Reconocimiento Automático del Habla 2023-2024

Preprocesamiento y Parametrización de la Voz

DSIC
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

MIARFID Máster en Inteligencia Artificial Reconocimiento de Formas e Imagen Digital

MIARFID-RAH mcastro@dsic.upv.es

Objetivos del preprocesamiento y la parametrización

Con el **preprocesamiento** se pretende acondicionar la señal:

- eliminar ruido,
- realzar la señal.

Con la **parametrización** se pretende obtener una buena representación de la señal

- compacta (poco redundante),

  (La señal muestreada requiere entre 50.000 y 125.000 bits por segundo, cuando la información transmitida como fonemas apenas llega a 50 bits por segundo.)

- y que recoge (y realza nuevamente) toda la información importante.

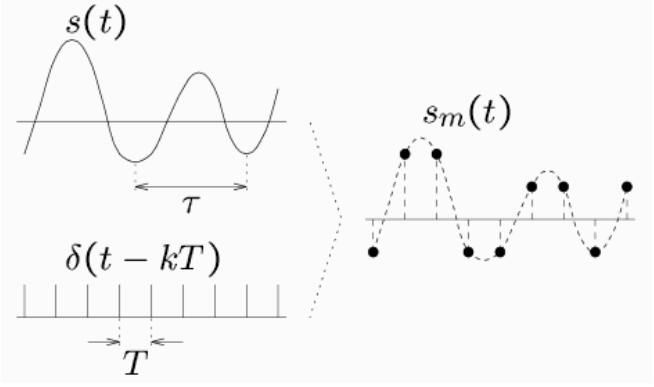
Adquisición de la señal

(1) Micrófono

Adquirimos la voz mediante un micrófono que transforma **ondas de presión** en **señal eléctrica** mediante un transductor.

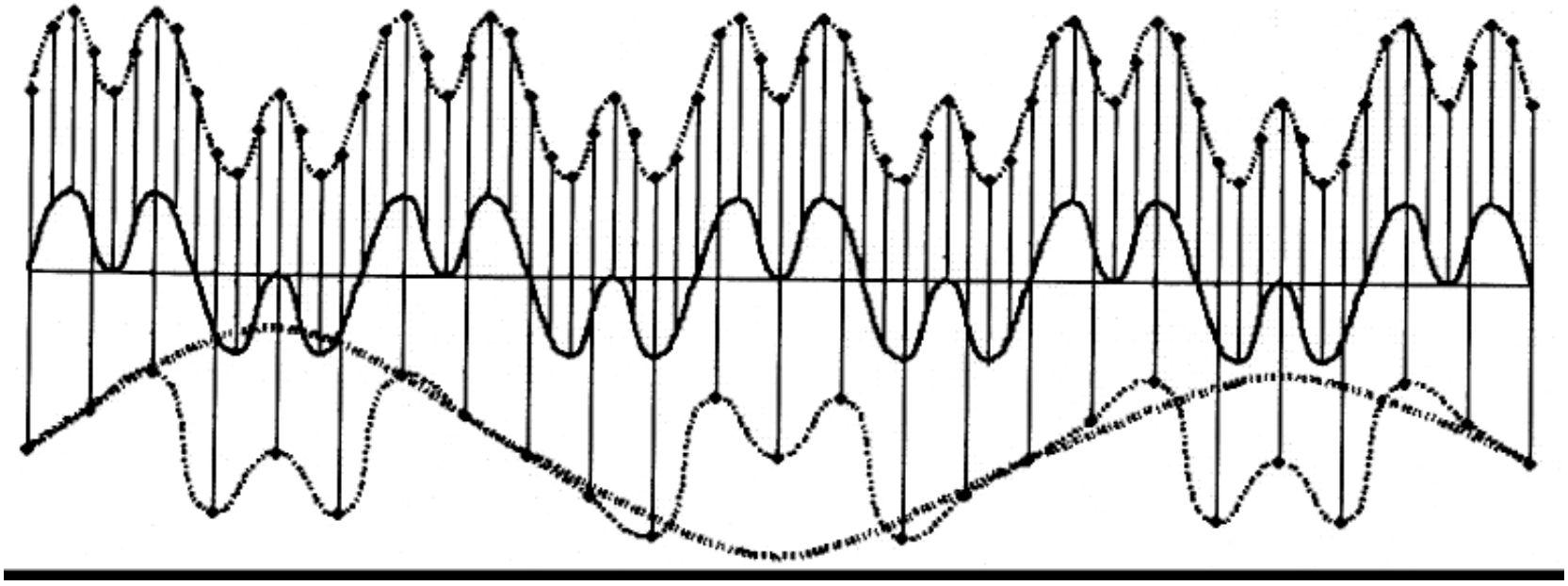
(2) Conversión A/D

La señal proveniente del micrófono es continua. Una etapa conversora A/D (analógico/digital) **discretiza** la señal a una **frecuencia dada** y con determinados niveles de c**uantificación**.



Discretización de la señal: La señal muestreada es un vector de valores: (s_0, s_1, s_2, . . .).

Frecuencia La señal acústica es muestreada a una frecuencia *Fs* (*sampling frequency*). Por el teorema de Nyquist, *Fs* ha de ser al menos el doble de la máxima frecuencia presente en la señal a muestrear o aparecen fenómenos de "aliasing".

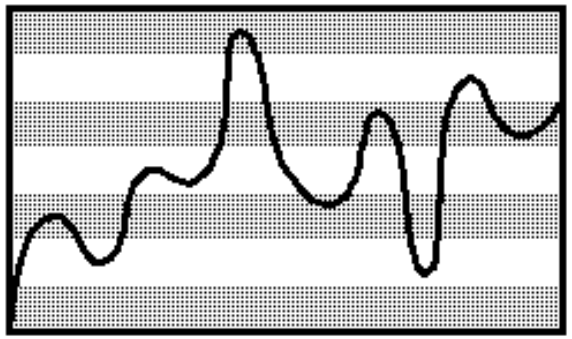


Fenómeno de "aliasing"

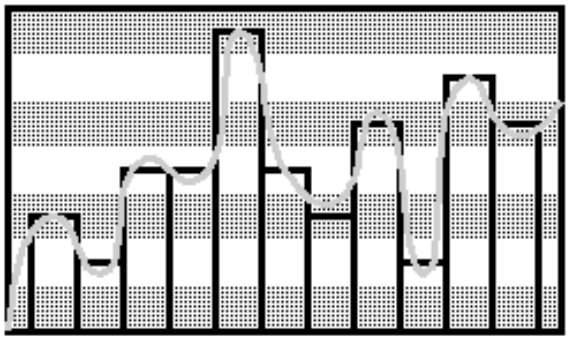
Si se escoge un valor de *T* (y por tanto una frecuencia *Fs*) que solapa los espectros repetidos, aparece el fenómeno de "aliasing".
La mayor parte de la energía en el habla se encuentra por debajo de 7 kHz.

- Típicamente *Fs* = 16 kHz en aplicaciones de reconocimiento de voz normales.
- En aplicaciones de reconocimiento de voz transportada sobre línea telefónica, la frecuencia de muestreo típica es de *Fs* = 8 kHz. (La energía está en la banda 300–3400 Hz.)

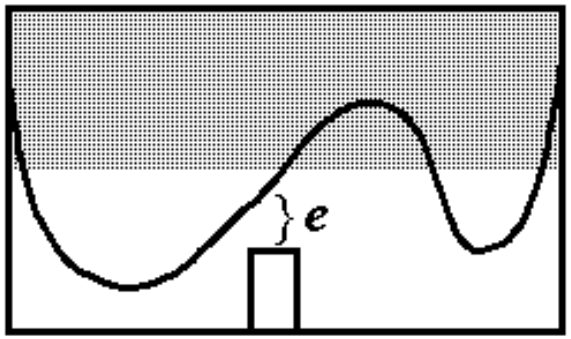
Cuantificación La cuantificación es el proceso de discretización de la señal continua. Añade cierto ruido *e* a la señal *s*: hay errores de cuantificación.



Señal continua



Cuantificación



Error de la cuantificación

El rango dinámico del oído es de aproximadamente 20 bits.

El número de bits por muestra es, típicamente, 16 (aunque 12 son suficientes para recoger el rango dinámico de la señal oral).

La señal telefónica tiene un rango dinámico de 12 bits, aunque cuantificados en 8 bits con una función de compresión no-lineal.

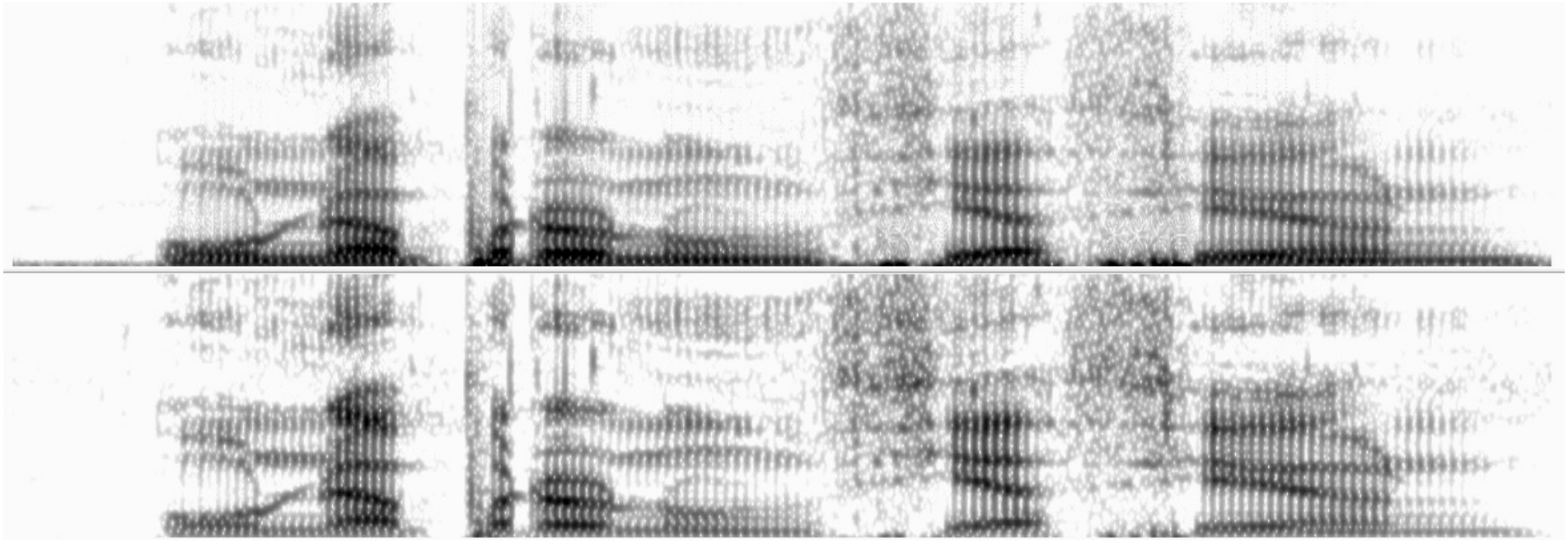
Preprocesamiento

Preénfasis

En el dominio de la frecuencia hemos observado que las características de la voz se manifiestan como **formantes** (picos debidos a resonancias del tracto vocal).

Los formantes de alta frecuencia tienen una amplitud menor que la de los formantes de baja frecuencia, aunque poseen información importante.

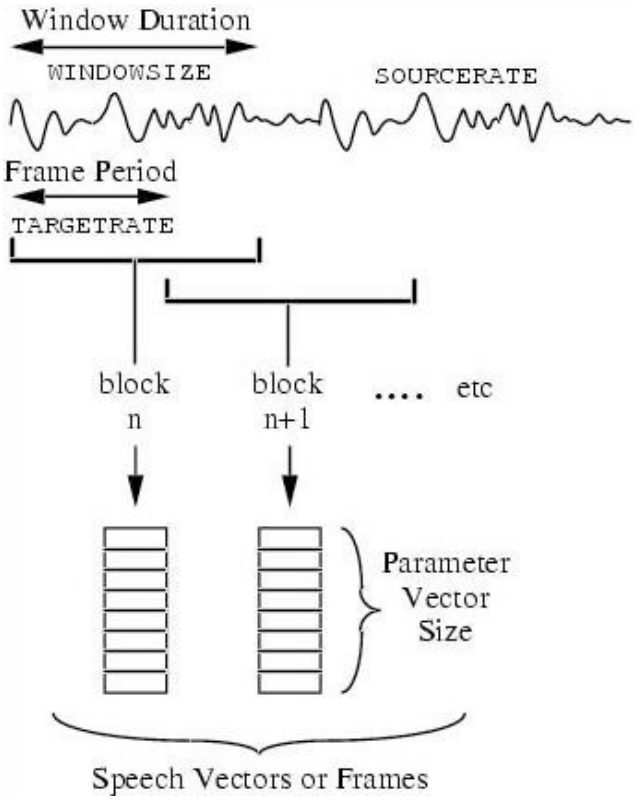
Es conveniente realzar las altas frecuencias con un proceso de **preénfasis**.



Sonograma de "una pronunciación". Arriba: sin preénfasis. Abajo: con preénfasis (factor a = 0.97)

Parametrización

(1) Análisis en bloques

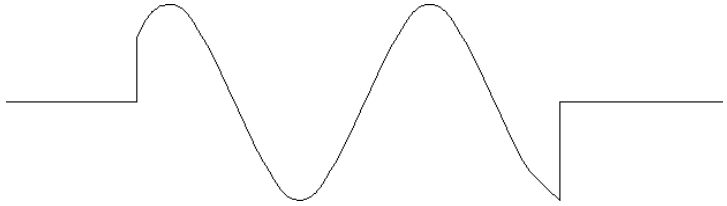


Fragmentación en bloques

Se "trocea" la señal en bloques contiguos con ciertos solapamiento. A la frecuencia con obtenemos los bloques se le denomina **frecuencia de submuestreo**. Con este proceso se convierte una señal en una **secuencia de vectores** (*frames*).

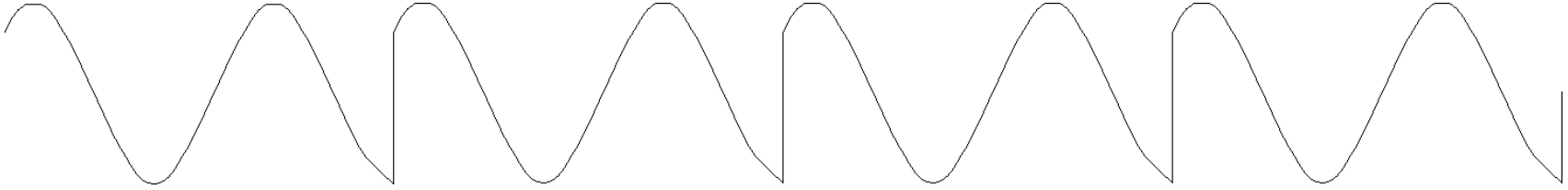
(2) Aplicación de ventanas de suavizado espectral

La forma abrupta con que empieza y acaba cada bloque se traduce en una fuerte distorsión de los espectros resultantes.



Señal recortada

Recuerda que se asume que la onda es periódica:

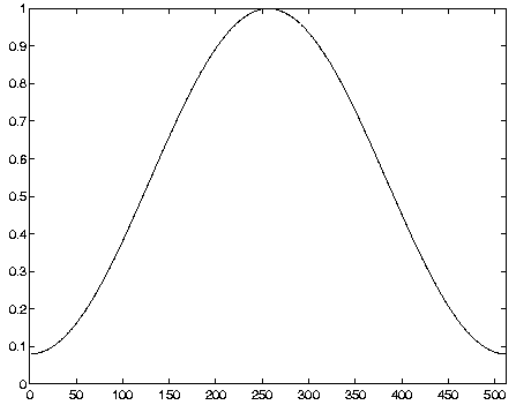


Visión de la señal recortada como función periódica

Los “saltos” verticales se modelan espectralmente introduciendo un ruido que interfiere con el “espectro real”.

El espectro se puede suavizar si aplicamos una *ventana* a cada bloque. Una ventana es una función de valor nulo fuera de cierto intervalo. Una familia de ventanas utilizada es la de *Hamming* generalizada:

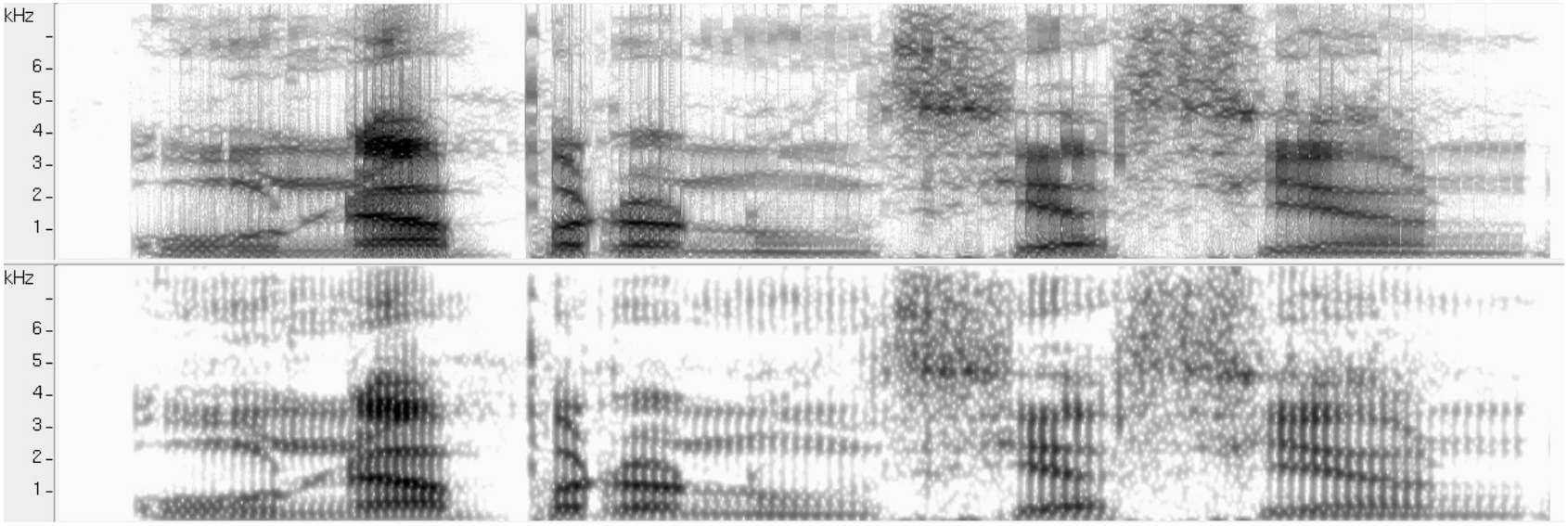
- Ventana de Hamming: si $\alpha = 0.54$. Es la utilizada normalmente.



Ventana de Hamming

- Ventana de Hanning (o de Hann): si $\alpha = 0.5$.

Ambas ventanas aplican pesos menores a los extremos del bloque.



Sononograma de “una pronunciación”. Arriba: sin ventana de suavizado. Abajo: con ventana de Hamming

La ventana se aplica entre la fragmentación en bloques y el cálculo de la FFT.

(3) Análisis en el dominio de la frecuencia: transformada discreta de Fourier

Se ha demostrado que conviene trabajar en el dominio frecuencial en lugar de en el dominio temporal.

La voz (función periódica) es una superposición de ondas senoidales.

Se puede calcular la **transformada discreta de Fourier** de cada fragmento "ventaneado" así:

$$s(t) = \frac{1}{N} \sum_{n=0}^{N-1} S(e^{i\omega n}) e^{i\omega n}.$$

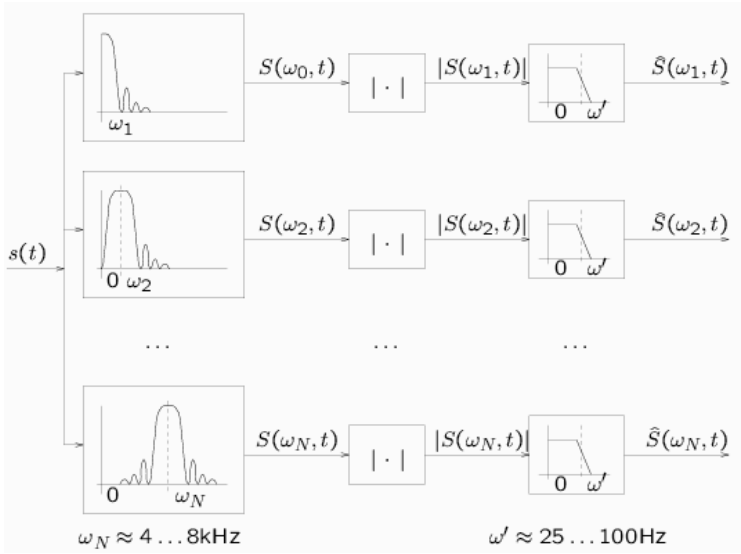
Los valores complejos

$$S(e^{0 \cdot i \frac{2\pi}{N}}), S(e^{i \frac{2\pi}{N}}), S(e^{2i \frac{2\pi}{N}}), \ldots, S(e^{(N-1)i \frac{2\pi}{N}}).$$

son los denominados coeficientes de Fourier o la transformada discreta de Fourier (DFT) de la señal *s(t)*. Los coeficientes de Fourier nos proporcionan una descripción de la señal en términos de funciones elementales (senos y cosenos).

El algoritmo FFT

El cálculo de TDF podría efectuarse mediante técnicas analógicas:



La TDF es la salida de un banco de filtros.

Dado que disponemos de una resolución de *N* frecuencias para *ω*, el cálculo trivial de un espectro en un computador es $O(N^2)$.

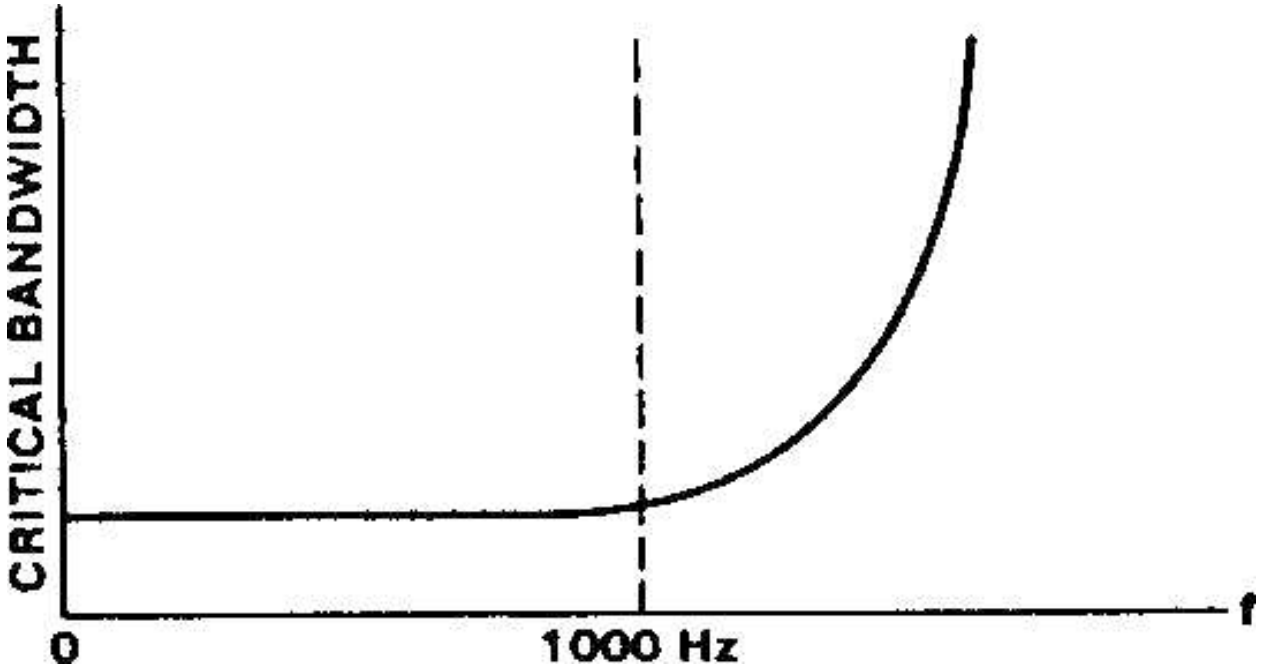
El algoritmo FFT (Fast Fourier Transform) sigue una estrategia “Divide y Vencerás” para calcular la transformada de Fourier en $O(N \log(N))$. Las versiones eficientes se basan en una transformación iterativa del algoritmo recursivo.

La FFT es susceptible de una implementación eficiente en hardware. La implementación de la FFT no es trivial. Afortunadamente hay muchas implementaciones eficientes disponibles (por ejemplo, FFTPACK).

(4) Análisis con banco de filtros

La FFT proporciona un **banco de filtros uniforme**. El oído humano no discrimina todos los rangos de frecuencias por igual. Un sistema que reproduzca la resolución no lineal de frecuencias del oído humano mejorará la calidad del reconocimiento:

Se usa un **bancos filtros en escala logarítmica** (similar a la del oído humano):



Anchos de banda en la escala perceptual

La escala es lineal a 1 KHz y logarítmica a partir de ella. La escala de Mel se asemeja a este modelo perceptual.

(5) Cepstrum

Los valores proporcionados por el banco de filtros se hallan fuertemente correlacionados. Esto da problemas cuando trabajamos con modelos estadísticos. Un modo de descorrelacionarlos es calcular la transformada inversa sobre el logaritmo de la salida del banco de filtros. Los coeficientes de esta transformación ya no presentan la correlación con el filtro. Son los denominados **coeficientes cepstrales** o **cepstrum**.

El resultado de todo el proceso se conoce por **MFCC** (Mel Frequency Cepstrum Coefficients) o **cepstrum**. La señales caracterizadas por combinaciones de armónicos se analizan mejor con el cepstrum. Por ejemplo, el cepstrum enfatiza los formantes vocálicos incluso en presencia de ruido.

El coeficiente c_0 es prácticamente la energía-log de la trama. Este coeficiente se suele calcular directamente sobre la señal. Los coeficientes cepstrales proporcionan un suavizado del espectro.

- Los coeficientes de la “parte baja” representan la macroestructura del espectro.
- Los coeficientes de la “parte alta” representan la microestructura del espectro.

Nota: hará falta cierta normalización de los coeficientes para igualar su “peso” en la descripción de la señal. La modificación de pesos en el cepstrum se denomina liftering.

Típicamente se trabaja con, a lo sumo, 15 coeficientes cepstrales (normalmente, 10 o 12).

(6) Coeficientes Delta y de Aceleración

Las prestaciones de los sistemas de reconocimiento mejoran sustancialmente si se añaden las derivadas temporales de los parámetros considerados.

La derivada se calcula tomando las diferencias de los valores en una ventana.

Típicamente K=3 para derivadas de primer orden. Si hemos obtenido Q parámetros, el resultado final es, pues, un vector con $3Q$ elementos.

Bibliografía

- Lawrence Rabiner, Biing-Hwang Juang: Fundamentals of speech recognition. Prentice Hall. 1993.
- Tony Robinson: Speech Analysis. http://svr-www.eng.cam.ac.uk/~ajr/SpeechAnalys
- Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein: Introduction to algorithms (2nd ed.). The Massachussetts Institute of Technology. 2001.
- Alan V. Oppenheim, Ronald W. Schaffer: Discrete-time Signal Processing. Prentice- hall. 1989.
- William H. Press, et al.: Numerical Recipes in C. Cambridge University Press. 1993. (Disponible en http://www.nr.com.)