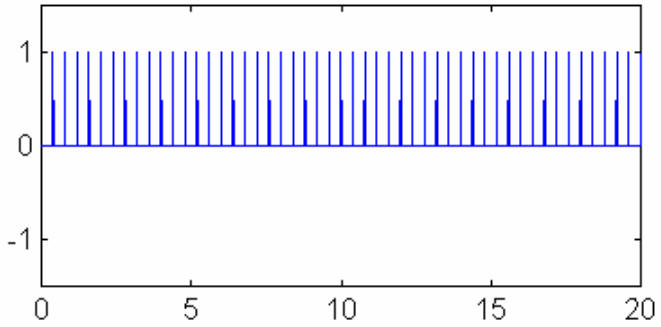
- **Propiedades más importantes**

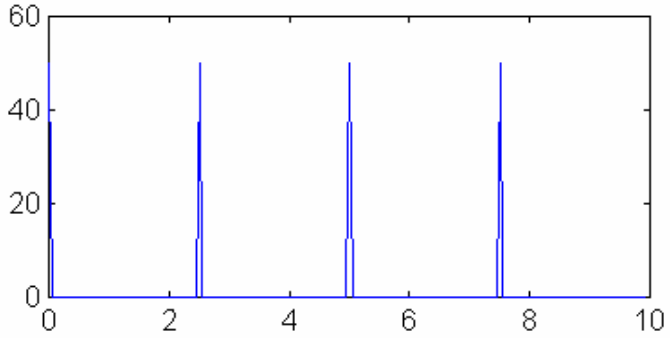
	tiempo $\rightarrow$	**frecuencia**
s. armónicas	seno-coseno (inf.) periodo T	línea espectral $f = 1/T$
linealidad	$a \cdot x_1(t) + b \cdot x_2(t)$	$a \cdot X_1(f) + b \cdot X_2(f)$
filtrado	$x(t) * h(t)$	$X(f) \cdot H(f)$
ventana	$x(t) \cdot w(t)$ (Δt)	$X(f) * W(f)$ $\Delta f = 1/\Delta t$
tren pulsos	$p(t)$ periodo T	$P(f)$ serie armónicos $F_0 = 1/T$
s. periódicas	$p(t) * h(t)$	$P(f) \cdot H(f)$ serie de armónicos

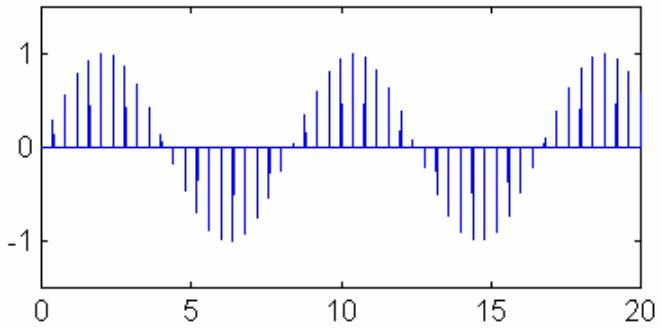
2.5.- El teorema de muestreo

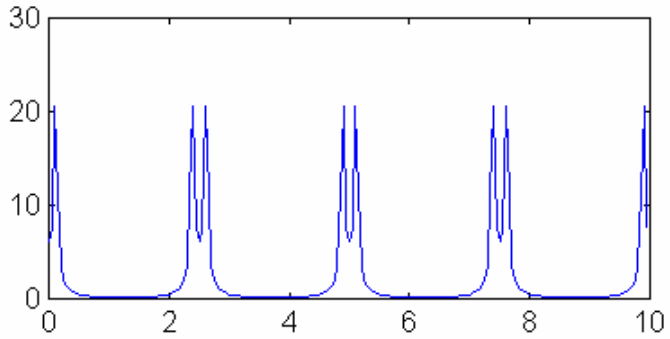












- **ENUNCIADO DEL TEOREMA DE MUESTREO:**

  - UNA SEÑAL LIMITADA EN BANDA A B Hz QUEDA REPRESENTADA POR SUS VALORES (MUESTRAS) TOMADOS A INTERVALOS REGULARES CON UNA FRECUENCIA DE MUESTREO NO INFERIOR A 2B Hz.

  - PARA RECUPERAR LA SEÑAL ORIGINAL, BASTA FILTRAR LA SEÑAL MUESTREADA CON UN FILTRO PASO-BAJA CON FRECUENCIA DE CORTE B Hz.

  - SI LA FRECUENCIA DE MUESTREO NO ES SUFICIENTEMENTE ALTA APARECEN COMPONENTES DE FRECUENCIA “FANTASMAS” (ALIASING)

  - PARA MUESTREAR:
    - Se debe seleccionar una frecuencia de muestreo suficientemente alta
    - O bien se debe filtrar paso-baja la señal antes de muestrear

2.6.- Señales analógicas y señales digitales

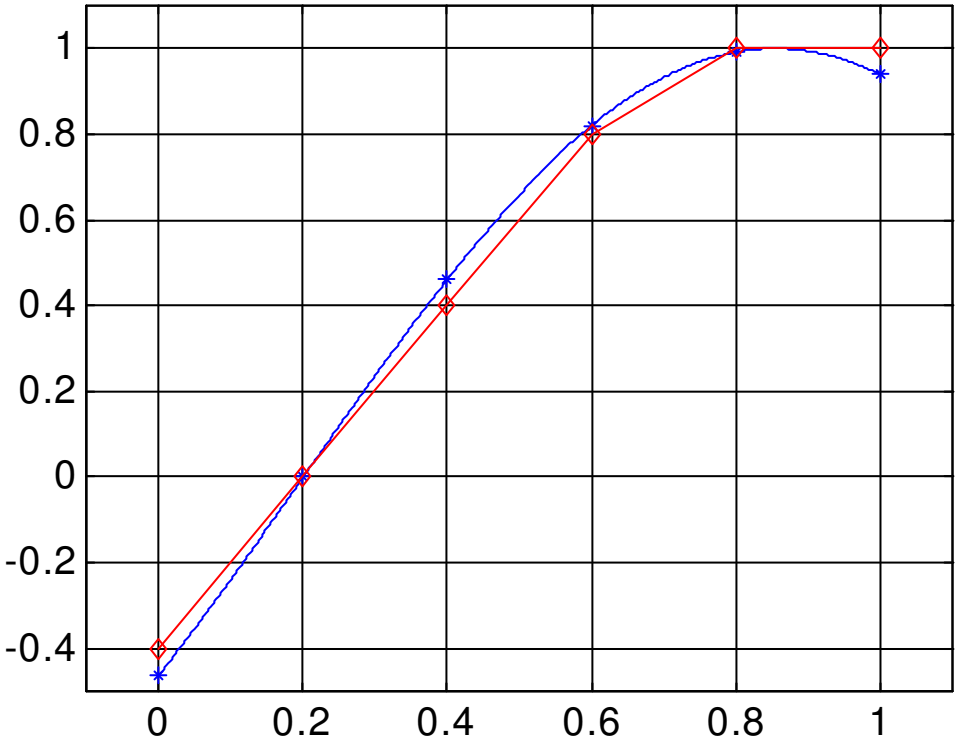
- **Señales físicas:**
  - Magnitud continua, variable continua: *x(t)*
  - Señal discreta
  - Variable discreta

- **Representación digital de una señal analógica:**
  - Discretización en el tiempo: MUESTREO
  - Discretización de la magnitud: CUANTIZACIÓN
  - El muestreo no supone pérdida de información (T. de muestreo)
  - La cuantización introduce un ruido (ruido de cuantización) (precisión limitada)

- **Adquisición de una señal:**
  - Amplificación y filtrado. Muestreo. Retención. Cuantización. Codificación

Adquisición de señales

- Muestreo: discretizar el tiempo:
  - *x(t)*
  - *x(n)*
- Cuantización: discretizar la magnitud:
  - *x(n)*
  - *x'(n)*
- Codificación (representación digital de las muestras)



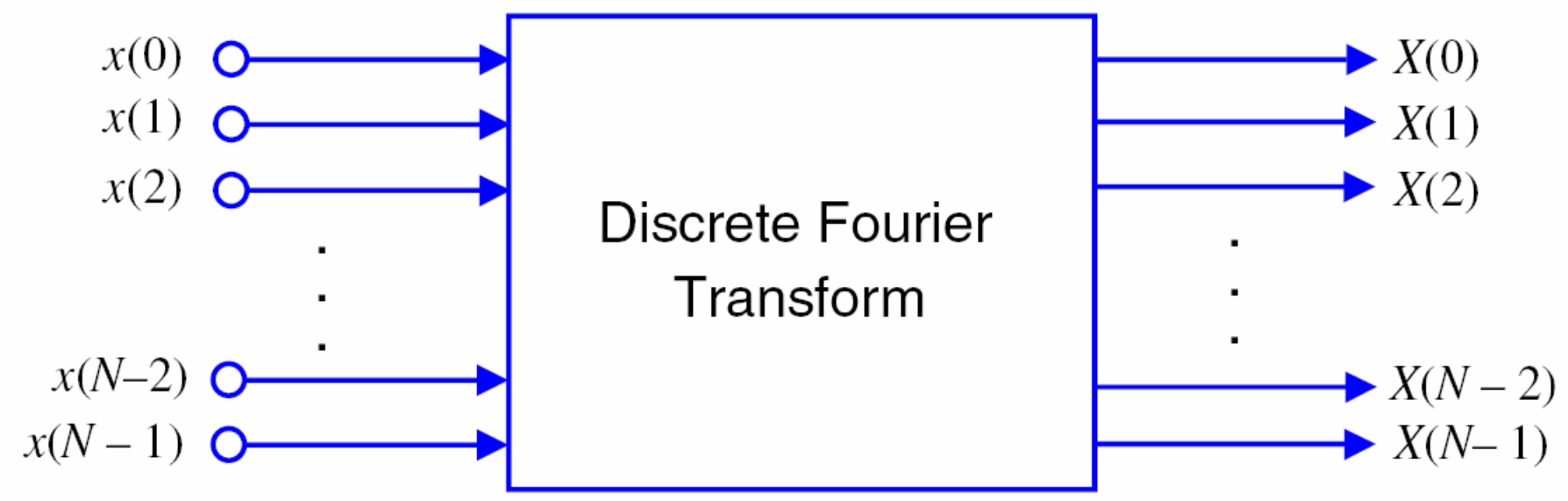
2.7.- La transformada discreta de Fourier

- **En la práctica, la transformada de Fourier no se utiliza para analizar señales:**
  - Señales infinitas, continuas, no periódicas: transformada de Fourier adecuada
  - FT requiere integración numérica
  - Señales digitales: muestreadas, finitas: Transformada discreta de Fourier (DFT)
- **Definición de la transformada discreta de Fourier :**
  - Para una señal discreta en el tiempo *x(m)*, finita con *N* muestras, la DFT se define como *N* muestras espectrales uniformemente espaciadas
  - Existe también una transformada discreta de Fourier inversa
  - Debido a la longitud finita de *x(m)* (equivale a aplicar ventana) no es necesario calcular el espectro para cualquier frecuencia (resolución espectral limitada)

$$X(k) = \sum_{m=0}^{N-1} x(m) e^{-j(2\pi/N)mk} \qquad x(m) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j(2\pi/N)mk}$$

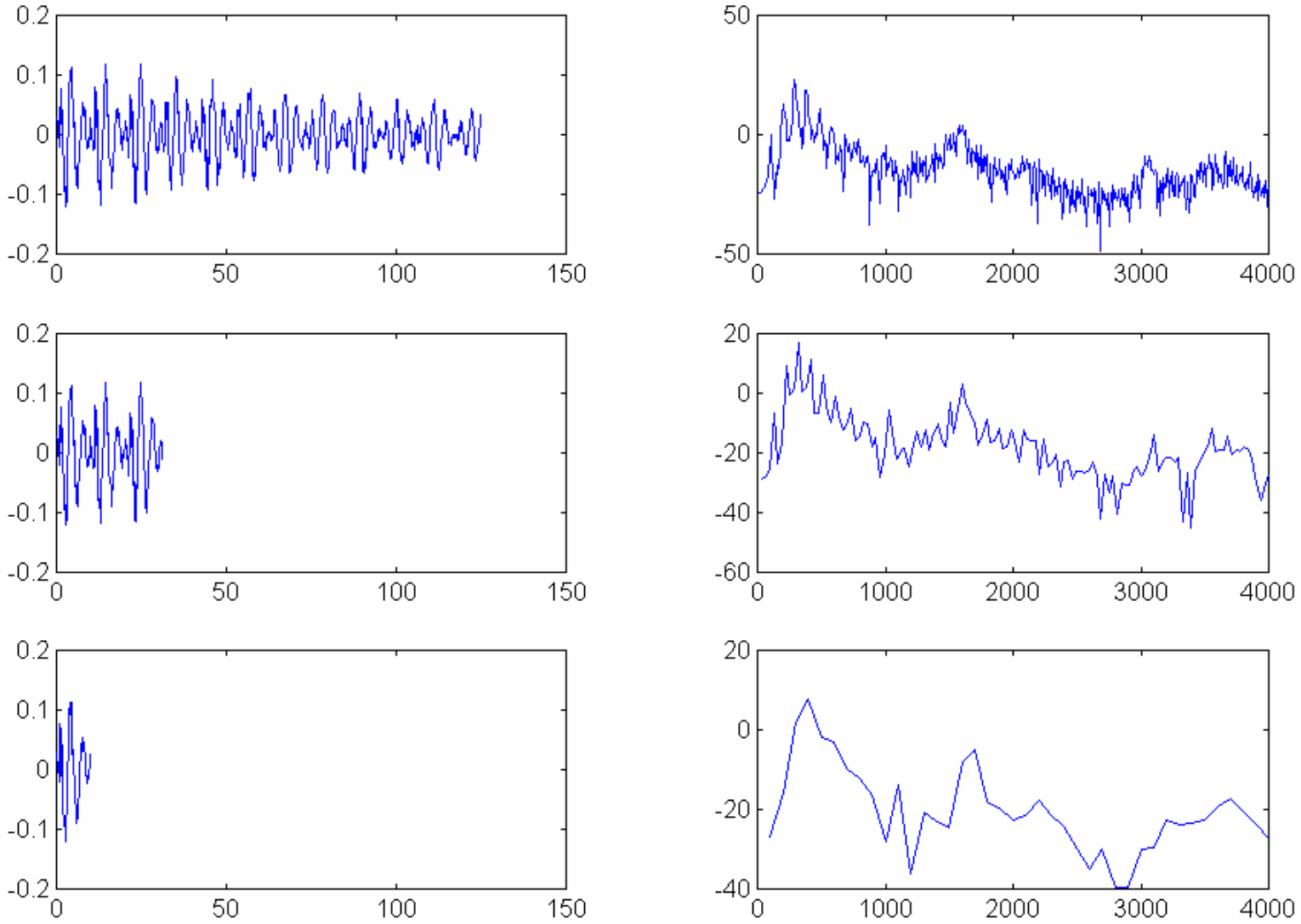
$$X(k) = \sum_{m=0}^{N-1} x(m) e^{-j(2\pi/N)mk}$$

$$x(m) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j(2\pi/N)mk}$$



- **Algoritmos DFT:**
  - Transformada discreta de Fourier (aplicar matriz sobre un vector) DFT
  - Algoritmo rápido (Fast Fourier Transform) FFT es el más usual (eficiente)
  - Forma de la ventana: rectangular, Hamming, Hanning, Gaussiana, etc.
  - Importancia de la longitud de la ventana: condiciona la resolución espectral

- **Tamaño de ventana y resolución espectral en FFT**



2.8.- Procesamiento digital de señales

- **Operaciones habituales en procesamiento digital de señales:**
  - Filtrado lineal (filtros IIR, filtros FIR); filtrado no lineal
  - Procesamiento en tramas (ventanas deslizantes)
  - Procesamiento mediante bancos de filtros
  - Extracción de características
  - Reducción de ruido (realce)
  - Normalización
  - Comparación con modelos
- **Las operaciones aplicadas dependen del tipo de señal y del tipo de información que se quiere obtener:**
  - Señal de audio
  - Identificación de locutor, reconocimiento de fonemas, análisis del tono fundamental, etc...
  - Métodos no paramétricos; basados en modelos; bayesianos; ANNs; HMMs....

TEMA 3

MODELO DIGITAL DE PRODUCCIÓN DE VOZ

Tema 3: MODELO DIGITAL DE PRODUCCIÓN DE VOZ

3.1.- Introducción.

3.2.- Producción de voz.

3.3.- Resonancias del tracto vocal.

3.4.- Modelo excitación – filtrado.

3.5.- Evolución temporal de los parámetros del modelo.

3.6.- Caracterización de los sonidos de voz:

- Tono, timbre, intensidad y duración.
- Frecuencia fundamental, formantes, evolución temporal
- Representación espectral de tiempo corto

3.1.- Introducción

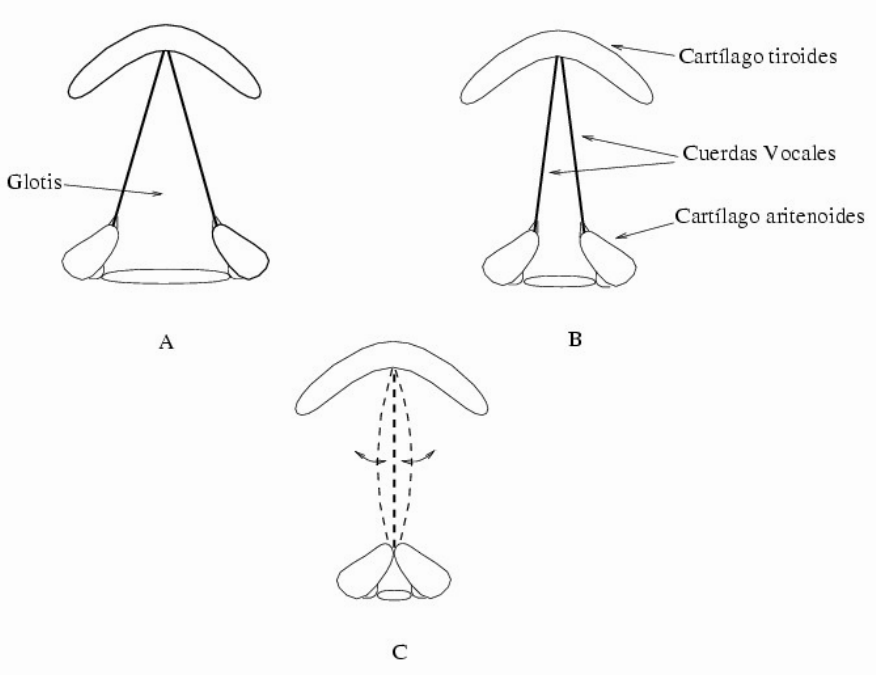
- **Las características de la voz están condicionadas por los mecanismos de producción:**
  - Generación del sonido
  - Resonancias
  - Radiación acústica

- **La voz es una onda acústica:**
  - Física acústica de la producción de voz

- **En este tema veremos un modelo digital de producción de voz:**
  - Excitación – Filtrado
  - Características de la señal de voz (es una señal de audio muy particular)

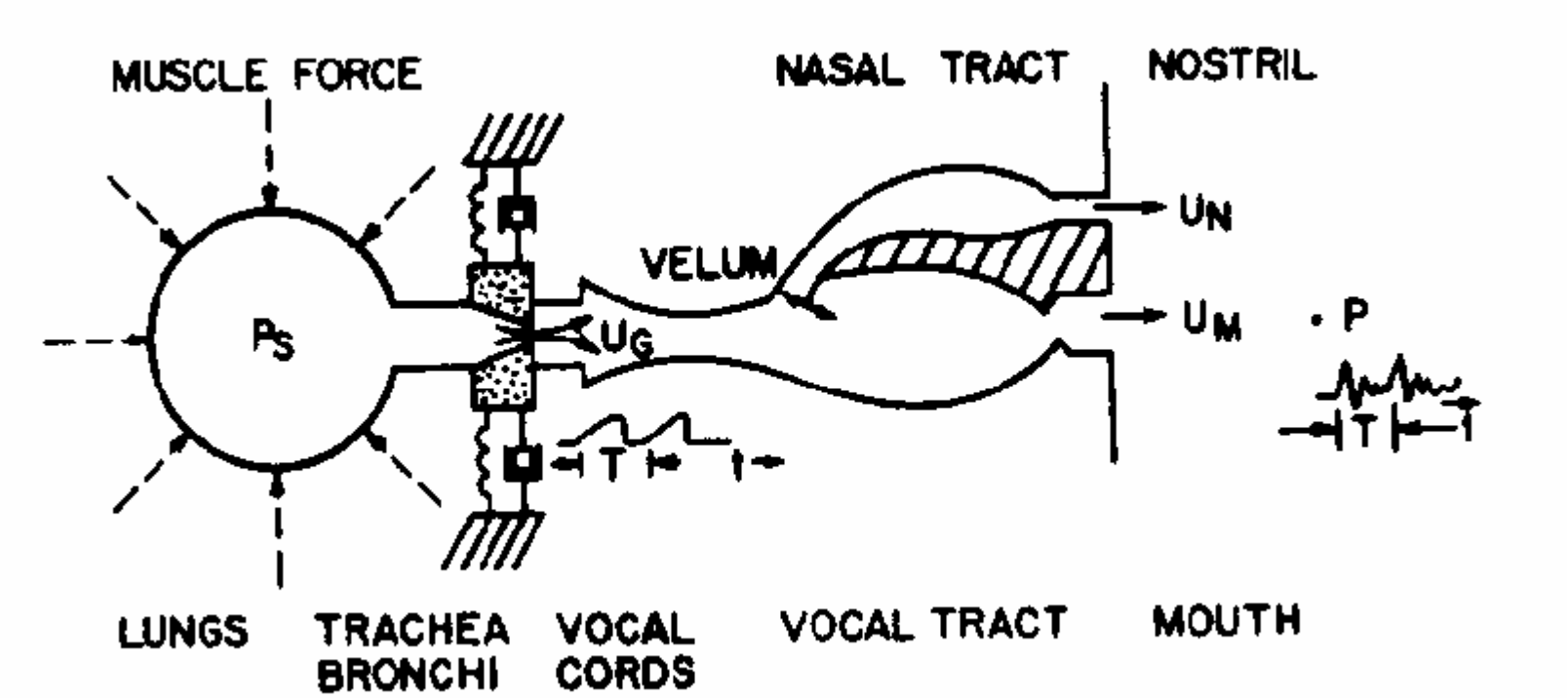
3.2.- Producción de la voz

- **Órganos de producción de voz:**
  - Cavidades infraglóticas
  - Cavidad laríngea (cuerdas vocales)
  - Cavidades supraglóticas



- **Provisión de aire**
- **Generación del sonido:**
  - Vibración cuerdas vocales (onda glotal) en fonemas sonoros
  - Flujo turbulento (fricativos)
  - Oclusión + apertura (oclusivos)
- **Filtrado del sonido**
- **Radiación del sonido**

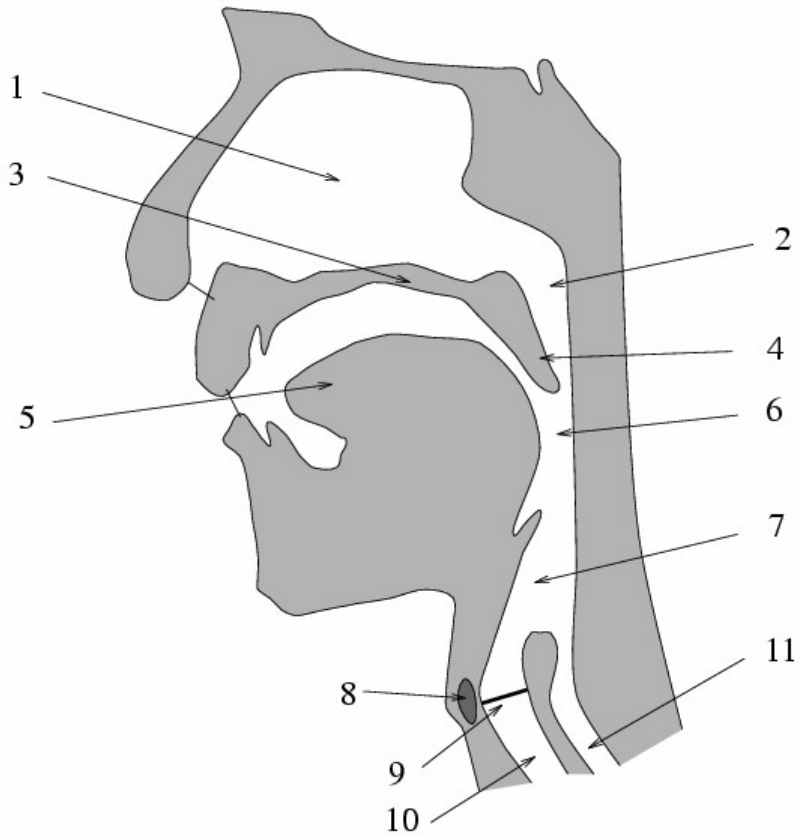
- **Cuerdas vocales:**
  - (A) Respiración
  - (B) Fonemas sordos
  - (C) Fonemas sonoros



- Pulmones
- Traquea, bronquios
- Cuerdas vocales

- Velo del paladar
- Tracto vocal
- Boca
- Tracto nasal
- Orificios nasales

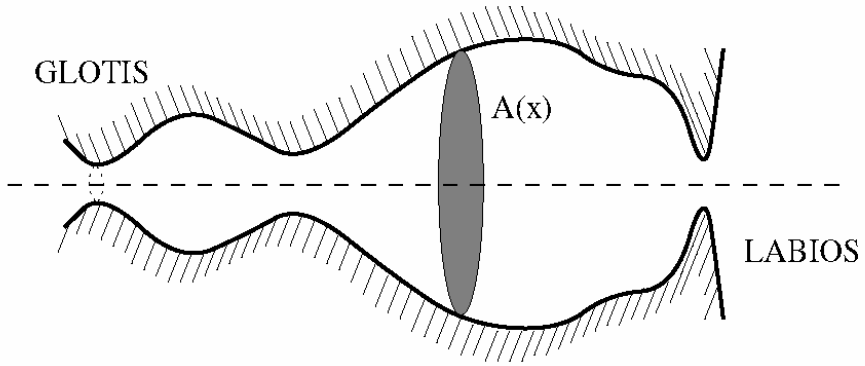
- **Cavidades supraglóticas (diversificación fonética):**
  - 1 Cavidad nasal
  - 2 Faringe nasal
  - 3 Paladar duro
  - 4 Paladar blando o velo del paladar
  - 5 Lengua
  - 6 Faringe oral
  - 7 Faringe laríngea
  - 8 Cartílago tiroides
  - 9 Cuerdas vocales
  - 10 Traquea
  - 11 Esófago

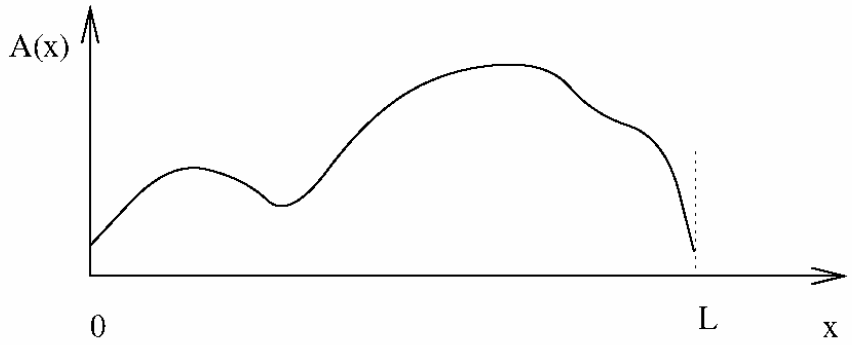


3.3.- Resonancias del tracto vocal

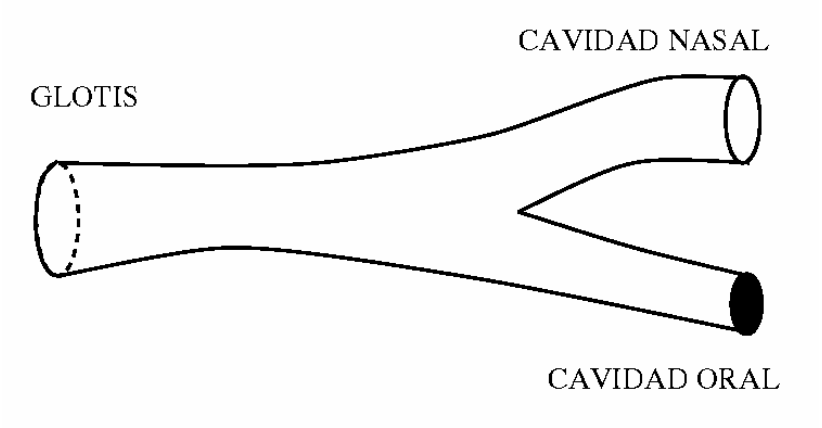
- **Voz: onda acústica (onda de presión que se propaga por el aire)**
  - Velocidad: $c = 350$ m/s
  - Longitud de onda: $\lambda = c / f$
    - Para 100 Hz, $\lambda = 3.5$ m
    - Para 4 kHz, $\lambda = 8.5$ cm
    - $\lambda > r$ (radio del tubo) $\Rightarrow$ aproximación de onda plana

- **La forma del tracto vocal condiciona las resonancias:**
  - El tracto vocal filtra del sonido generado
  - El tracto vocal queda descrito por la "función de área" $A(x,t)$
  - Variación del tracto vocal: se producen entre 5 y 20 fonemas por segundo (varía lentamente)
  - Acoplamiento del tracto nasal (velo del paladar)





Función de área

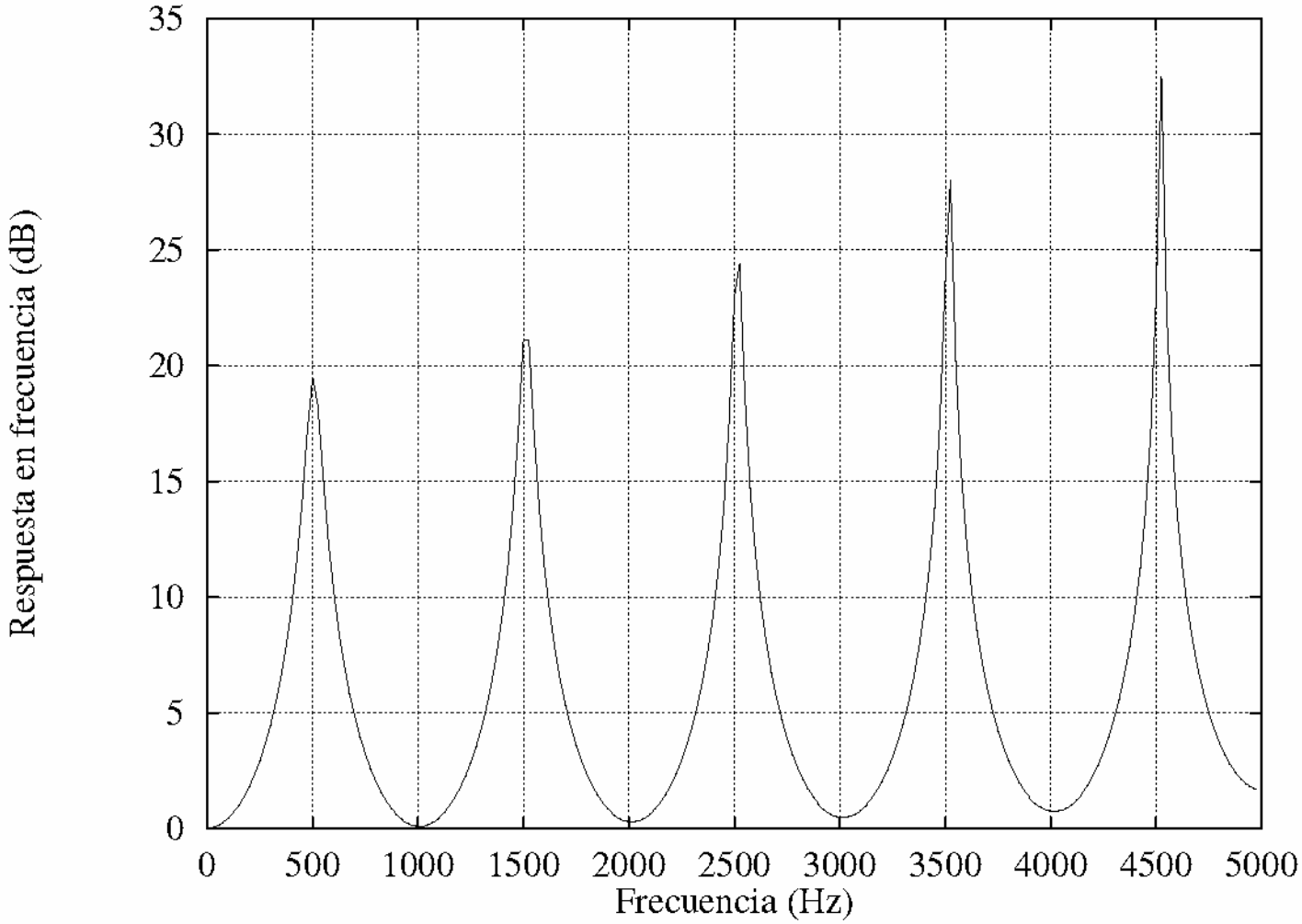


Acoplamiento del tracto nasal

- **Simplificaciones para estudiar las resonancias del tracto vocal**
  - Aproximación de onda plana (onda unidimensional)
  - Estacionariedad (función de área invariante en el tiempo $A(x,t) = A(x)$)
  - Tubo de sección constante a trozos
  - Tubo de sección constante
  - Impedancia acústica nula en los labios (ignorar radiación)
  - Pérdidas despreciables (ignorar pérdidas por viscosidad, conducción térmica, etc.)
  - Tubo rígido (ignorar elasticidad del tracto vocal)

- **La función de área** $A(x)$
  - Si $A(x)$ es sencilla, se pueden obtener soluciones exactas para las resonancias
  - Si $A(x)$ es compleja, métodos numéricos
  - $A(x)$ se puede medir por diversos métodos: Rx, TAC, RMN, articulógrafos

Respuesta en frecuencia del tracto vocal: formantes



L = 17.5 cm

A = 5.0 cm^2 (cte)

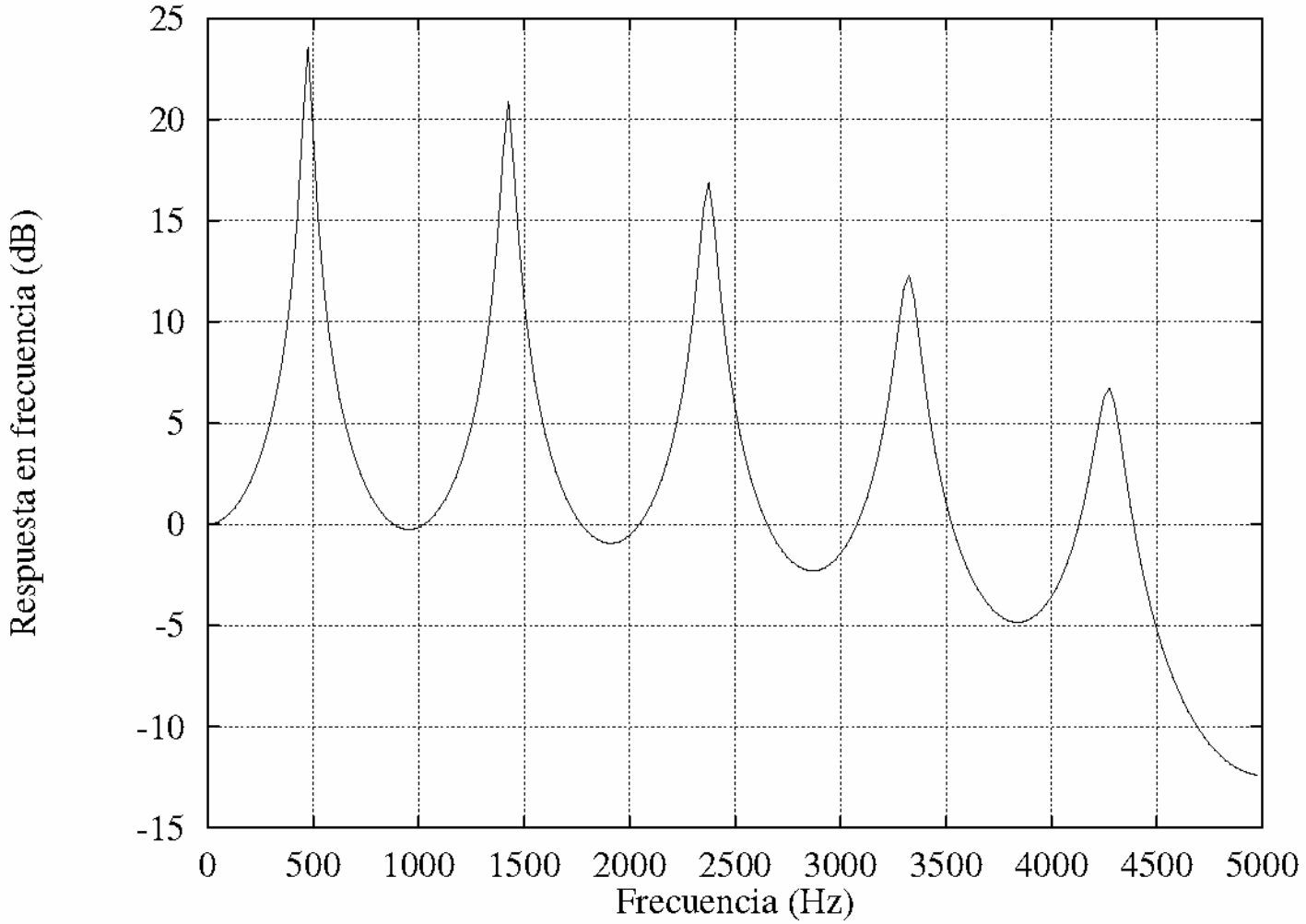
Paredes elásticas

1 resonancia / kHz

- **Formantes**
  - Los formantes son resonancias del tracto vocal
  - Debido a las dimensiones del tracto vocal y a la velocidad de propagación, aparece en promedio 1 formante por cada kHz
  - El tracto vocal “filtra” el sonido generado:
    - Amplifica algunas frecuencias (correspondientes a los formantes)
    - Atenúa otras frecuencias

- **Pérdidas por radiación**
  - Impedancia acústica del bafle
    - Abertura circular en plano infinito
    - Abertura circular en esfera
    - Labios
  - Pérdidas dependientes de la frecuencia
  - Caída para altas frecuencias: aproximadamente 6 dB / década

Respuesta en frecuencia del tracto vocal



L = 17.5 cm

A = 5.0 cm^2 (cte)

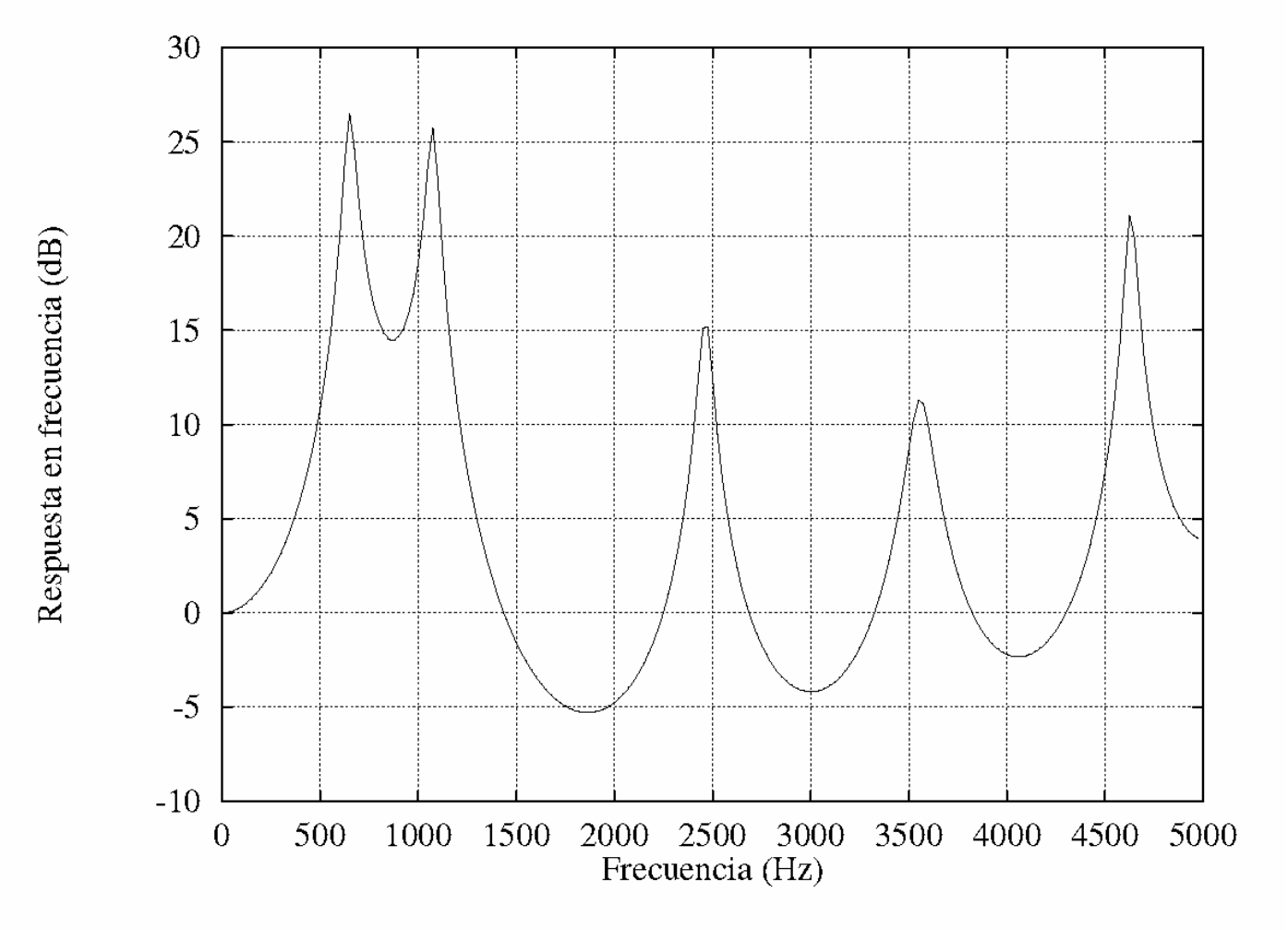
Paredes elásticas

Pérdidas por radiación

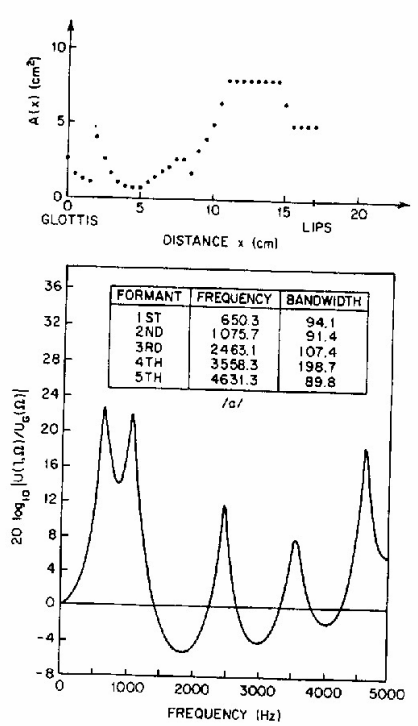
1 formante / kHz

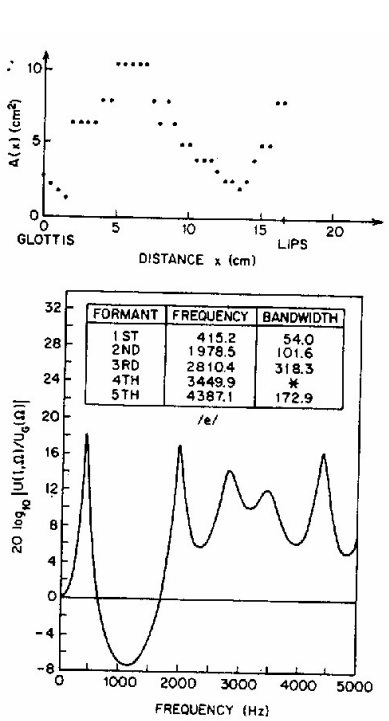
Caída de 6 dB / dec

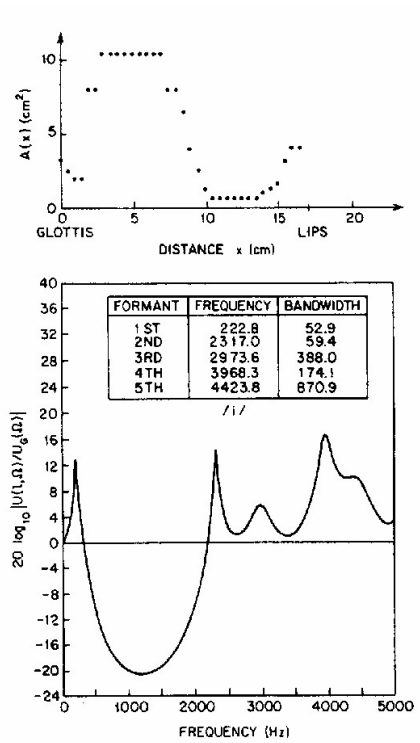
Respuesta en frecuencia del tracto vocal: fonema */a/*

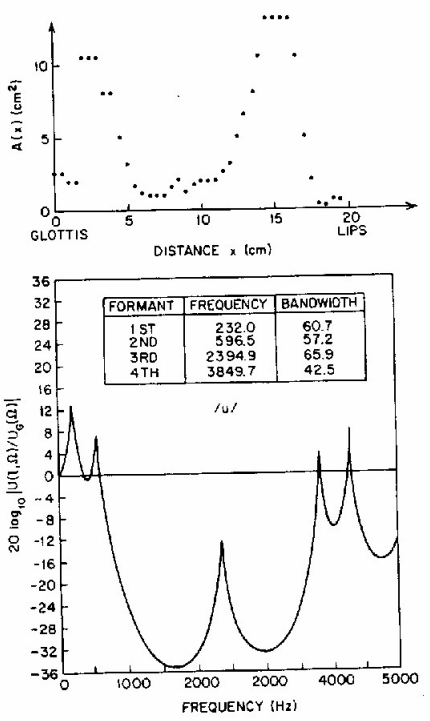


Funciones de área y formantes





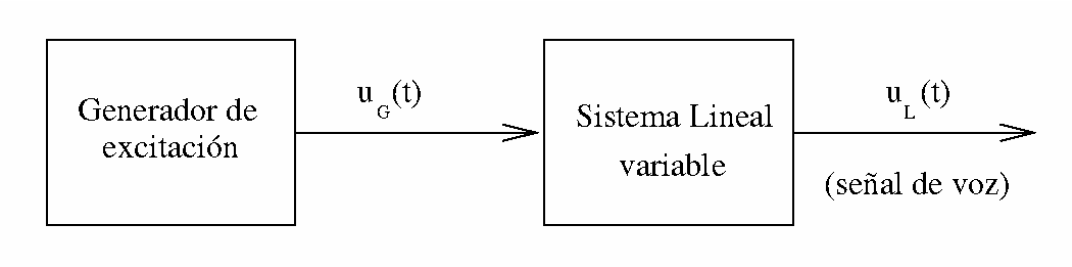


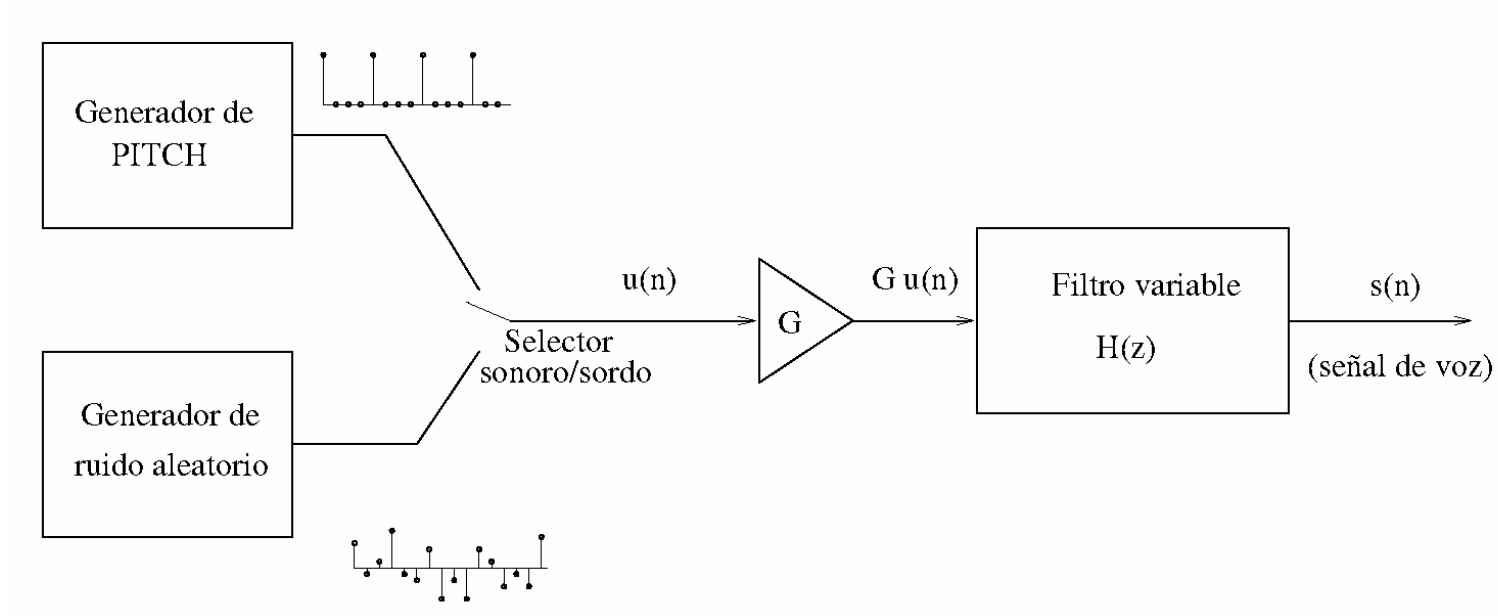


3.4.- Modelo excitación – filtrado

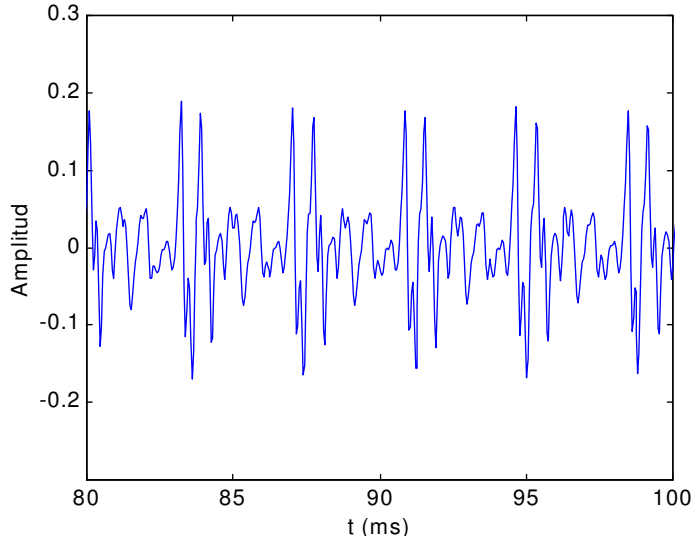
- **Excitación:**
  - Fonemas sonoros: vibración de cuerdas vocales
    - Tren de pulsos glotales (pitch)
    - Frecuencia fundamental f_0, periodo del pitch T_0
  - Fonemas fricativos: flujo turbulento en un estrechamiento del tracto
  - Fonemas oclusivos: oclusión + apertura
  - Combinación de mecanismos de excitación
- **Filtrado:**
  - Función de área del tracto vocal y nasal (acoplamiento nasal)
  - Radiación
  - Un formante por kHz y caída promedio de 6 dB / década
- **Evolución temporal de la excitación y el filtrado**

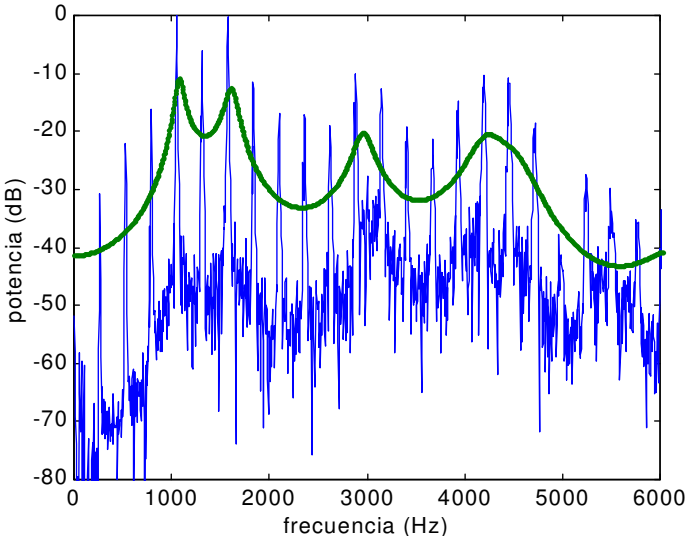
Modelo digital de producción de voz

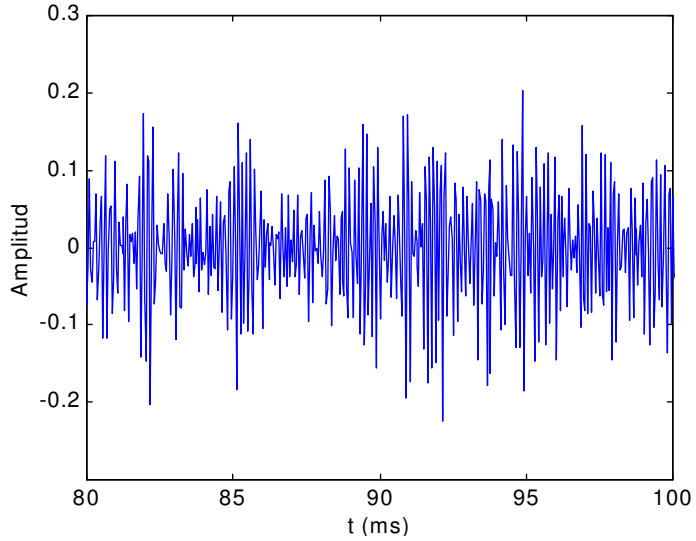


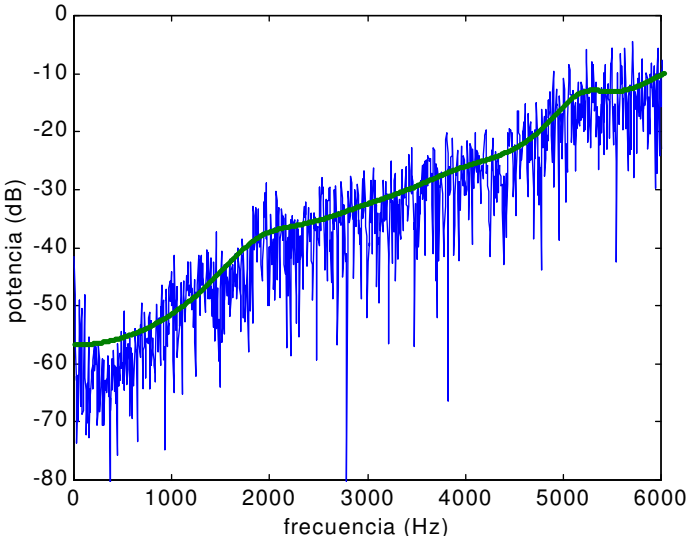


Ejemplo de fonemas sonoro y sordo: /a/ /s/









3.5.- Evolución temporal de los parámetros del modelo

- **Limitaciones fisiológicas:**
  - Variaciones en la presión de aire suministrada
  - Variaciones en la tensión de las cuerdas vocales
  - Variaciones en la conformación del tracto buco-nasal
  - Limitación en la velocidad de producción de fonemas: de 5 a 20 fonemas / seg

- **Cuasiestacionariedad de la voz:**
  - En segmentos cortos, la voz se puede considerar una señal estacionaria:
    - Excitación constante (intensidad constante, frecuencia fundamental constante)
    - Resonancias constantes
  - Ventana de análisis: entre 25 ms y 50 ms
  - Problemas de esta aproximación:
    - Coarticulación
    - Consonantes no estacionarias (oclusivas, africadas)

3.6.- Caracterización de los sonidos de voz

Tono, timbre, intensidad y duración:

- **Tono:**
  - Relacionado con vibración de cuerdas vocales
  - En fonemas sonoros
  - Periodicidad en el tiempo; serie de armónicos en frecuencia
- **Timbre:**
  - Relacionado con los formantes (o resonancias o conformación del tracto vocal)
  - Envolvente espectral
  - Patrón repetido en el dominio del tiempo
- **Intensidad:**
  - Relacionada con la presión de aire suministrada por pulmones
- **Duración:**
  - Evolución en el tiempo de las características anteriores (ataque, caída, etc.).

Frecuencia fundamental, formantes y evolución temporal:

- **Frecuencia fundamental:**
  - Entre 50 Hz y 400 Hz
  - Rizado espectral (o estructura fina)
- **Formantes:**
  - Un formante por kHz
  - Envolvente espectral
  - Caída de 6 dB por década (filtrado digital de pre-énfasis)
  - Potencia despreciable por encima de 6 kHz
- **Evolución temporal:**
  - En periodos cortos la señal es cuasiestacionaria
  - Evolución lenta (velocidad de producción de fonemas)

Representación espectral de tiempo corto:

- **No tiene sentido estudiar el espectro de un segmento con varios fonemas (espectro promedio)**
- **En periodos cortos (inferiores a 50 ms) señal cuasiestacionaria:**
- **Análisis espectral de tiempo corto:**
  - Segmentación en tramas (longitud entre 20 ms y 50 ms)
  - Análisis espectral de cada trama:
    - Espectro FFT (espectrograma)
    - Banco de filtros
    - Espectro LPC
    - Representaciones cepstrales
- **Análisis típico de señales de voz:**
  - Adquisición de señal digital (amplificación, filtrado y muestreo)
  - Pre-énfasis y segmentación en tramas usando ventanas deslizantes (Hamming)
  - Análisis por tramas (análisis espectral; otros tipos de análisis)

- **Resolución espectral y temporal:**
  - **Resolución temporal:** depende del tamaño de la ventana y del solapamiento entre ventanas
  - **Resolución espectral:** depende del tamaño de la ventana y del tipo de análisis
- **Excitación:**
  - Rizado espectral (estructura fina del espectro)
  - Resolución espectral
  - Ventanas largas (que incluyan varios periodos de pitch) en el espectrograma para resolverlo en frecuencia (mayores de 30 ms, Narrow Band Spectrogram)
  - Ventanas cortas (inferiores a un periodo de pitch) para resolverlo en el dominio del tiempo (menores de 8 ms, Wide Band Spectrogram)
- **Resonancias:**
  - Envolvente espectral
  - Ventanas cortas
  - Técnica de suavizado espectral

TEMA 4

REPRESENTACIÓN DE LA SEÑAL DE VOZ

Tema 4: REPRESENTACIÓN DE LA SEÑAL DE VOZ

4.1.- Introducción.

4.2.- Características de la señal de voz.

4.3.- Representación de la forma de onda.

4.4.- Energía de tiempo corto.

4.5.- Tasa promedio de cruces por cero.

4.6.- Función de autocorrelación de tiempo corto.

4.7.- Estimación del tono fundamental.

4.8.- Análisis de Fourier de tiempo corto. Espectrograma (WB y NB).

4.9.- Linear Prediction Coding: Análisis LPC.

4.10.- Análisis basado en banco de filtros.

4.11.- Procesamiento homomórfico. Cepstrum (FFT, LPC y MFCC).

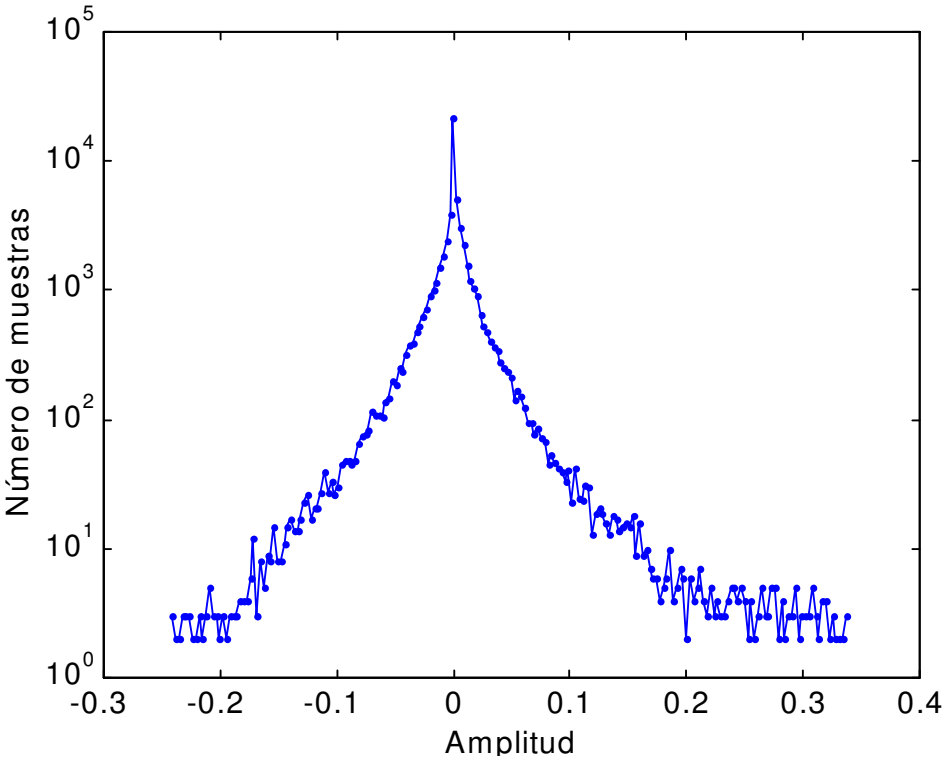
4.1.- Introducción

- **Modelo de producción de voz: excitación + filtrado**
  - Excitación: rizado espectral
  - Filtrado: envolvente espectral
- **Parámetros del modelo: varían lentamente (5 – 20 fonemas / seg)**
- **Representaciones basadas en análisis de tiempo corto**
  - Análisis basado en tramas solapadas
  - Segmentación en tramas y aplicación de ventanas
  - Cada trama se representa por uno o varios parámetros
- **En este tema estudiamos distintas representaciones de la señal de voz, que resultarán útiles para distintos propósitos**

4.2.- Características de la señal de voz

- **Frecuencia de la señal de voz:**
  - Rango de frecuencias audibles: de 20 Hz a 20 kHz
  - Rango mínimo de frecuencias para la voz: de 350 Hz a 3.5 kHz (telefonía)
  - Rango razonable para la voz: de 60 Hz a 6 kHz
  - Caída de 6 dB/octava para frecuencias altas (algo más: 8-10 dB/octava)
  - El rango de frecuencia elegido condiciona la frecuencia de muestreo
- **Intensidad de la señal de voz:**
  - Rango típico: de 50 dBA a 70 dBA
  - Rango extendido: de 30 dBA (susurro) a 90 dBA (gritos a corta distancia)
  - Rango de 60 dB: equivalentemente $A_{max} / A_{min} = 1000$
  - Cuantización uniforme: requiere del orden de 2000 niveles si la ganancia está perfectamente ajustada (11 bits).
    - Típicamente se usan 12 o 16 bits con cuantización uniforme
    - 8 bits para cuantización con compresión instantánea (ley-mu o ley-A)

- **Modelo estadístico de la señal de voz:**
  - Distribución de amplitudes: aproximadamente Gamma o Laplaciana
  - Distribución espectral de la energía: caída de unos 8 dB -10 dB por octava



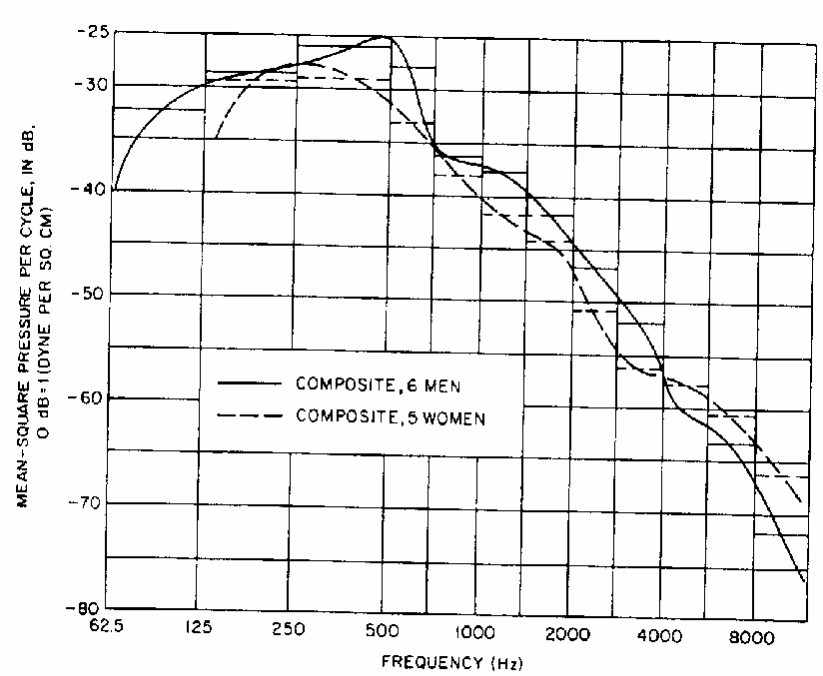


Fig. 5.6 Long-time power density spectrum for continuous speech. (After Dunn and White [5].)

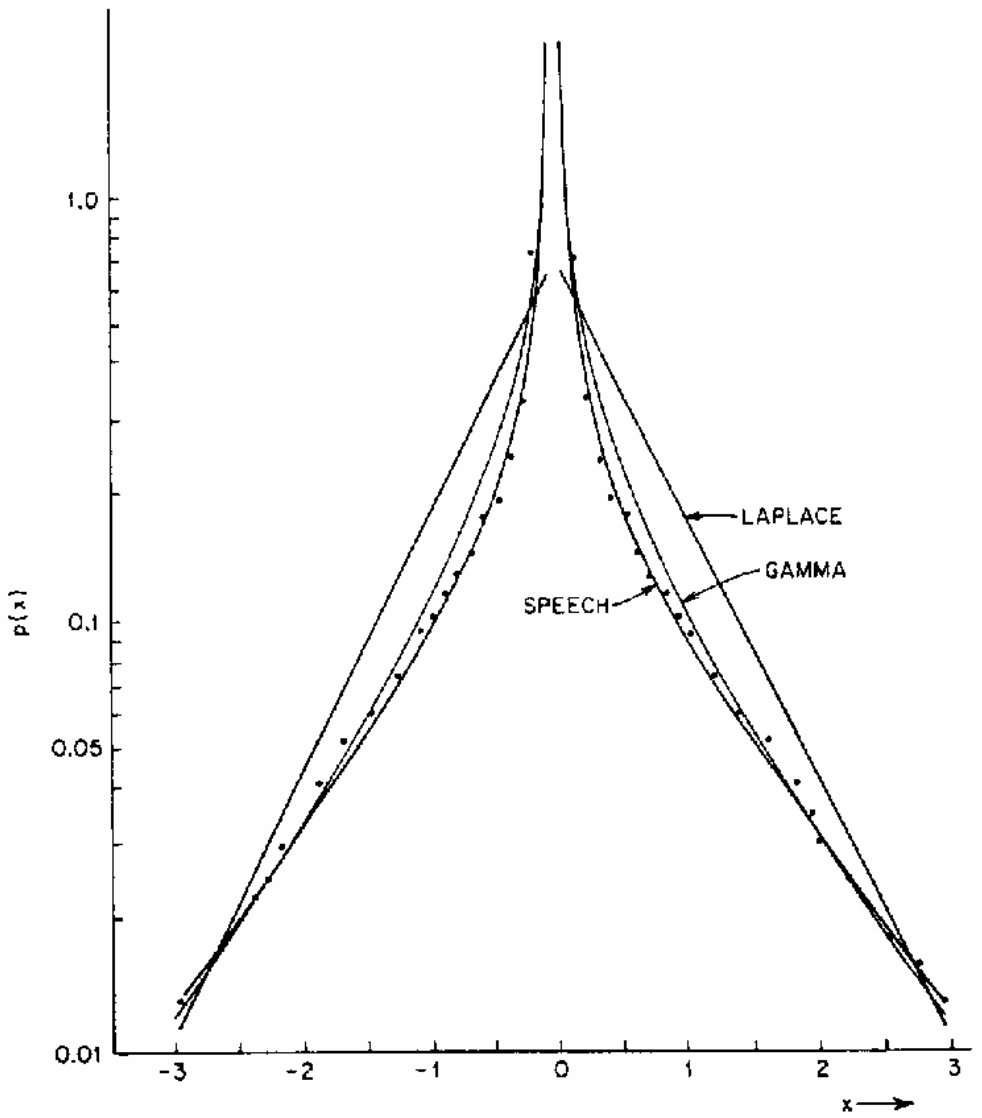
- Distribución de amplitudes

Distribución laplaciana

$$p(x) = \frac{1}{\sqrt{2}\sigma_x} e^{-\frac{\sqrt{2}|x|}{\sigma_x}}$$

Distribución gamma

$$p(x) = \left\{ \frac{\sqrt{3}}{8\pi\sigma_x|x|} \right\}^{1/2} e^{-\frac{\sqrt{3}|x|}{2\sigma_x}}$$



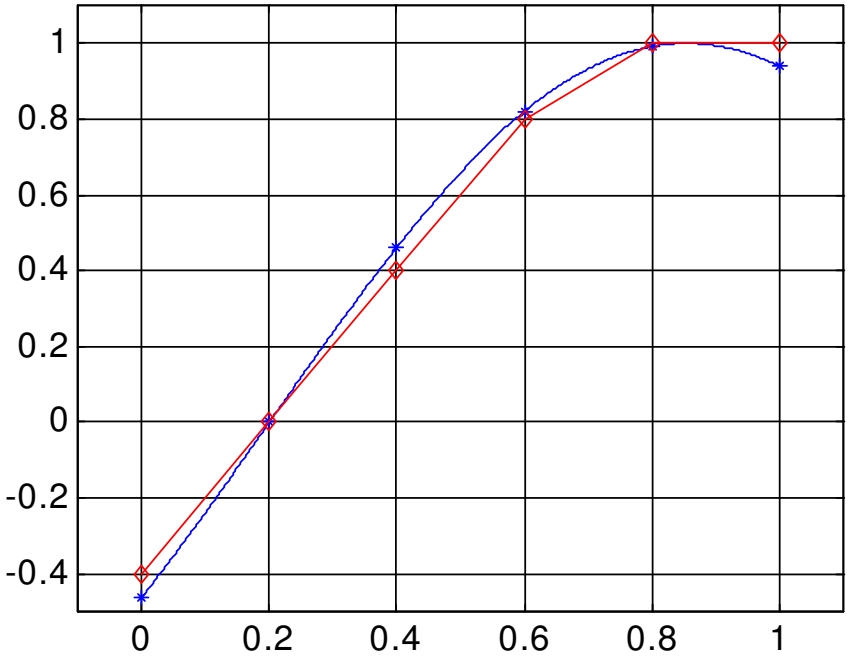
4.3.- Representación de la forma de onda

- **Proceso de representación:**
  - Amplificación y filtrado (analógico)
  - Muestreo
  - Cuantización de las muestras
  - Codificación de las muestras (representación digital de la forma de onda)
- **Pérdida de información asociada a la adquisición:**
  - Filtrado paso – baja previo al muestreo (eliminación de componentes de alta frecuencia)
  - Aliasing (si la frecuencia de muestreo no es suficientemente alta)
  - Ruido de cuantización
  - Otras fuentes de error (ruido del entorno, ruido de adquisición, ruido en la transmisión, errores de bit, errores de redondeo en las operaciones, etc.)
- **Evaluación del error: relación señal a ruido:**
  - $e = x' - x$ $\quad SNR = \sigma_x^2 / \sigma_e^2$ $\quad SNR(dB) = 10 \log_{10}(SNR)$

- Importancia de la relación señal ruido en audio y voz:
  - SNR = 80 dB Mínimo en equipos HiFi
  - SNR = 60 dB Típico en audio no HiFi. (El ruido se percibe si se presta atención)
  - SNR = 40 dB El ruido se percibe claramente
  - SNR = 25 dB Típico en comunicación telefónica de calidad
  - SNR = 15 dB El ruido es desagradable
  - SNR = 10 dB El ruido dificulta la inteligibilidad de la voz
  - SNR = 0 dB El ruido hace muy difícil la inteligibilidad de la voz

- **Amplificación:**
  - Permite adaptar la amplitud de la señal de entrada al rango dinámico del conversor analógico digital (compensar características del micrófono, separación entre el locutor y el micrófono, etc.)
- **Filtrado:**
  - Debe eliminar componentes de frecuencia superior a la mitad de la frecuencia de muestreo (para evitar aliasing). Filtrado analógico
  - Rizado en banda pasante; banda de transición; atenuación mínima
  - Para evitar problemas en el diseño, margen de frecuencia

- **Muestreo:**
  - Frecuencia de muestreo adecuada
  - Frecuencias típicas: 8 kHz; 11,025 kHz; 16 kHz; 20 kHz; 22,05 kHz; 44,1 kHz
- **Cuantización:**
  - Necesaria para representación digital de las muestras
  - Introduce un error de cuantización
  - Error de saturación
  - Cuantización uniforme:
    - 8 bits: SNR = 40 dB
    - 12 bits: SNR = 64 dB
    - 16 bits: SNR = 88 dB
  - Compresión instantánea:
    - Ley – mu ; Ley – A
    - 8 bits: SNR = 40 dB
  - Cuantización adaptable
  - Cuantización diferencial
  - Cuantización con predicción
- **Codificación de las muestras**



4.4.- Energía de tiempo corto

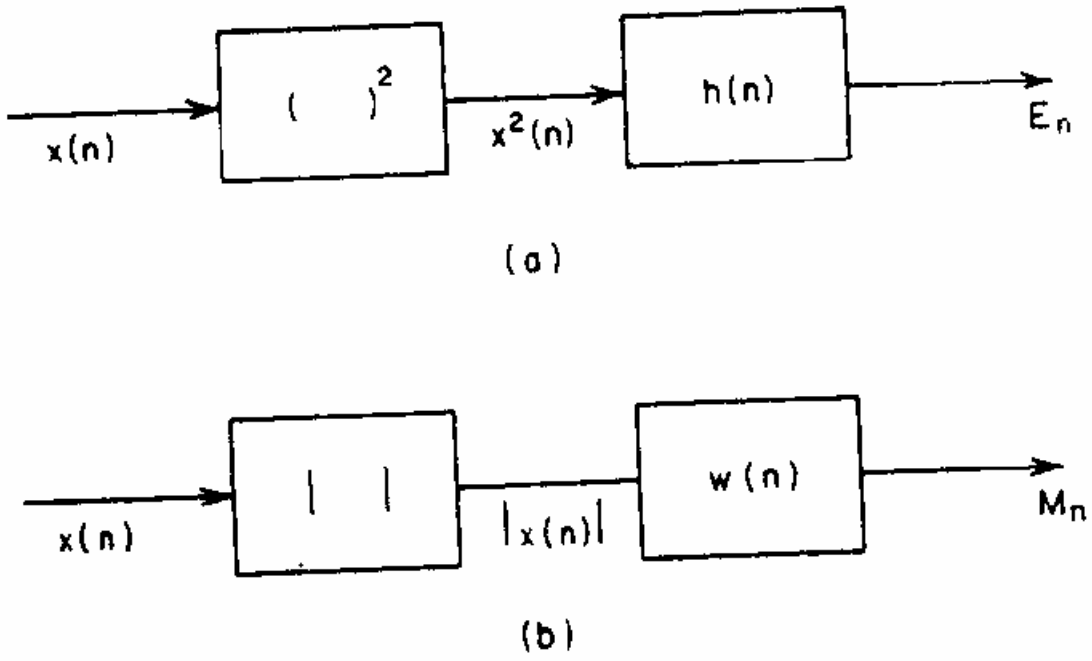
- **La energía de la señal varía en el tiempo:**
  - Fonemas sordos menor energía que fonemas sonoros
  - Consonantes sonoras menor energía que vocales
- **La energía de tiempo corto pone de manifiesto estas variaciones:**

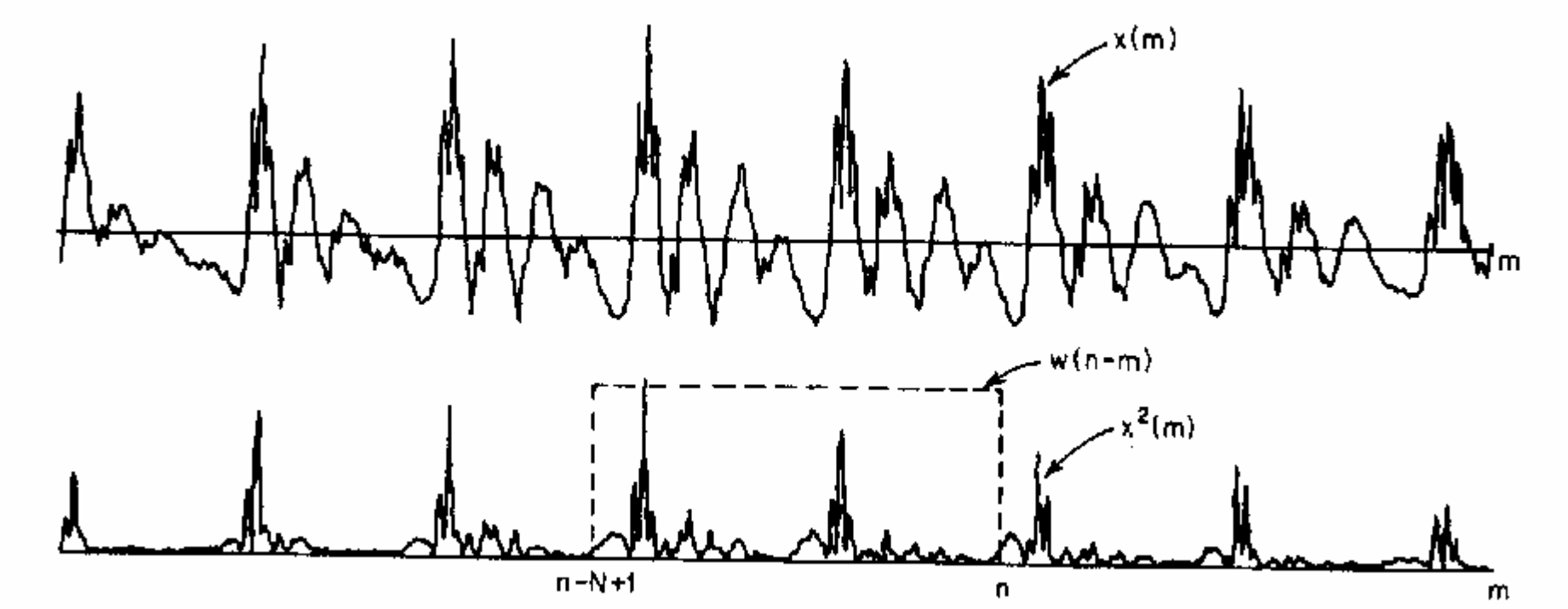
$$E_n = \sum_{m=-\infty}^{\infty} [x(m)\, w(n-m]^2 \qquad E_n = \sum_{m=-\infty}^{\infty} x^2(m) \cdot h(n-m)$$

- **Magnitud promedio de tiempo corto:**
  - Es una medida alternativa a la energía de tiempo corto
  - Es menos sensible a la amplitud de las muestras

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

- **Diagrama de bloques para estimar la energía o magnitud promedio:**
  - **Energía de tiempo corto:** elevar al cuadrado y ventana deslizante
  - **Magnitud promedio de tiempo corto:** valor absoluto y ventana deslizante
  - La ventana deslizante equivale a filtrado paso – baja
  - Como la variación de estos parámetros es lenta, no es necesario calcularla muestra a muestra (se puede calcular trama a trama)



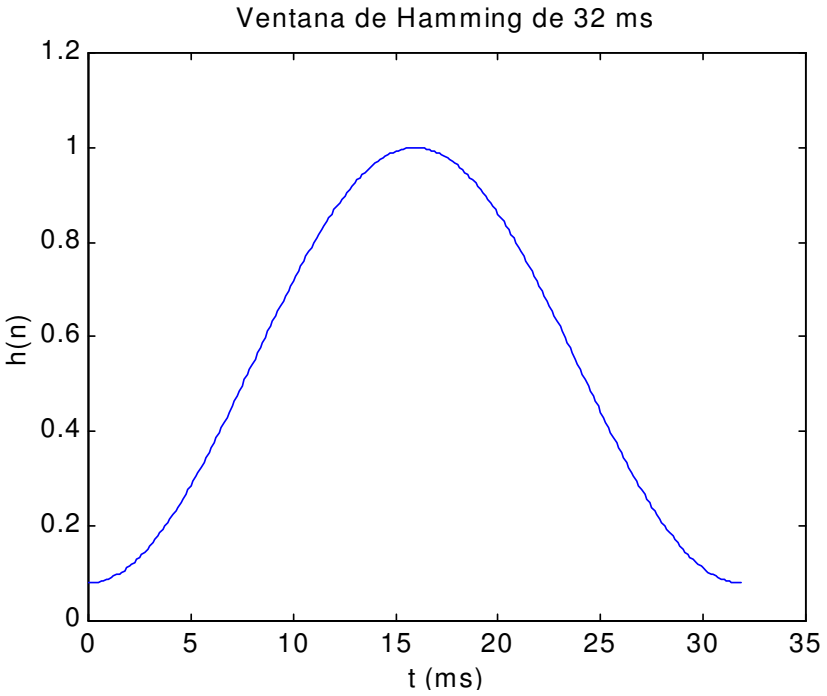


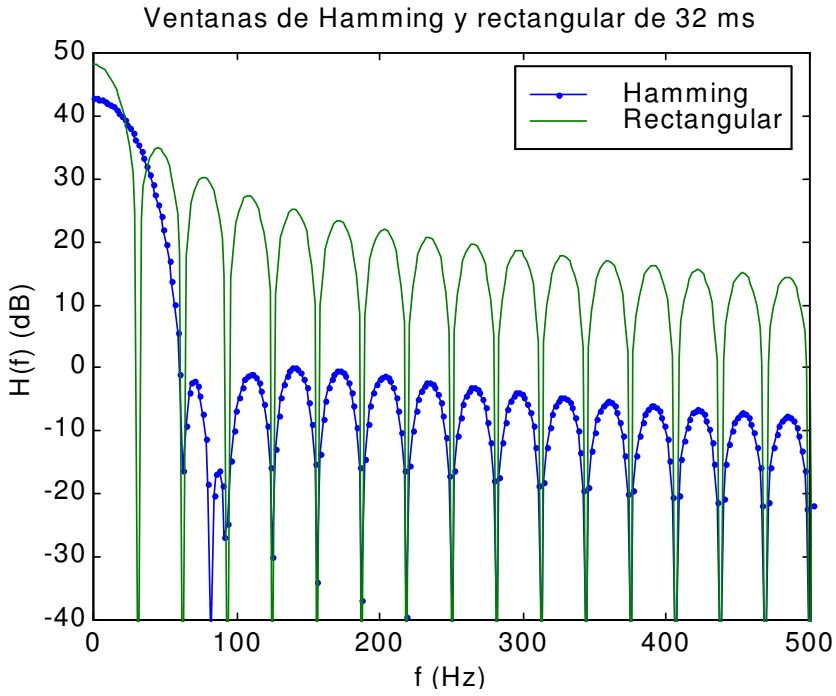
$$E_n = \sum_{m=-\infty}^{\infty} x^2(m) \cdot h(n-m)$$

- **Ventana aplicada:**
  - Forma de la ventana:
    - Rectangular
    - Hamming
  - Longitud de la ventana:
    - 4 ms; 8 ms; 16 ms; 32 ms; 64 ms

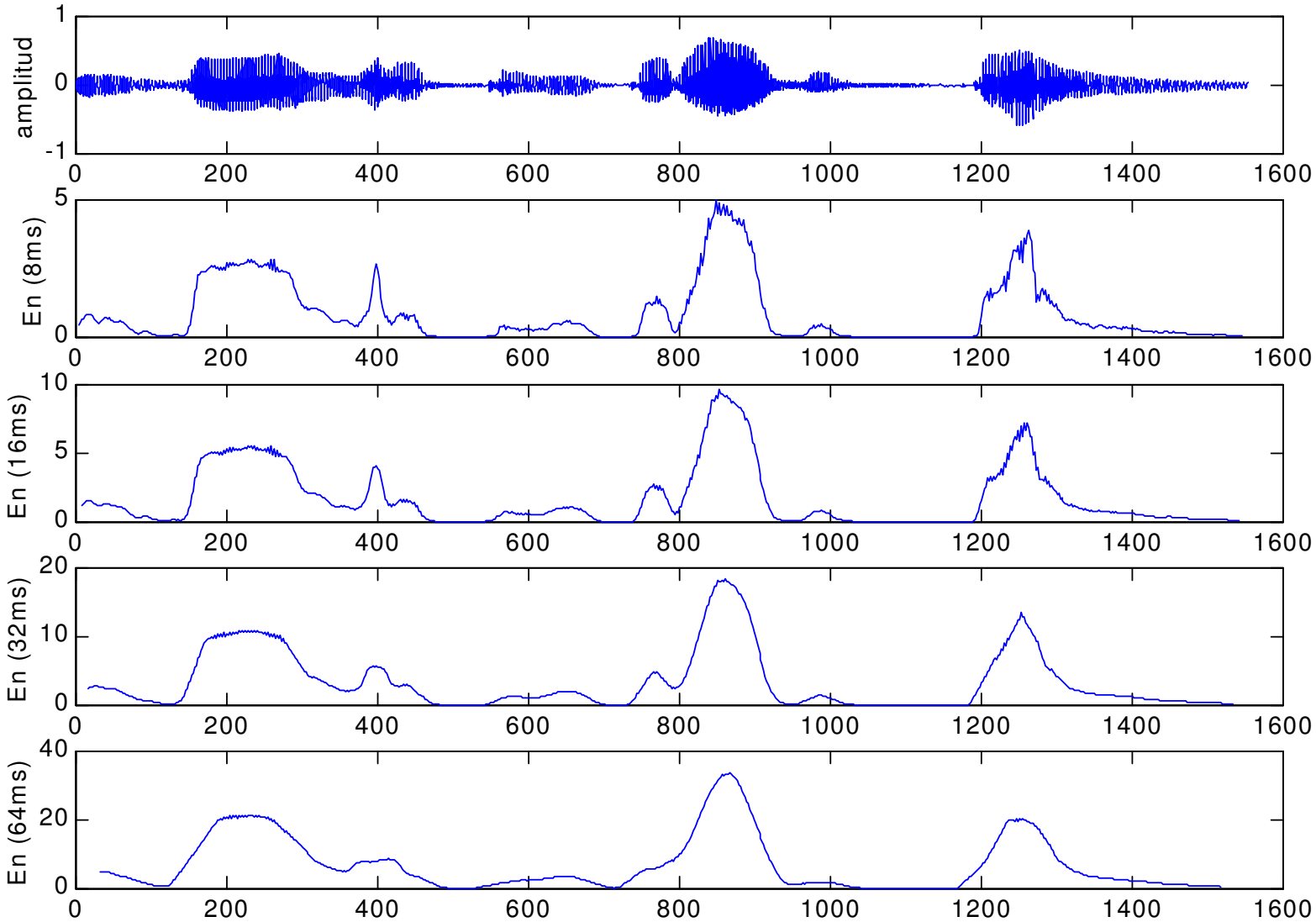
$$h(n) = 1 \quad 0 \leqslant n \leqslant N-1$$
$$= 0 \quad otherwise$$

$$h(n) = 0.54 - 0.46\cos(2\pi n/(N-1)), \quad 0 \leqslant n \leqslant N-1$$
$$= 0 \quad otherwise$$

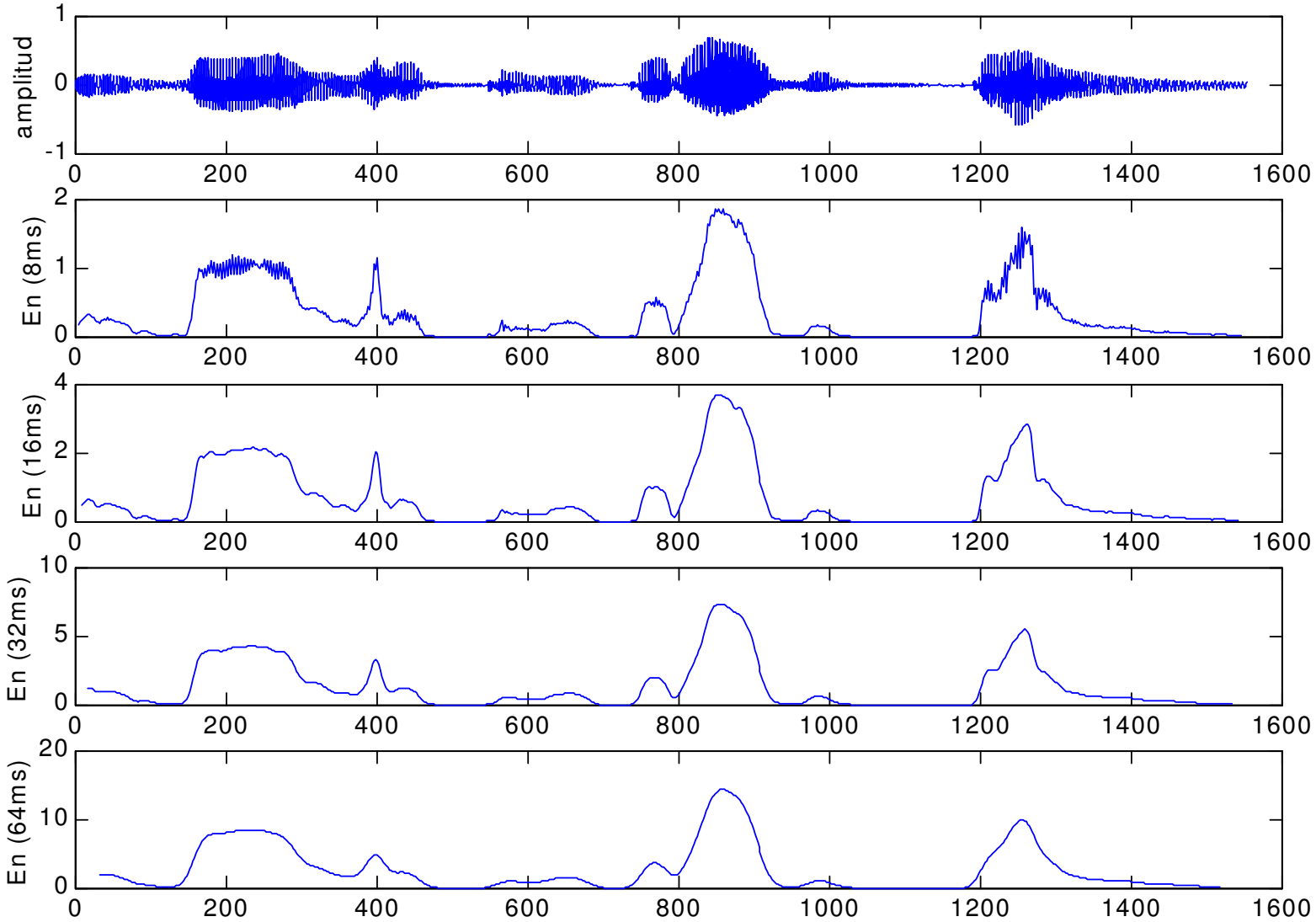




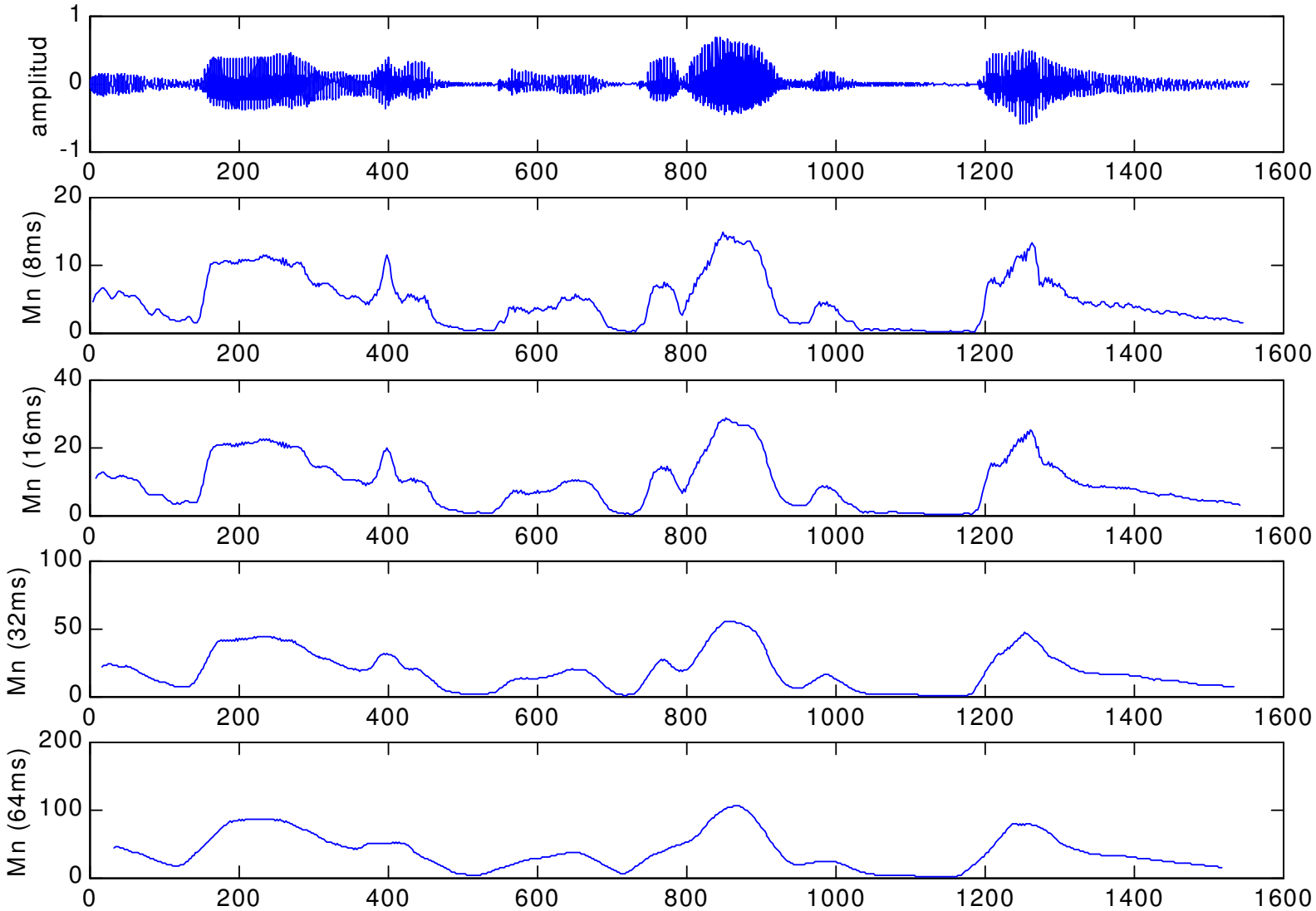
Energía de tiempo corto, ventana rectangular de distintas longitudes



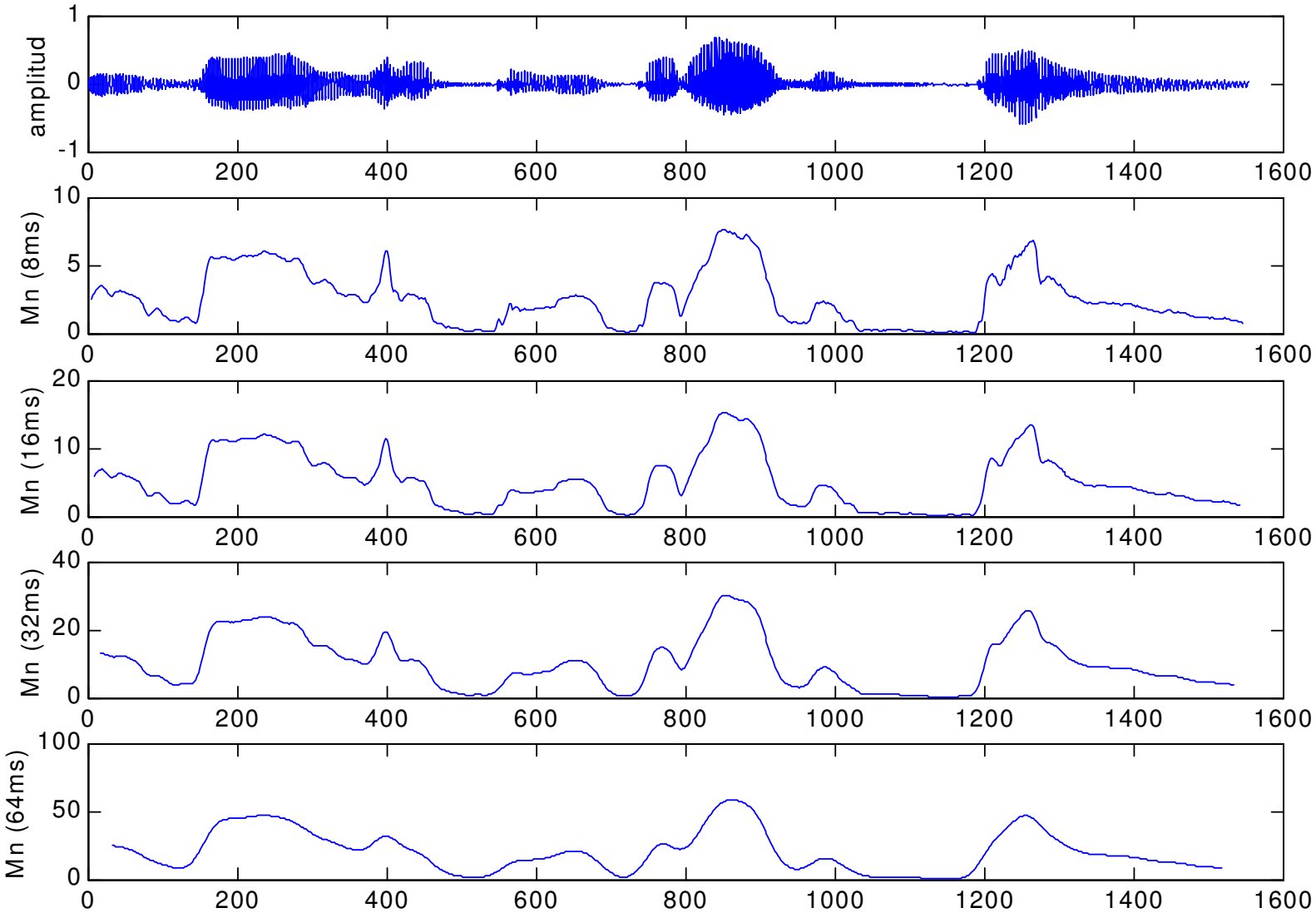
Energía de tiempo corto, ventana de Hamming de distintas longitudes



Magnitud promedio de tiempo corto, ventana rectangular de distintas longitudes



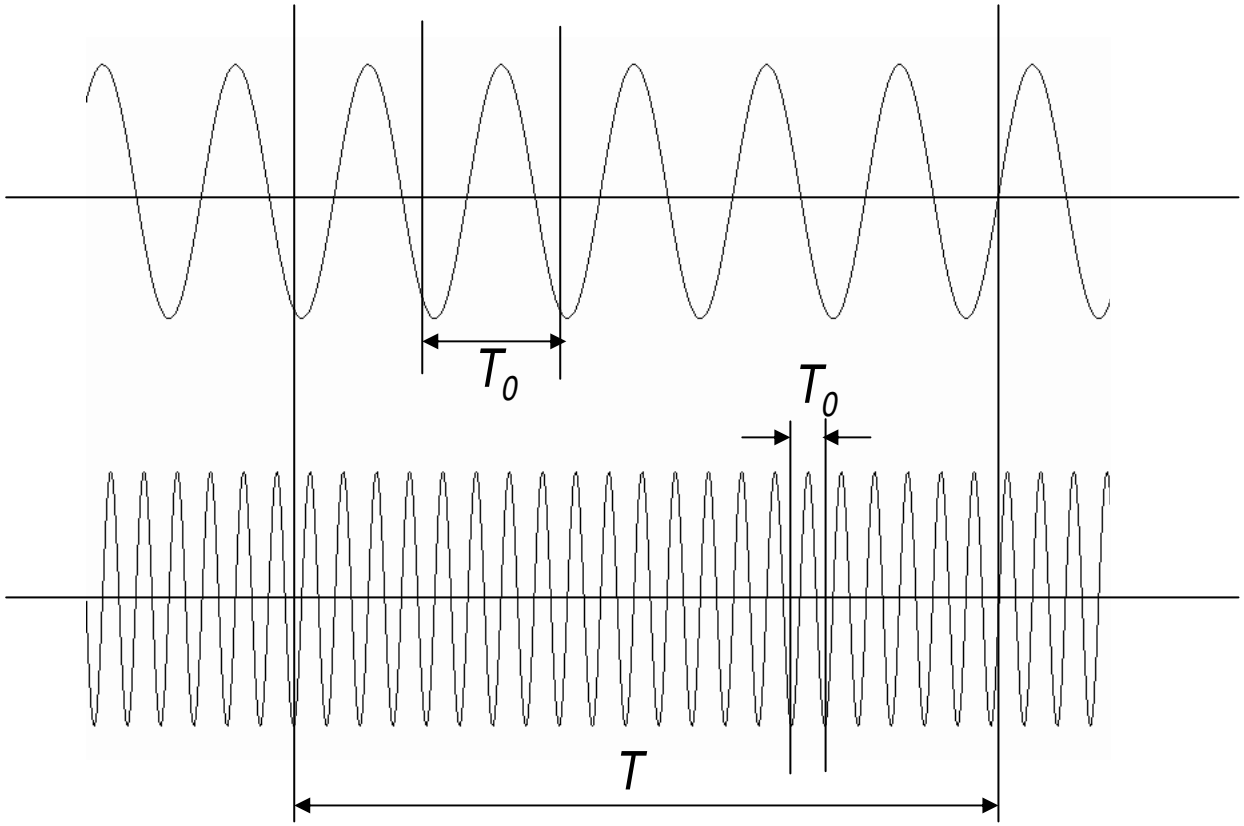
Magnitud promedio de tiempo corto, ventana de Hamming de distintas longitudes



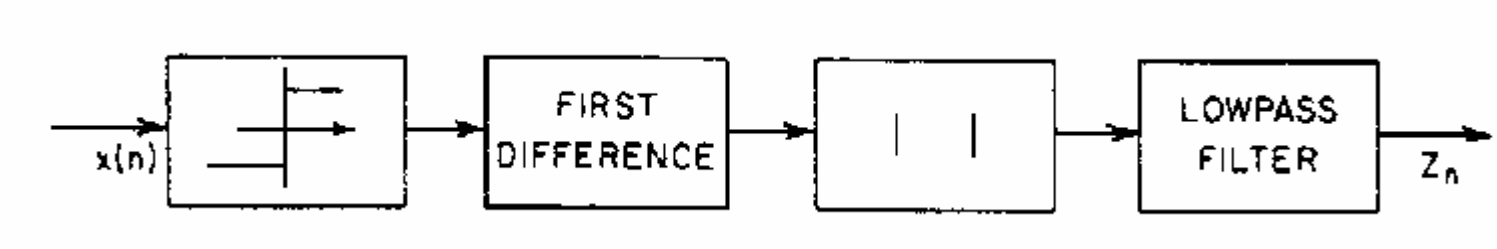
4.5.- Tasa promedio de cruces por cero

- **La tasa promedio de cruces por cero es proporcional a la frecuencia de la señal (para señales de banda estrecha):**

$$Z = 2\,T / T_0 = 2\,f_0\,T$$

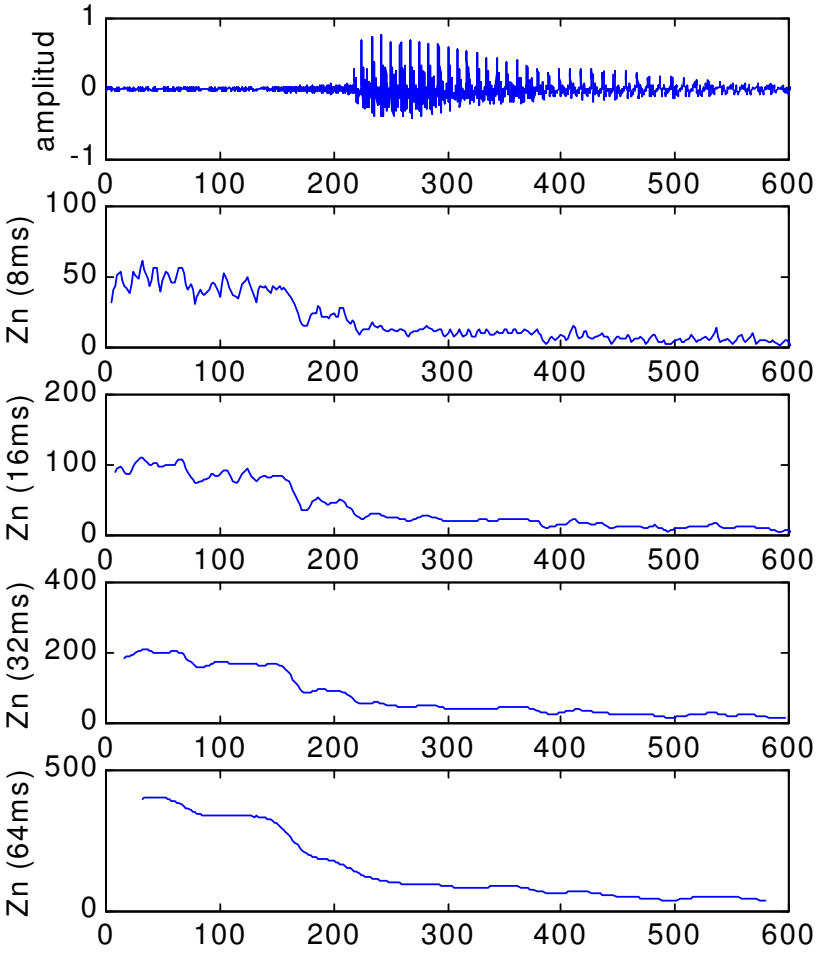


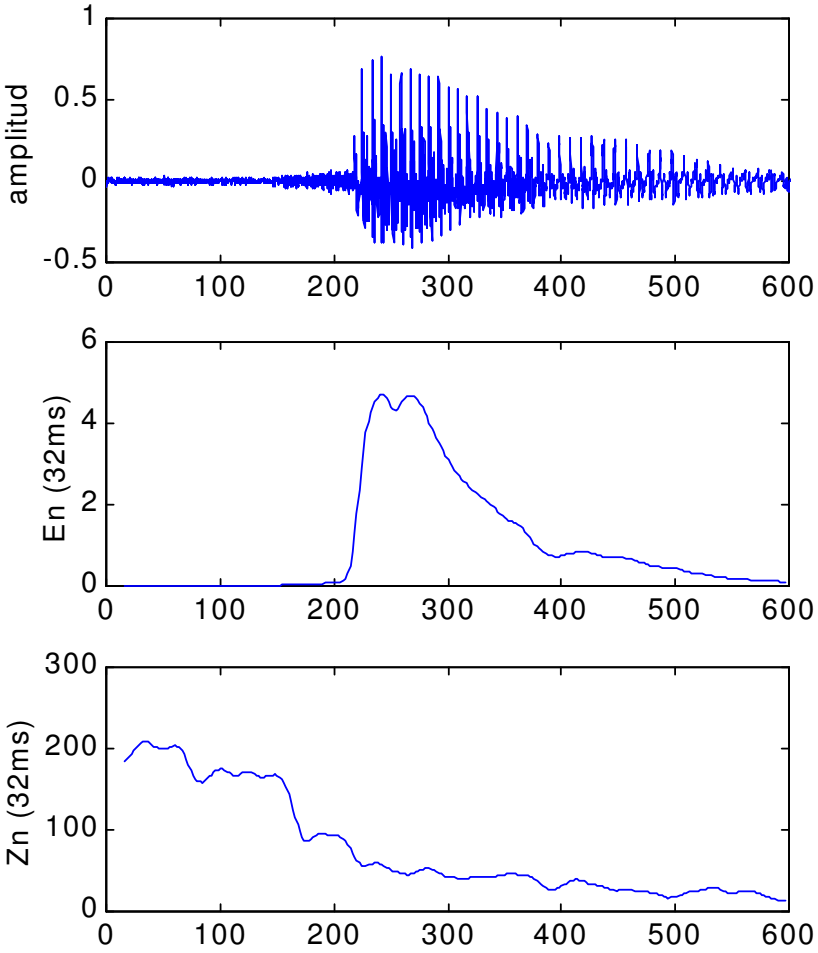
$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]| w(n-m)$$



- La tasa de cruces por cero mide el contenido de frecuencias
- Mayor para fonemas sordos que para sonoros
- Es independiente de la amplitud de la señal
- Es sensible al ruido
- Es sensible a nivel de continua
- Útil para detección de actividad de voz en combinación con energía
- Al igual que ocurría con la energía, el número de cruces por cero se cuenta sobre una ventana deslizante

Tasa promedio de cruces por cero para varias longitudes de ventana





4.6.- Función de autocorrelación de tiempo corto

- **Autocorrelación:**
  - Es la correlación de una señal consigo misma desplazada *k* muestras
  - Diversas expresiones
    - Señales deterministas (estacionarias)
    - Estimación trama a trama
    - Para evitar problemas de extremos de la trama

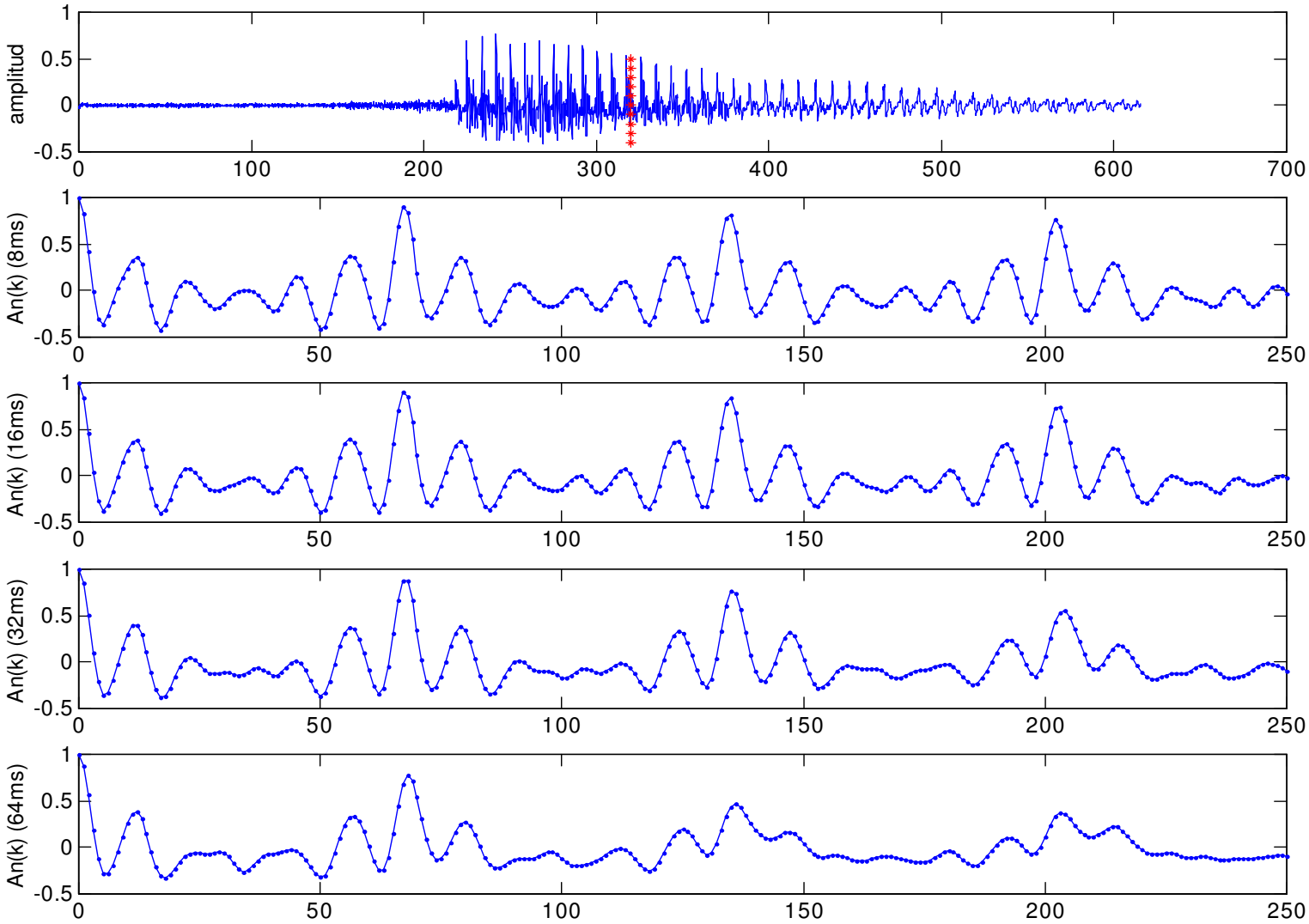
$$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k)$$

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)\,w(n-m)\,x(m+k)\,w(n-k-m)$$

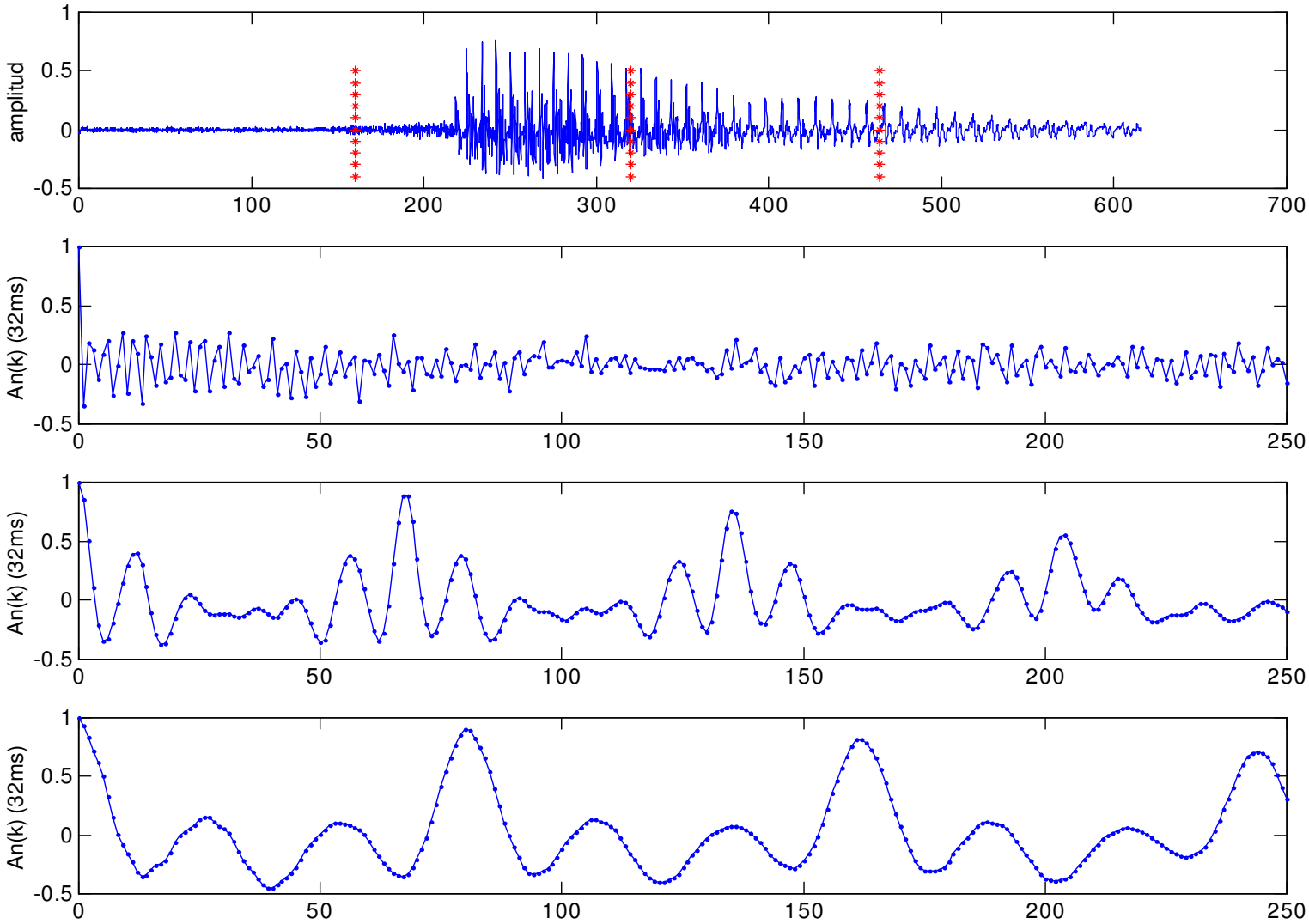
$$\hat{R}_n(k) = \sum_{m=0}^{N-1} x(n+m)x(n+m+k) \qquad 0 \leqslant k \leqslant K$$

- **Características de la función de autocorrelación:**
  - Si x(n) es periódica de periodo P, entonces también lo es la función de autocorrelación
  - Para $k = 0$ la función de autocorrelación proporciona la energía
- **La función de autocorrelación depende del tamaño de la ventana**
- **Primeros términos relacionados con envolvente espectral**
- **En segmentos sonoros es cuasi-periódica**
  - **Útil para medir la frecuencia fundamental (máximo después de k=0)**
- **En segmentos sordos cae muy rápidamente**

Función de autocorrelación de tiempo corto para distintas longitudes de ventana



Función de autocorrelación de tiempo corto para distintos fonemas



4.7.- Estimación del tono fundamental

- La periodicidad de la señal asociada a excitación periódica (pulsos glotales, vibración cuerdas vocales) da lugar al tono fundamental
- Si tenemos un fonema sonoro, la señal es cuasi – periódica, (P_0 f_0)
  - P_0 y f_0 están definidos para fonemas sonoros (no para fonemas sordos)
- La función de autocorrelación de tiempo corto de una señal periódica de periodo P_0, es periódica con periodo P_0:

$$R_n(k) = R_n(k+P) \qquad R_n(0) = R_n(P)$$

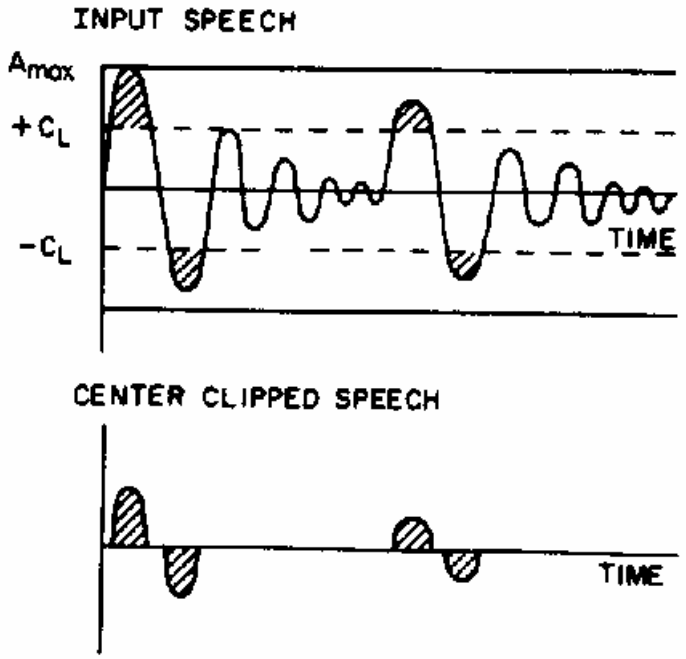
  - Como la señal es cuasiperiódica:

$$R_n(0) \approx R_n(P)$$

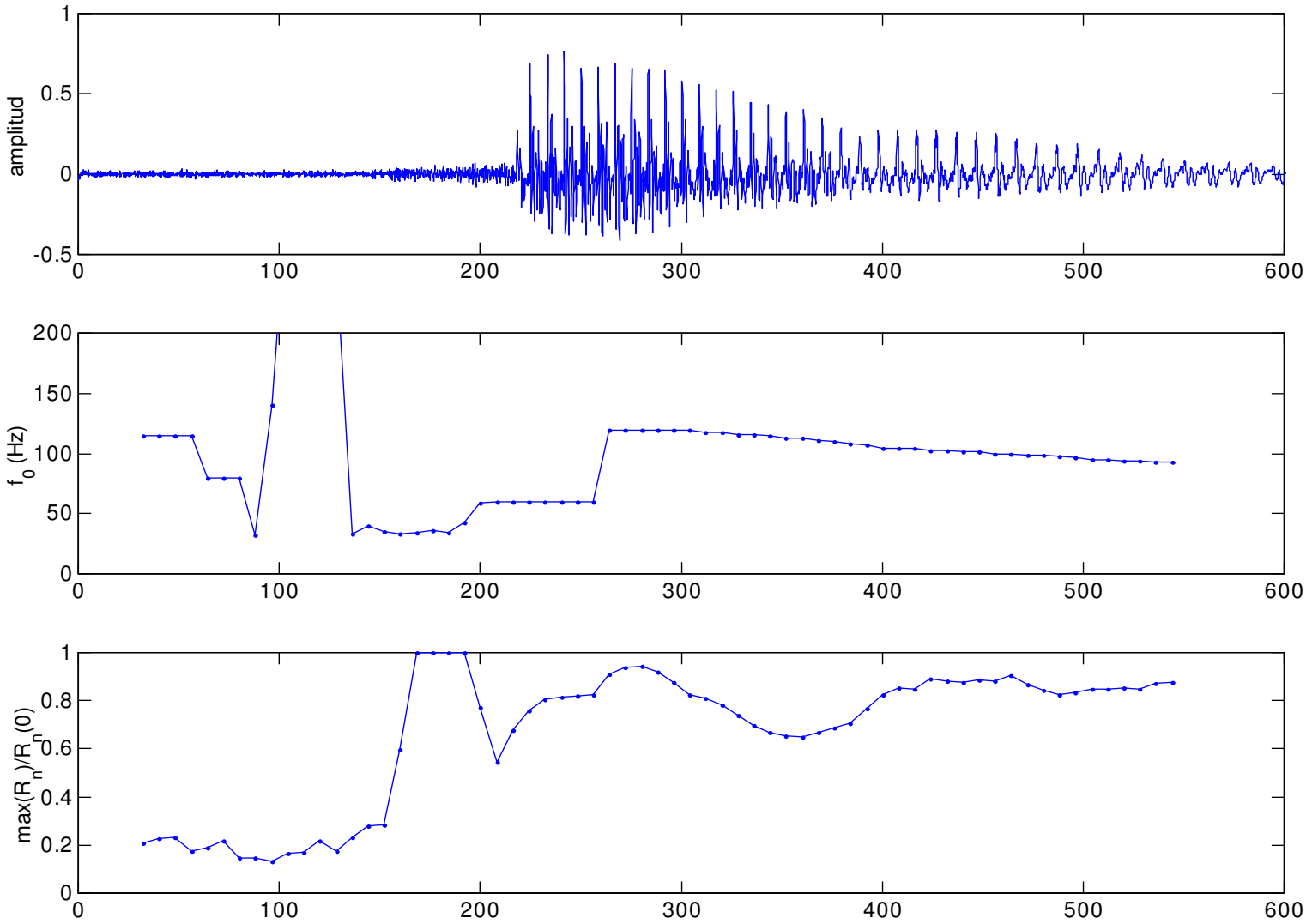
  - El pico más importante de la función de autocorrelación (después de $R_n(0)$) se presenta en la muestra correspondiente al periodo del pitch
  - La función de autocorrelación de tiempo corto permite estimar el periodo del pitch (y el tono fundamental) $f_0 = 1 / P_0$

- **Problemas de la estimación basada en la función de autocorrelación:**
  - Pico principal asociado a la cuasi – periodicidad
  - Picos al principio asociados a la respuesta del tracto vocal
  - Los picos del principio pueden ser mayores que el asociado a periodicidad (en algunos segmentos) dando lugar a estimaciones erróneas del tono fundamental

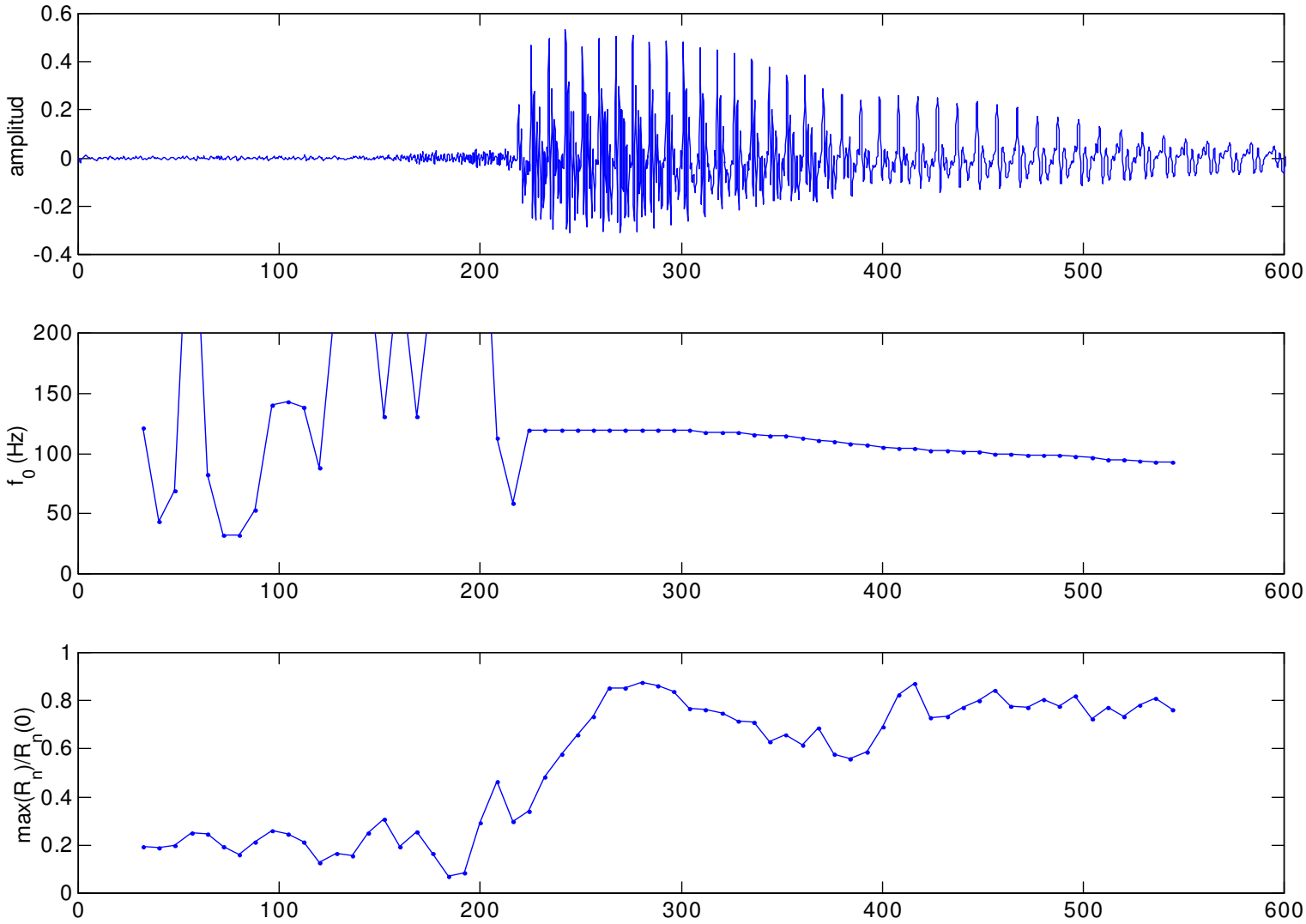
- **Soluciones:**
  - Filtrar la señal paso – baja o paso – banda (50 Hz – 900 Hz)
  - Center – clipping (nivel de clipping suele establecerse como porcentaje del máximo)
  - Autocorrelación sobre la señal obtenida tras el center – clipping
  - Autocorrelación normalizada (con *Rn(0)*) y umbral para determinar si el segmento es sordo o sonoro



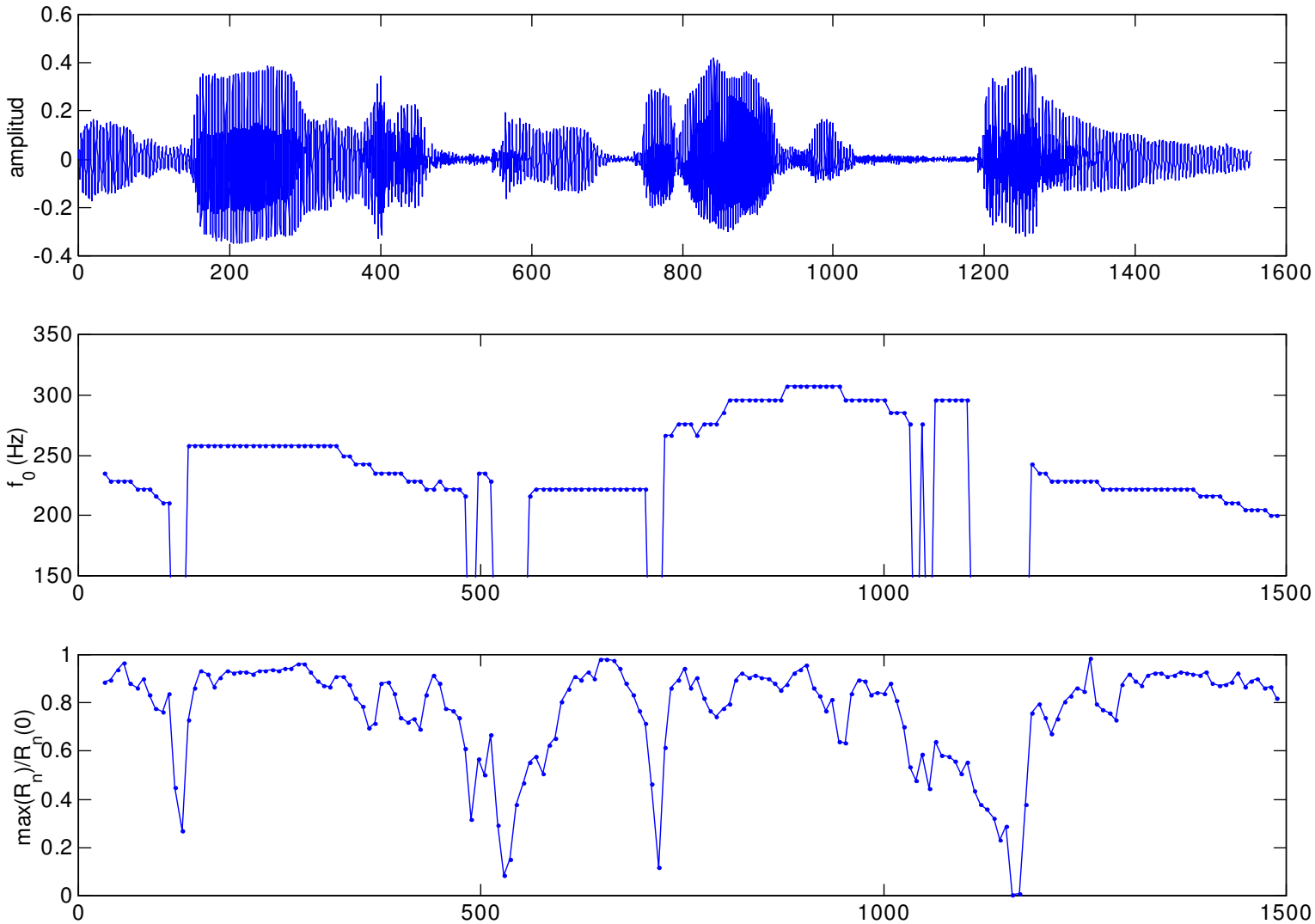
Estimación tono fundamental sin filtrado ni clipping



Estimación tono fundamental con filtrado y clipping

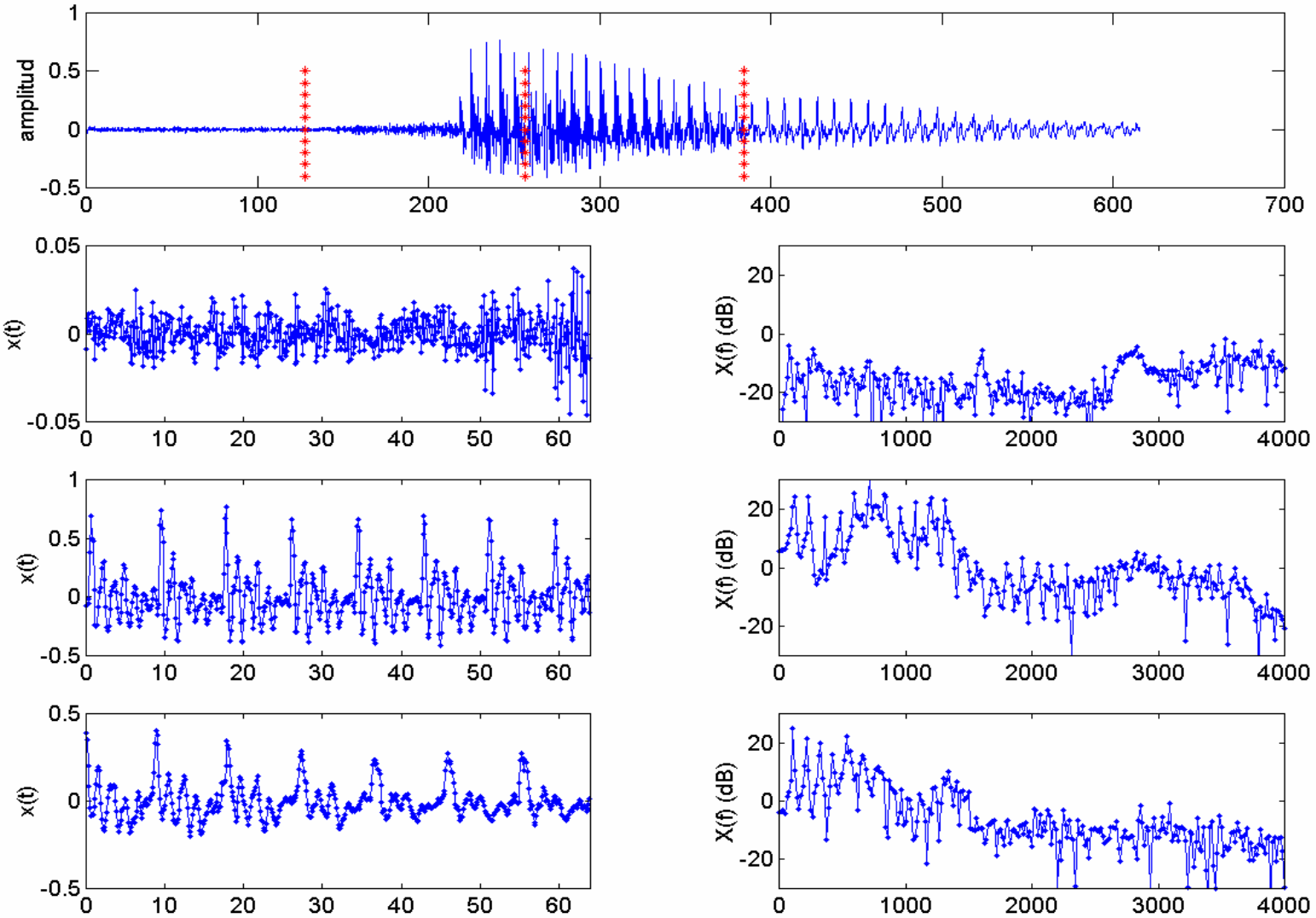


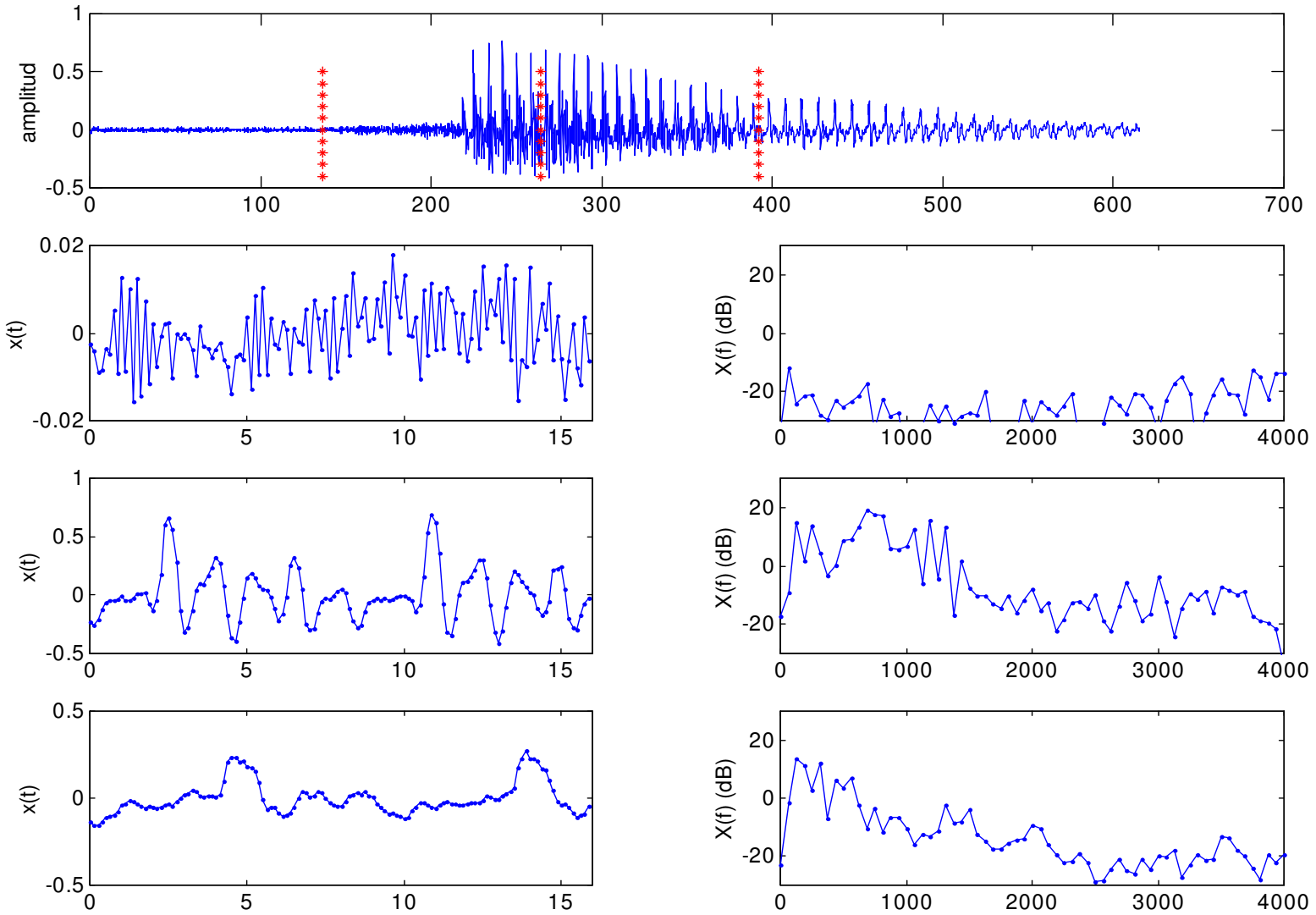
Estimación tono fundamental con filtrado y clipping (frase más larga)

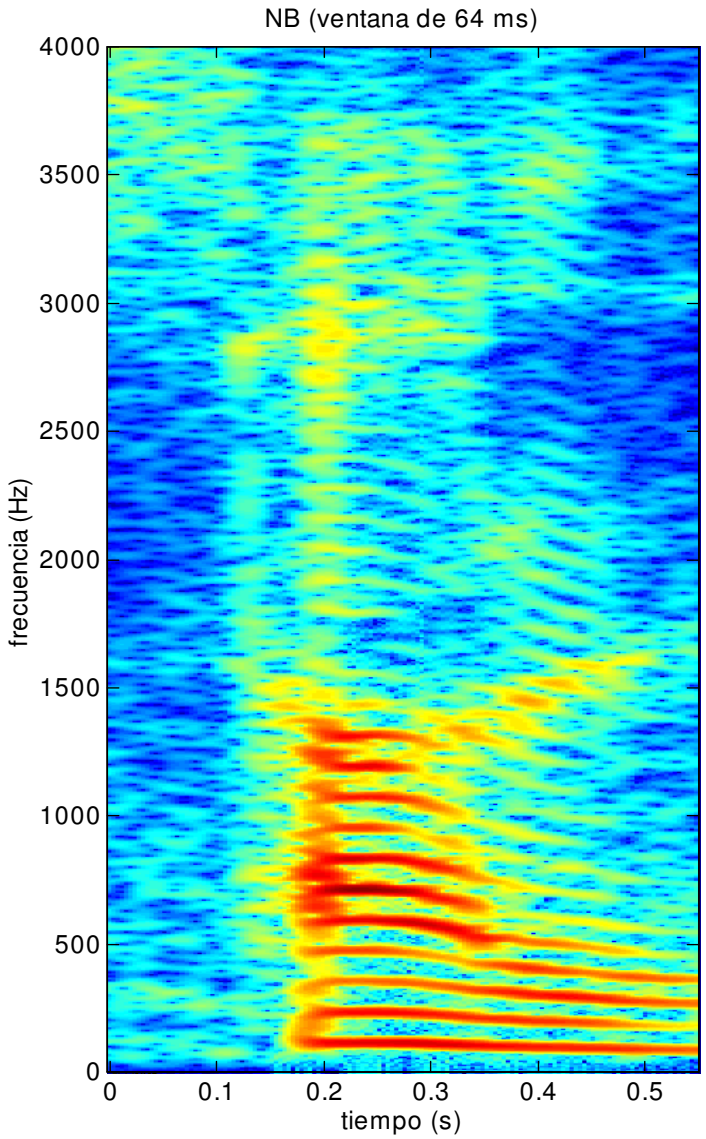


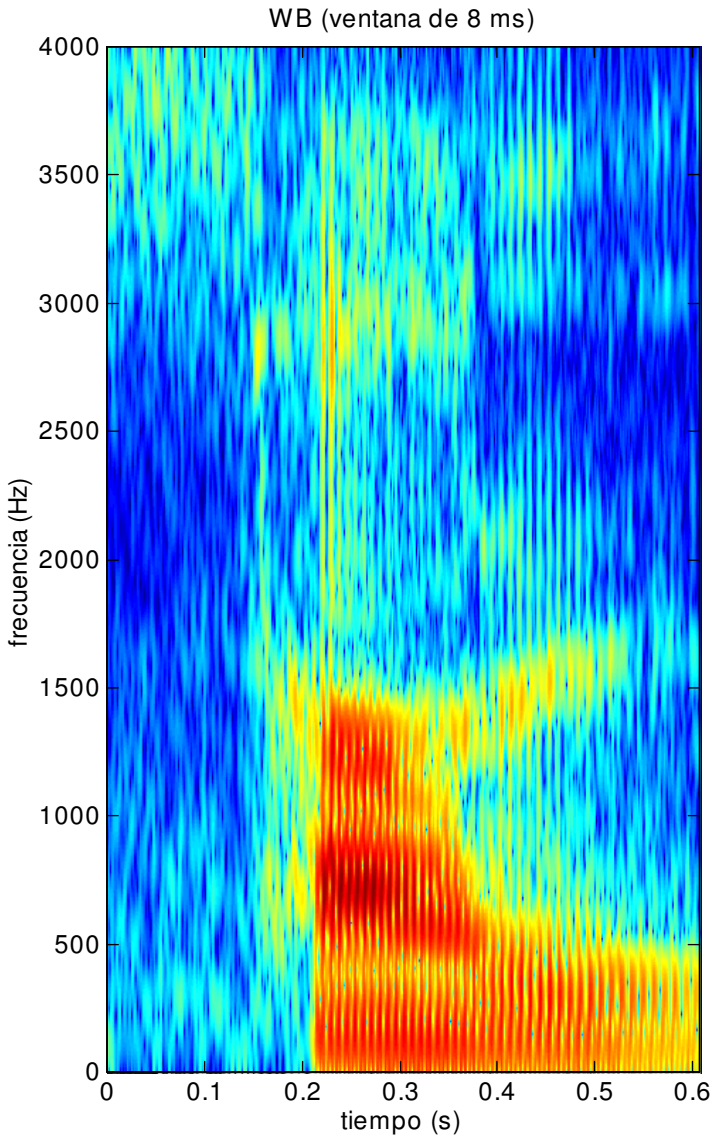
4.8.- Análisis de Fourier de tiempo corto: espectrogramas NB y WB

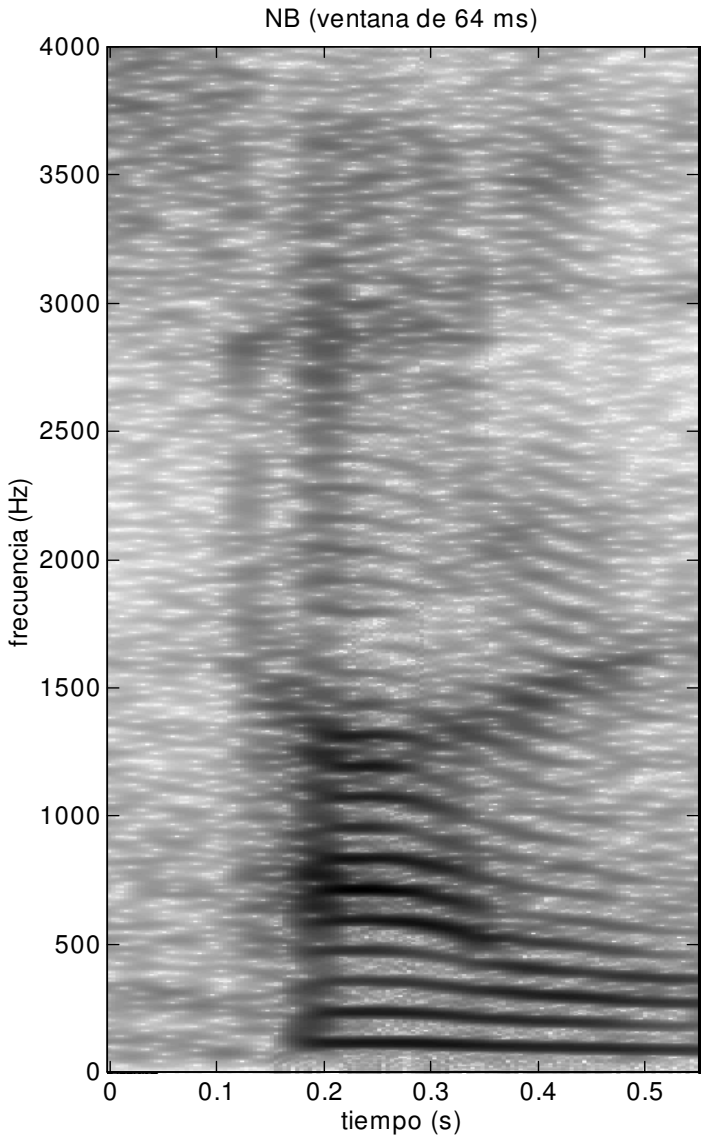
- **La FFT de una ventana proporciona el espectro de la porción de señal contenida en la ventana**
- **Espectrograma: representación del espectro de las distintas ventanas. Es una representación tridimensional:**
  - Eje de tiempo (para cada ventana)
  - Eje de frecuencia (para cada punto de la FFT)
  - Eje de amplitud (se suele representar el módulo de la FFT en dB)
- **Espectrograma típico:**
  - Eje horizontal para el tiempo
  - Eje vertical para la frecuencia
  - Amplitud representada mediante un mapa de color o nivel de gris

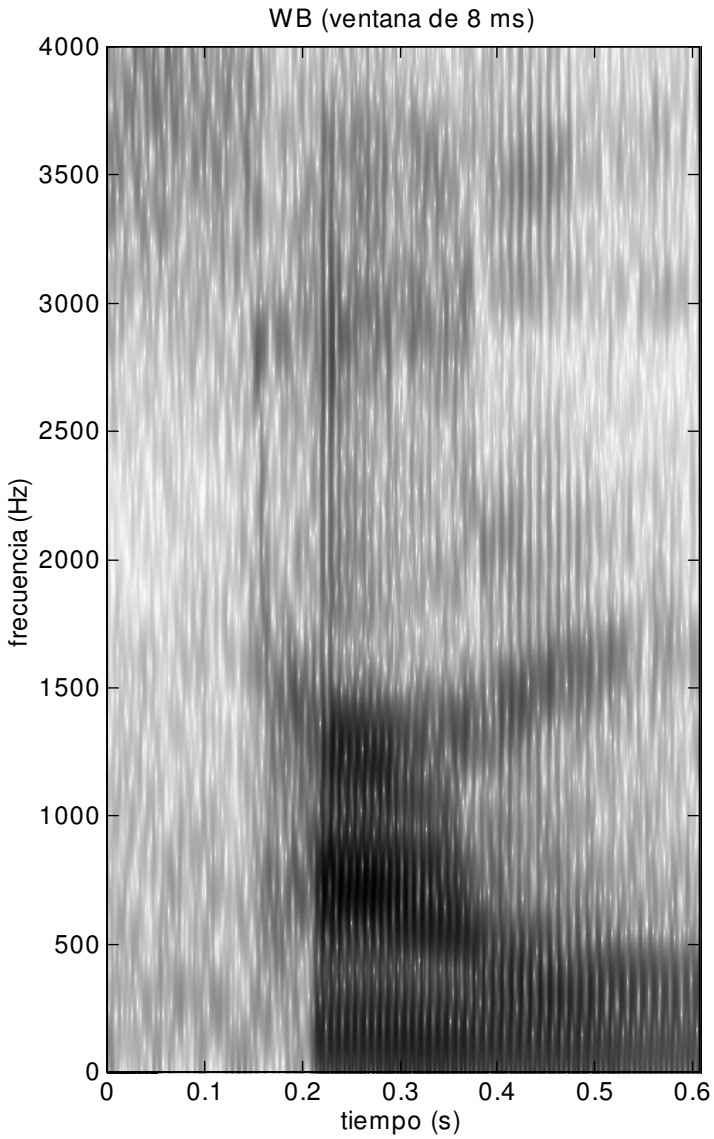

amplitud
x(t)
X(f) (dB)


amplitud
x(t)
X(f) (dB)

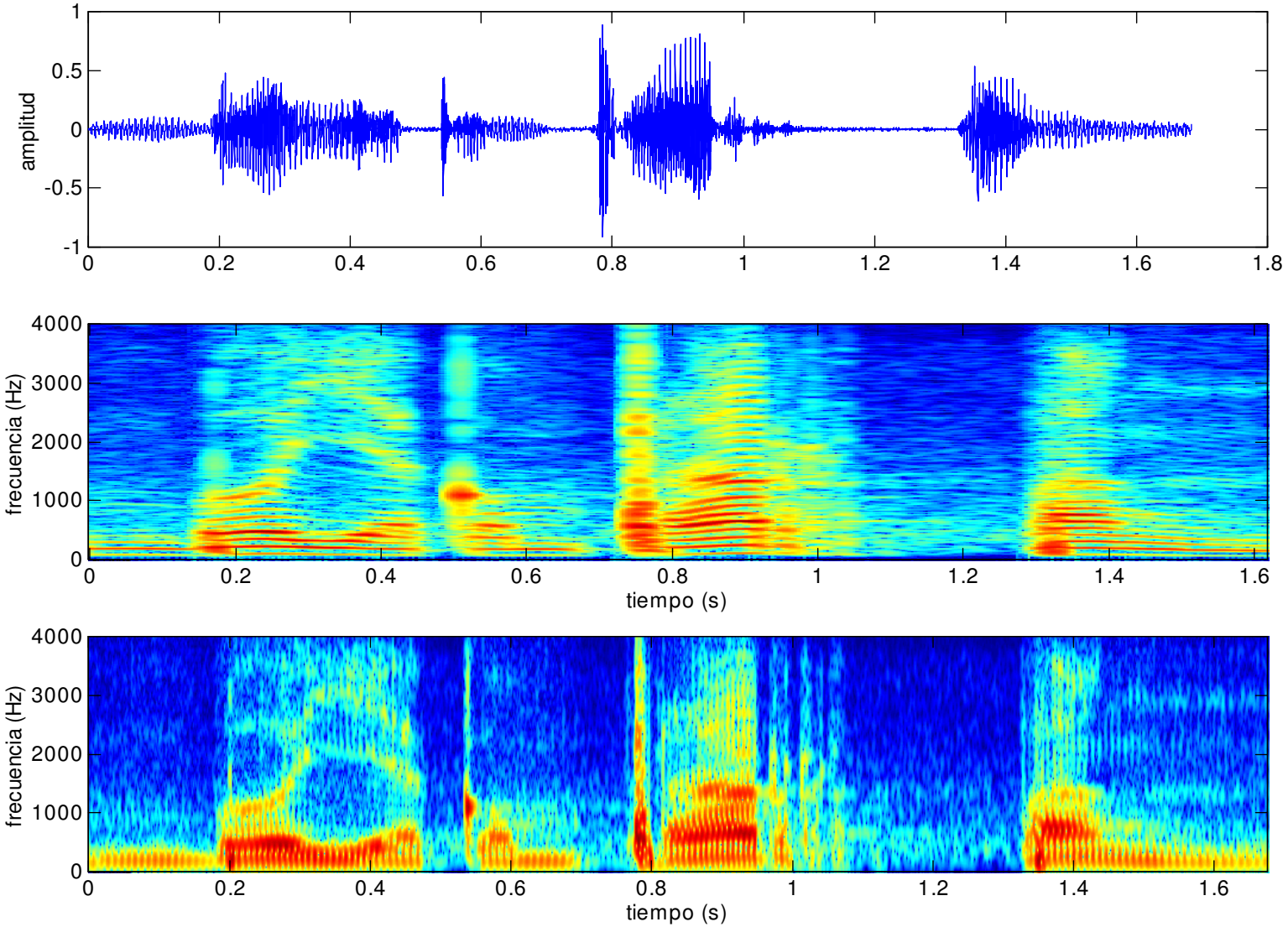

NB (ventana de 64 ms)
frecuencia (Hz)
tiempo (s)
0
500
1000
1500
2000
2500
3000
3500
4000
0
0.1
0.2
0.3
0.4
0.5


WB (ventana de 8 ms)
frecuencia (Hz)
tiempo (s)
0
500
1000
1500
2000
2500
3000
3500
4000
0
0.1
0.2
0.3
0.4
0.5
0.6


NB (ventana de 64 ms)
frecuencia (Hz)
4000
3500
3000
2500
2000
1500
1000
500
0
0
0.1
0.2
0.3
0.4
0.5
tiempo (s)


WB (ventana de 8 ms)
frecuencia (Hz)
4000
3500
3000
2500
2000
1500
1000
500
0
0
0.1
0.2
0.3
0.4
0.5
0.6
tiempo (s)

- **Utilidad del espectrograma:** representación global de la señal de voz
  - Características espectrales de tiempo corto (formantes)
  - Evolución de las características espectrales
  - Tono fundamental / periodo fundamental:
    - Representado en el dominio del tiempo (WB)
    - Representado en el dominio de la frecuencia (NB)
  - Se pueden identificar ('leer') fonemas del espectrograma
- **Importancia del tamaño de la ventana:**
  - Determina la resolución espectral
  - Determina la resolución temporal
  - Df = 45 Hz Dt = 22 ms NB
  - Df = 300 Hz Dt = 3.3 ms WB
- **Forma de la ventana: usualmente Hamming**


amplitud
1
0.5
0
-0.5
-1
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
frecuencia (Hz)
4000
3000
2000
1000
0
tiempo (s)
frecuencia (Hz)
4000
3000
2000
1000
0
tiempo (s)

- **Problema del espectrograma:**
  - Hay un compromiso entre resolución temporal y resolución espectral (incertidumbre)
  - En general interesan ventanas de unos 20 o 30 ms (cuasi-estacionariedad)
  - Dependiendo del propósito, interesan resoluciones espectrales peores que 50 Hz o 30 Hz (que corresponderían a NB)
  - Si queremos estudiar la envolvente espectral (resonancias) convendría suavizar el espectro

- **Técnicas de suavizado espectral:**
  - Espectro LPC
  - Banco de filtros
  - Procesamiento homomórfico (cepstrum)

4.9.- Linear Prediction Coding: Análisis LPC

- El análisis LPC trata de caracterizar el filtro *H(z)* que representa al tracto vocal, de acuerdo con el modelo excitación – filtrado
- Filtro todo – polos para modelar las resonancias (dos polos por cada resonancia). *A(z)* es un polinomio en z^{-1}

$$H(z) = \frac{G}{A(z)} \qquad A(z) = \sum_{k=0}^{p} a_k z^{-k} \quad (a_0 = 1)$$

- Salida del filtro cuando se presenta una excitación *u(n)*:

$$s(n) = Gu(n) - \sum_{k=1}^{p} a_k s(n-k)$$

- Predictor lineal: obtiene una predicción de s(n) en base a las p últimas muestras:

$$\tilde{s}(n) = -\sum_{k=1}^{p} \alpha_k s(n-k)$$

- Coeficientes de predicción lineal (LPC): los que minimizan el error de predicción:

$$e(n) = s(n) - \tilde{s}(n) = s(n) + \sum_{k=1}^{p} \alpha_k s(n-k)$$

$$\frac{\partial E}{\partial \alpha_k} = 0 \quad (k = 1, \ldots, p) \qquad \text{con} \qquad E = \sum_{n} e^2(n)$$

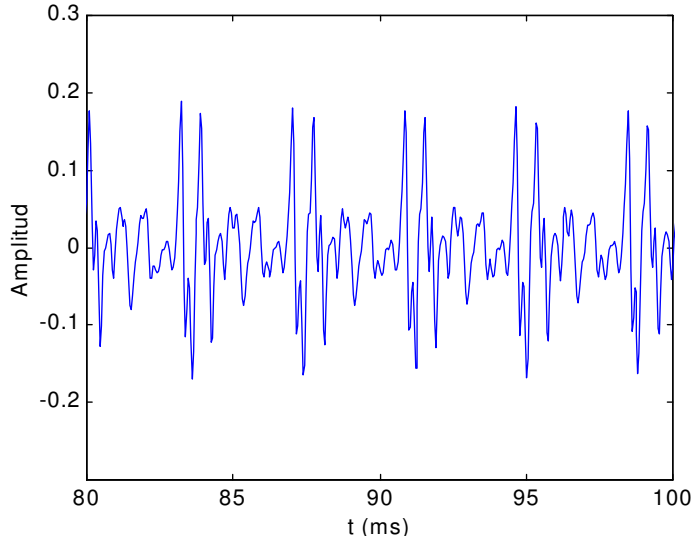
(la suma se extiende a la ventana de análisis)

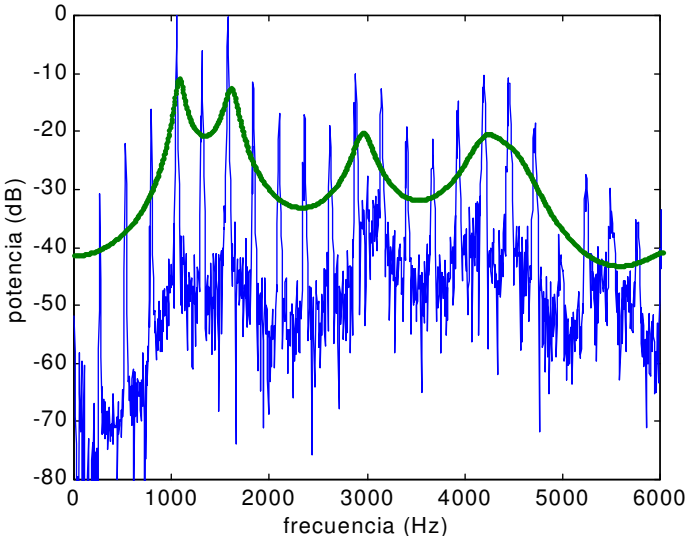
- Los coeficientes del predictor se pueden identificar con los coeficientes del filtro
  - El error de predicción es $G\,u(n)$
  - Para sonidos sonoros, la excitación $u(n)$ es un tren de impulsos (que vale cero para la mayor parte de las muestras). Esto es consistente con calcular los coeficientes del filtro como aquellos que minimizan la energía residual.
  - Si u(n) es un impulso simple o ruido blanco estacionario, el modelo AR (autoregresivo) garantiza que los coeficientes que minimizan la energía residual coinciden con los del filtro
- De este modo, el análisis LPC proporciona una estimación del filtro que representa el tracto vocal en el modelo excitación – filtrado

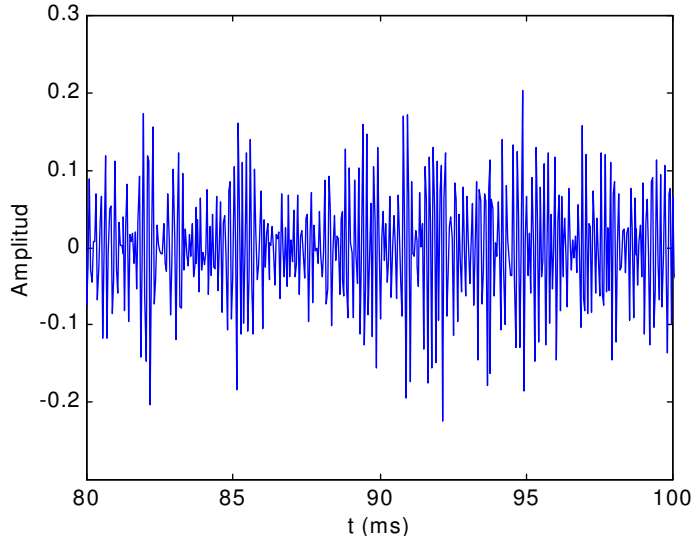
- **Espectro LPC**: es la respuesta en frecuencia del filtro. Para la frecuencia f se obtiene evaluando $H(z)$ en $z = e^{j2\pi f}$

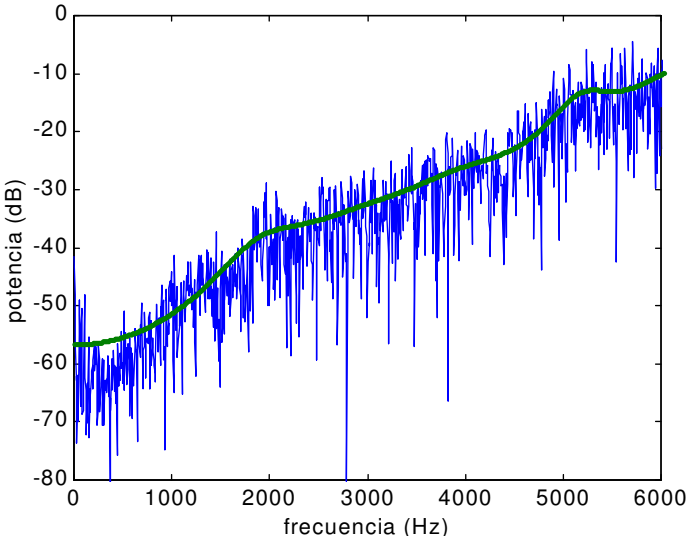
$$H(z) = \frac{G}{A(z)} \qquad A(z) = \sum_{k=0}^{p} a_k z^{-k} \quad (a_0 = 1)$$

Espectro FFT y LPC para fonemas sonoro y sordo: /a/ /s/ (ventana de 180 ms)





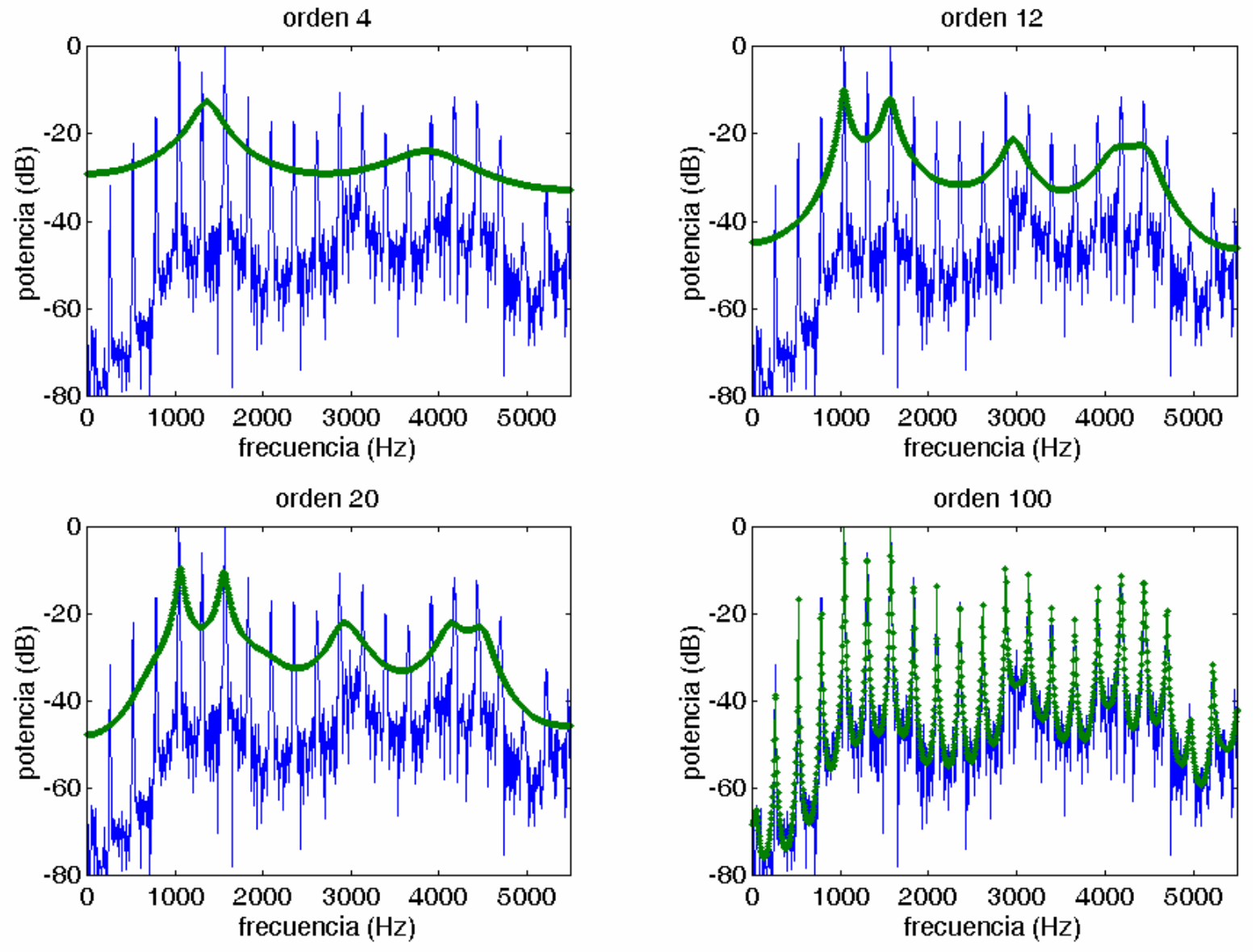




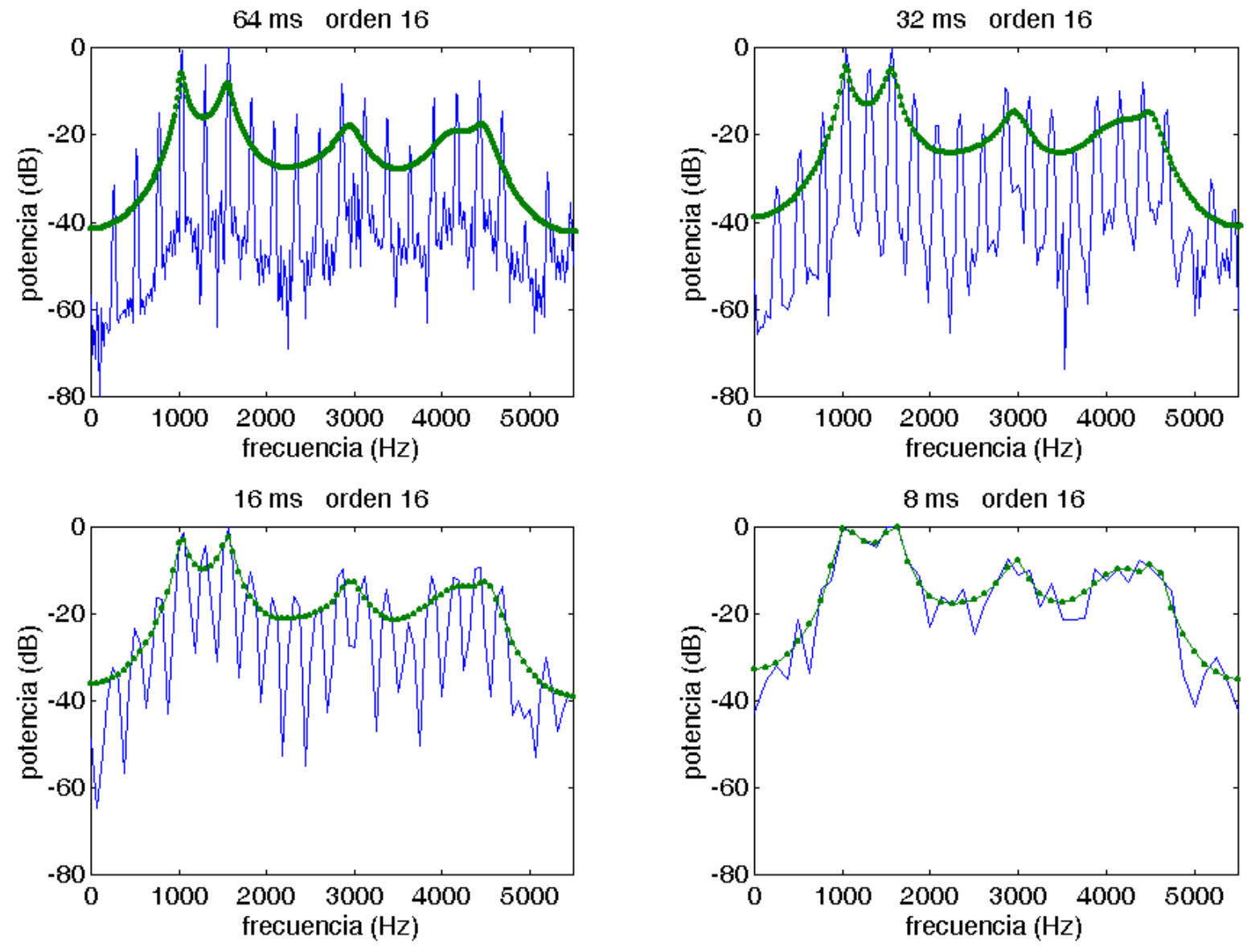
ORDEN DE PREDICCIÓN ADECUADO:

- El análisis LPC busca una resonancia por cada 2 polos
- Si buscamos la envolvente espectral, (los formantes), debemos utilizar un orden aproximado de 2 x Numero-de-formantes (algo mayor)
- Como hay aproximadamente 1 formante por cada kHz, para frecuencia de muestreo de 8 kHz (se analiza entre 0 y 4 kHz) se debe usar aproximadamente orden 8
- Para ajustar mejor el espectro LPC a la envolvente espectral, se puede usar un orden un poco mayor (por ejemplo, 12 o 14)
- Si se usa un orden excesivo, el espectro LPC se ajusta a los armónicos

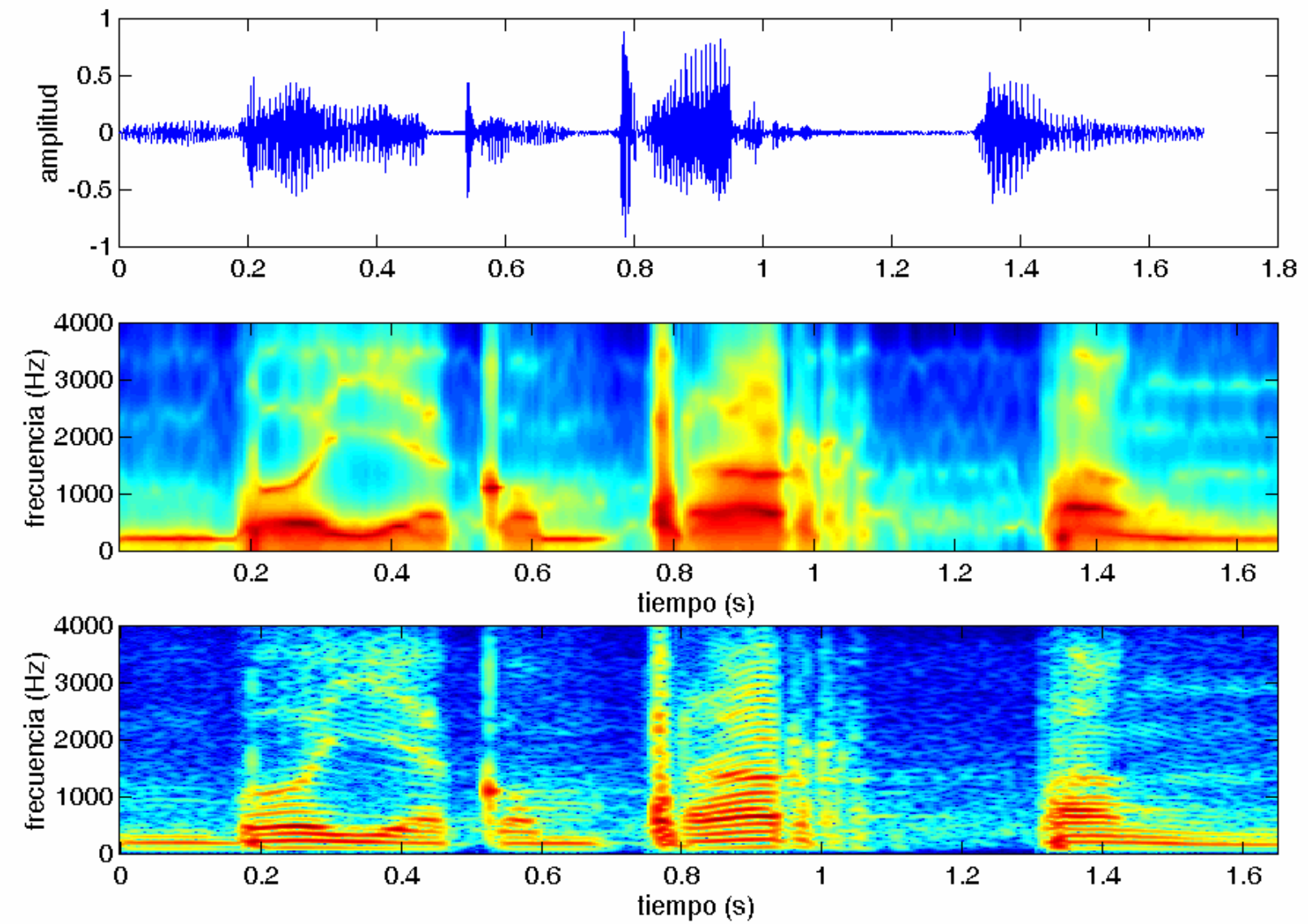
Influencia del orden de predicción (señal muestreada a 11 kHz)



Influencia del tamaño de la ventana (señal muestreada a 11 kHz; orden LPC 16)



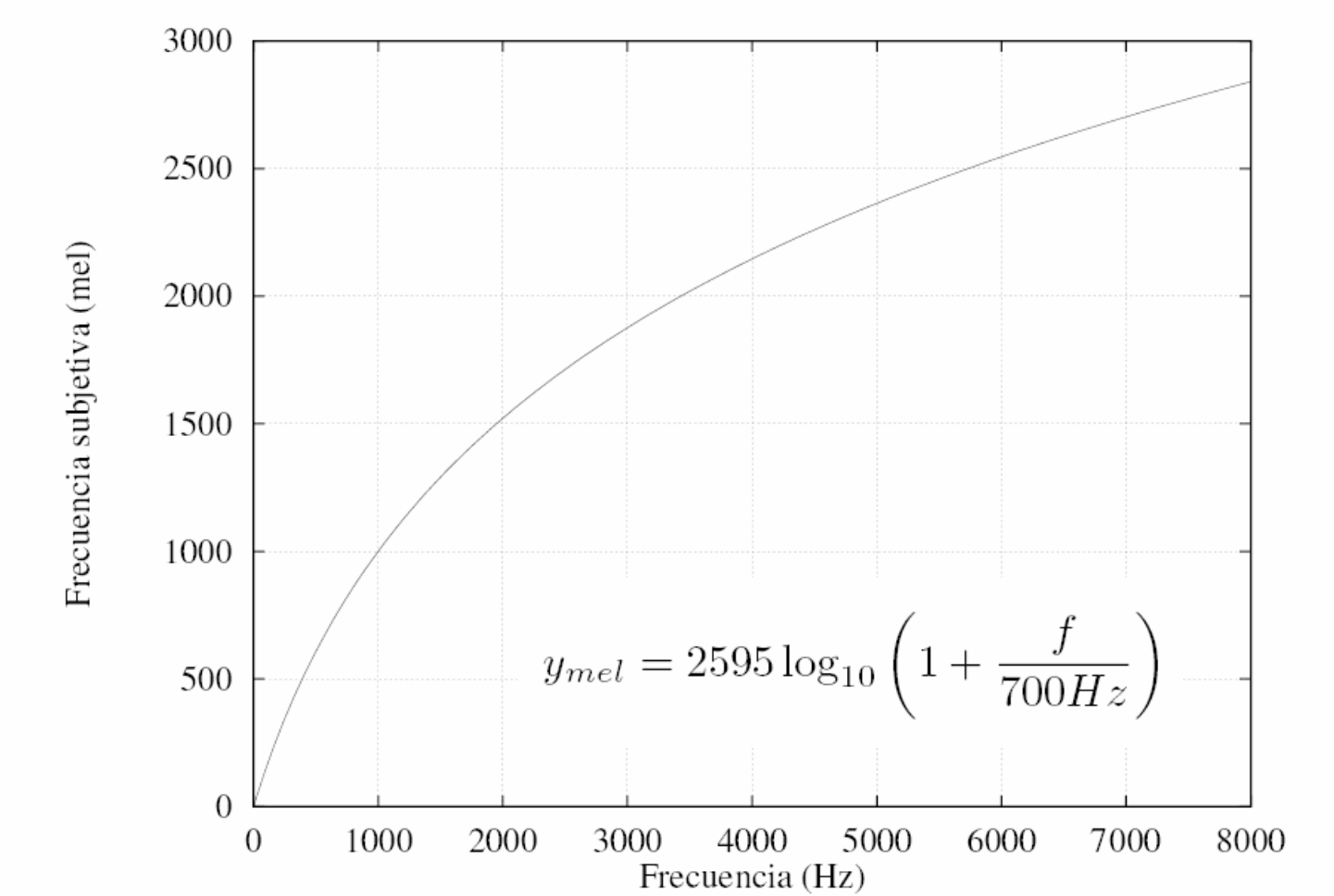
Espectrogramas FFT y LPC (ventana de 32 ms; orden LPC 12)

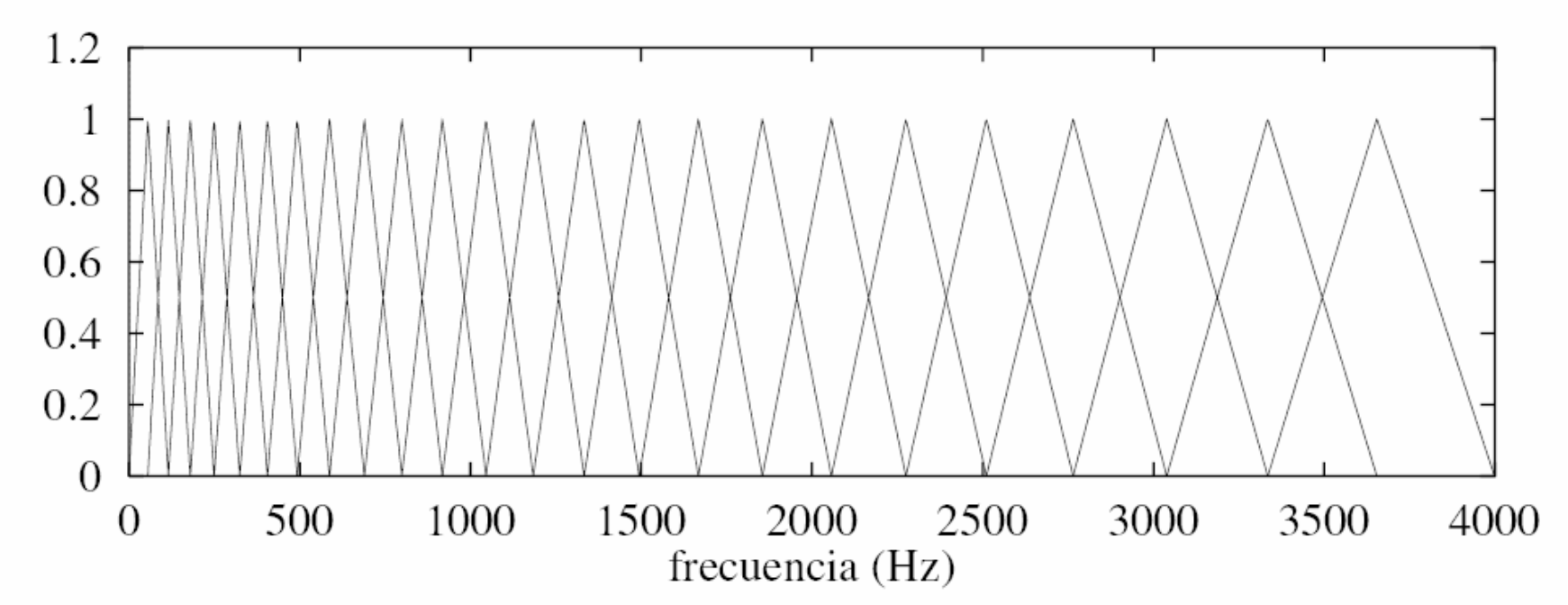


4.10.- Análisis basado en banco de filtros

- El análisis basado en banco de filtros proporciona un espectro suavizado
- La resolución espectral viene determinada por el número de filtros y el ancho de banda de éstos
- El banco de filtros se puede diseñar con distintos criterios:
  - Distribución de los filtros a lo largo del eje de frecuencia
  - Ancho de banda para cada frecuencia
- Los filtros se pueden implementar por distintos métodos:
  - En el dominio del tiempo
  - En el dominio de la frecuencia (segmentación en tramas y FFT)
- Dependiendo de la aplicación, se puede usar la salida de cada filtro, la envolvente de la salida, o la potencia de salida

- Escala Mel
  - Distribuye los filtros de forma uniforme desde un punto de vista perceptual
  - Compresión del eje de frecuencia

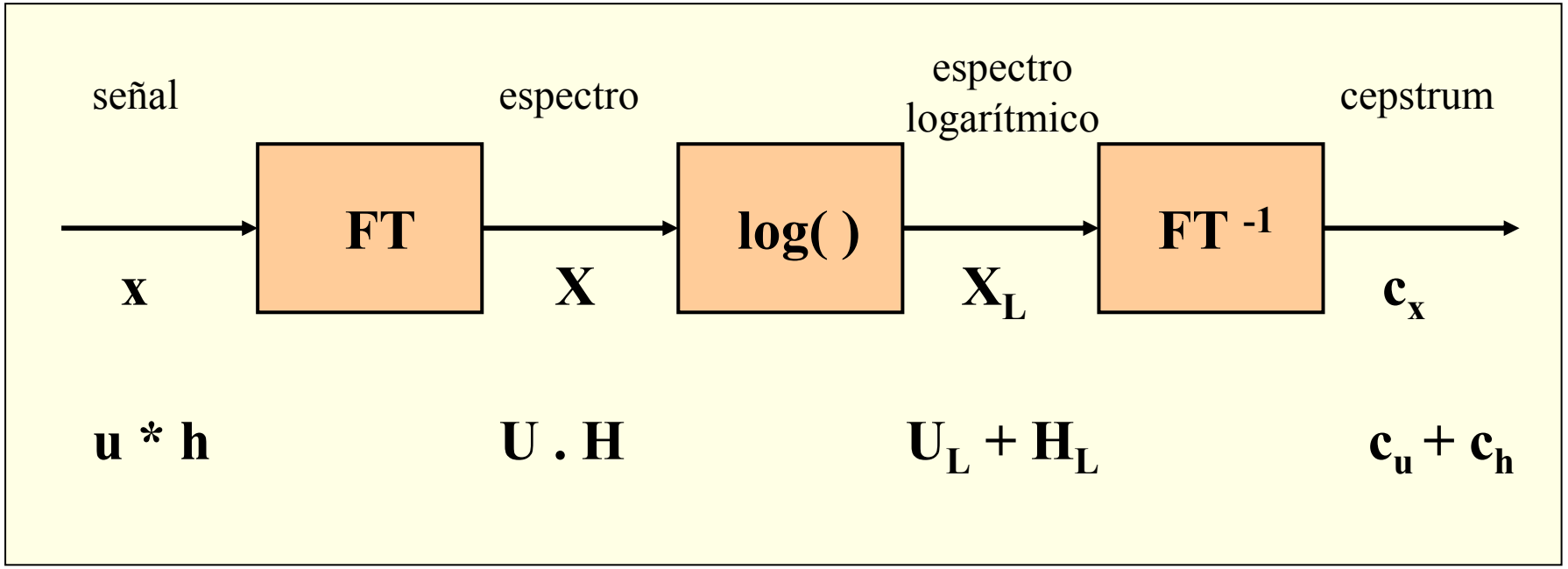




- Banco de filtros triangulares en escala Mel (para aplicar sobre el espectro FFT)
- Cada trama de voz quedaría representada por M energías de salida (una por cada filtro del banco)
- Típicamente se usan alrededor de 20 – 30 filtros solapados
- Se suele representar la potencia de salida en escala logarítmica

4.11.- Procesamiento homomórfico. Cepstrum (FFT, LPC y MFCC)

- El procesamiento homomórfico aplica operaciones no lineales
  - Objetivo: separar señales que se mezclan de forma complicada
  - Procedimiento: transformar señales para que en el dominio transformado la mezcla complicada se transforme en una mezcla aditiva
  - Dificultad: los sistemas no lineales son difíciles de estudiar
  - Para determinadas aplicaciones el procesamiento homomórfico es muy útil
- Cepstrum: procesamiento homomórfico que incluye:
  - Transformación al dominio de la frecuencia
  - Logaritmo
  - Transformación al dominio del tiempo
- El cepstrum convierte la convolución en una suma
  - El filtrado es la convolución de la excitación y la respuesta impulsiva del filtro
  - En el dominio cepstral se pueden separar las partes correspondiente a excitación y a filtrado



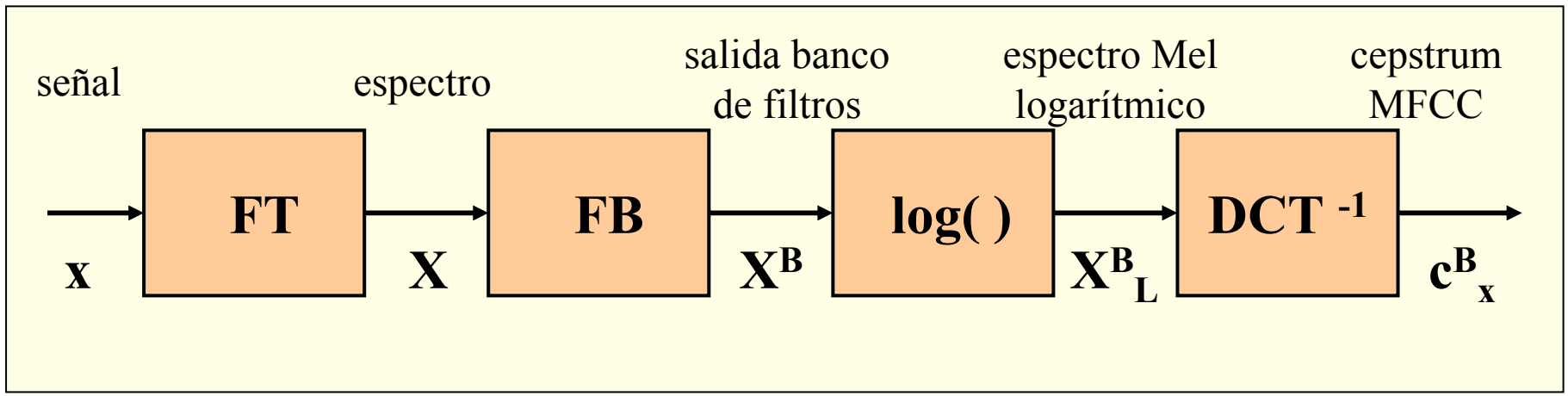
- **En el dominio cepstral es más fácil separar la excitación y el filtrado:**
  - Se mezclan de forma aditiva
  - La excitación (corresponde al rizado espectral) está en los términos de orden alto del cepstrum
  - El filtrado (corresponde a la envolvente espectral) está en los términos de orden bajo del cepstrum
  - Liftering: filtrado en el dominio del cepstrum
  - Se podría hacer transformación inversa, para recuperar *u* ó *h*

- **Cepstrum FFT:**
  - Cada trama representada por unos pocos coeficientes cepstrales (envolvente espectral)
  - El espectro FFT es un conjunto de números complejos
  - Logaritmo del espectro debe ser un logaritmo complejo

- **Cepstrum LPC:**
  - El cepstrum se puede calcular a partir del espectro LPC (también complejo)
  - También se puede calcular a partir de los coeficientes LPC

$$c(n) = \begin{cases} 0 & si \quad n \leq 0 \\ -a_1 & si \quad n = 1 \\ -a_n - \sum_{k=1}^{n-1} \frac{k}{n} c(k) a_{n-k} & si \quad n > 1 \end{cases}$$

- **Cepstrum MFCC:**
  - Mel Frequency Cepstral Coefficients
  - El espectro se estima mediante banco de filtros en escala Mel
  - La transformada inversa se sustituye por una transformada discreta del coseno (DCT) inversa
  - Reduce las operaciones con números complejos
  - Cada trama representada por un vector de coeficientes cepstrales



TEMA 5
ANÁLISIS DE SEÑALES DE VOZ

Tema 5: ANALISIS DE SEÑALES DE VOZ

5.1.- Introducción.

5.2.- La forma de onda.

5.3.- Las vocales.

5.4.- Consonantes estacionarias sonoras y sordas.

5.5.- Consonantes no estacionarias.

5.6.- Coarticulación.

5.7.- Variabilidad.

5.8.- La señal de voz en presencia de ruido:

- Ruido blanco y ruido coloreado
- Ruido no estacionario
- Detección de actividad de voz