Señales y Sistemas

Procesamiento de la señal de voz

H. Leonardo Rufiner y Gastón Schlotthauer

lrufiner@bioingenieria.edu.ar

Laboratorio de Cibernética – Facultad de Ingeniería
Universidad Nacional de Entre Ríos
http://www.bioingenieria.edu.ar/grupos/cibernetica/

30 de abril de 2013

Contenidos

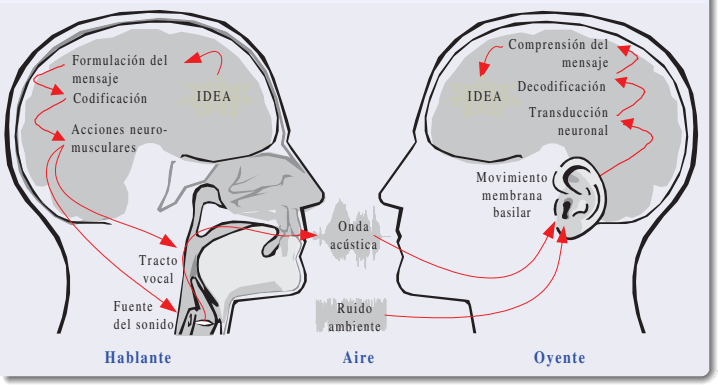
1. Introducción
2. Aparato fonador
3. Percepción de la voz
4. Análisis Cepstral
5. Estimación de la F_0

Contenidos

1. Introducción
2. Aparato fonador
3. Percepción de la voz
4. Análisis Cepstral
5. Estimación de la F_0

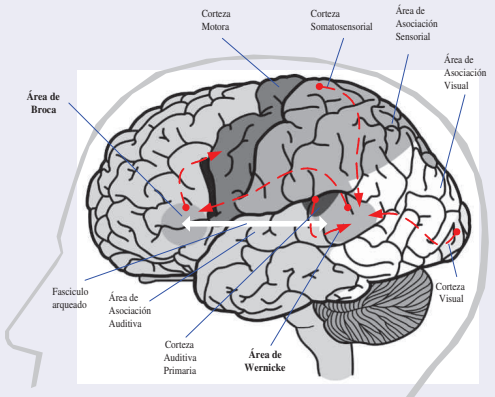
Procesamiento de la señal de voz

Introducción



Procesamiento de la señal de voz

Mecanismo de producción del habla



Producción del habla

Restricciones para la comunicación

- Acústico: Tracto vocal humano.
- Fonológico: No todos los sonidos posibles de generar son fonemas.
- Fonético: La realización particular de un fonema depende del contexto.
- Morfológico: las unidades de significación se combinan apropiadamente para formar palabras (plurales, conjugaciones).
- Léxico: Define las palabras válidas y su sentido.

Producción del habla

Restricciones para la comunicación

- Sintáctico: No todas las combinaciones de palabras tienen sentido.
- Prosódico: Relativo a la acentuación y entonación.
- Semántico: No todas las frases gramaticalmente válidas tienen sentido.
  Is the baby crying
  Is the bay bee crying
- Contextual/pragmático: Una respuesta debe cobrar sentido en la conversación.
  It is easy to recognize speech
  It is easy to wreck on nice beach

Contenidos

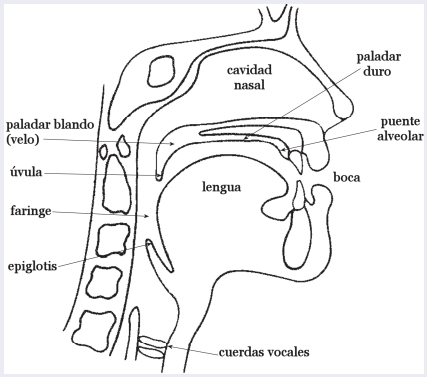
1 Introducción
2 Aparato fonador
3 Percepción de la voz
4 Análisis Cepstral
5 Estimación de la F_0

Aparato fonador

- Transforma energía muscular en energía acústica.
- Puede modelizarse como un sistema de filtros que modifica a una o más fuentes de sonidos.
- Si $H(\omega)$ es la función de transferencia del filtro que representa el tracto vocal y $X(\omega)$ la fuente de excitación (perturbación acústica de la corriente de aire pulmonar), el sonido resultante puede representarse como $Y(\omega) = H(\omega)X(\omega)$.
- Puede agregarse el efecto de radiación producido en los labios (pasa-altos).

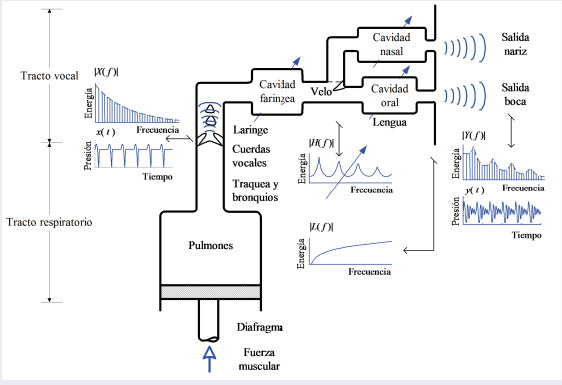
Aparato fonador

Corte esquemático del aparato fonador



Aparato fonador

Diagrama esquemático del aparato fonador



Aparato fonador

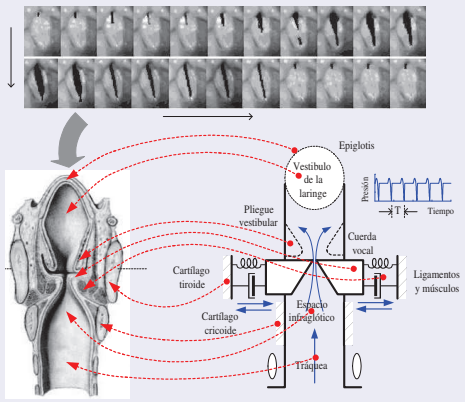
Mecanismos de excitación del tracto vocal

El sistema respiratorio es la principal fuente de energía para producir sonidos en el aparato fonador, en forma de flujo de aire y presiones. Existen tres mecanismos de excitación del tracto vocal:

1. Generación de pulsos cuasiperiódicos: las cuerdas vocales modulan el flujo de aire.
2. Generación de ruido de banda ancha: el flujo procedente de los pulmones pasa por una constricción del tracto vocal.
3. Excitación de tipo plosivo: el flujo produce una presión en un punto de oclusión total del tracto vocal y se libera rápidamente.

Aparato fonador

Laringe



Aparato fonador

Función fonatoria de la laringe

- Intervienen las cuerdas vocales, los cartílagos en los que se insertan y los músculos laríngeos intrínsecos.
- La forma de onda de los pulsos generados puede representarse en forma simplificada como una onda triangular. Un modelo más adecuado es el de Rosemberg, donde estos pulsos tienen la siguiente forma:

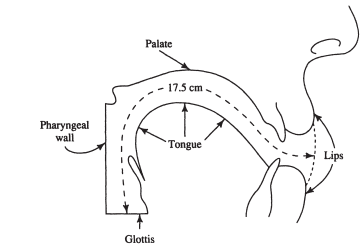
$$u_g[n] = \begin{cases} \frac{1}{2}\left[1 - \cos\left(\pi n / N_1\right)\right] & 0 \leq n \leq N_1 \\ \cos\left[\pi (n - N_1)/2N_2\right] & N_1 \leq n \leq N_1 + N_2 \\ 0 \quad \text{en otro caso.} \end{cases}$$

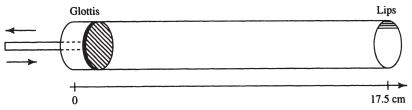
Aparato fonador

Función fonatoria de la laringe

- En el hombre, la frecuencia de esta onda varía entre 100 y 170 Hz, en las mujeres entre 180 y 280 Hz y en los niños puede superar los 300 Hz.
- Los valores de esta frecuencia glótica se modifican en forma voluntaria y son los responsables de la frecuencia fundamental (denominada F_0) que se percibe como "tono".

Aparato fonador





Función transferencia $H(\omega)$

$$H(\omega) = \frac{U_l(\omega)}{U_g(\omega)} = \frac{1}{\cos(\frac{\omega L}{c})}$$

donde $U(\omega)$ es la transformada de Fourier de $u(t)$:
$U(\omega) = \mathfrak{F}\{u(t)\}$
$u_g(t)$: flujo en la glotis.
$u_l(t)$: flujo en los labios.
L: longitud del tracto vocal ($\approx 17{,}5$ cm).
c : velocidad del sonido en aire húmedo a 37 ^oC (≈ 350 m/s).

Sonidos y fonemas

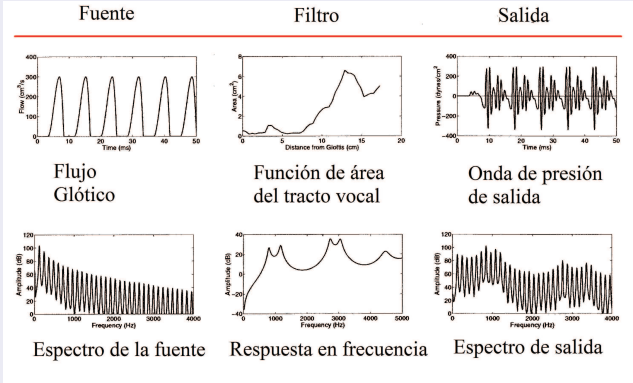
Fonema

Es el conjunto mínimo de unidades que permite decir cualquier palabra en un idioma dado. Dos fonemas son diferentes si al cambiar uno por otro, cambia la palabra. Ejemplo *boda* vs. *moda*.

Vocales:	/a/ /e/ /i/ /o/ /u/	
Fricativos:	/f/ /s/ /j/ /y/	Consonantes
Africados:	/ch/	
Oclusivos:	/b/ /d/ /g/ /p/ /t/ /k/	
Nasales:	/n/ /m/ /ñ/	
Vibrantes:	/r/ /rr/	
Laterales:	/l/ /ll/	

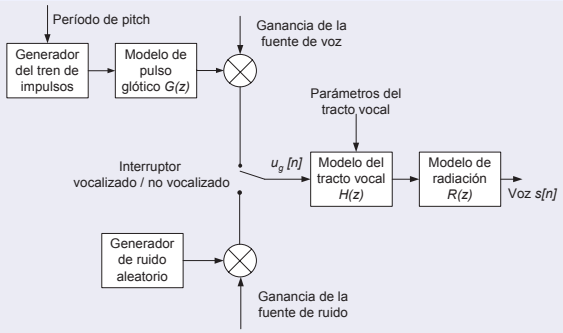
Producción de la señal de voz

Laringe



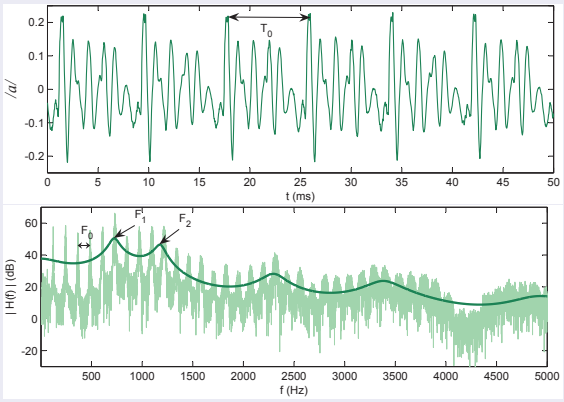
Producción de la señal de voz

Modelo Fuente - Filtro



Análisis de la señal de voz

Vocal sostenida - Período y Frecuencia fundamental (F_0)- Formantes



Análisis de la señal de voz

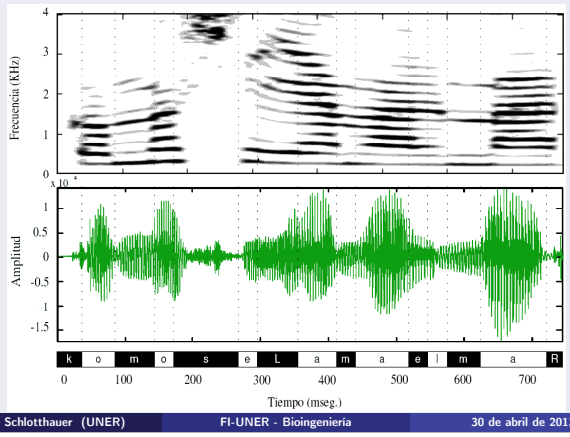
Vocal sostenida - Período y Frecuencia fundamental (F_0) - Formantes

La frecuencia fundamental F_0 corresponde a la frecuencia glótica, presente en los fonemas sonoros, y es una componente importante de la entonación en el habla.

Las frecuencias formantes $(F_1, F_2, F_3, \cdots)$ permiten discriminar entre las vocales. Su variación temporal posibilita también diferenciar entre los diferentes fonemas sonoros.

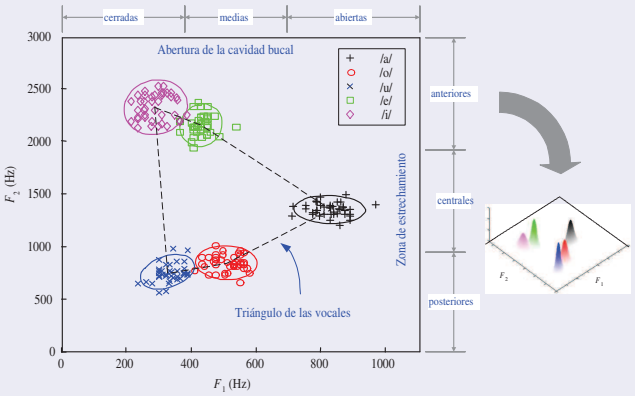
Análisis de la señal de voz

Sonograma y espectrograma



Análisis de la señal de voz

Formantes



Modelo AR del tracto vocal

Modelo

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

Análisis por tramos

Tomamos ventanas de la señal $s(n)$: $s_m(n) = s(m+n)w(n)$, $s_m(n)$ idénticamente cero fuera de $0 \leq n \leq N-1$, $w(n)$ es una ventana, como por ejemplo la ventana de Hamming. Estimamos:
$r_m(k) = \sum_{n=0}^{N-1-k} s_m(n) s_m(n+k)$,

Modelo AR del tracto vocal

Solución

Se resuelve hallando la solución de la ecuación

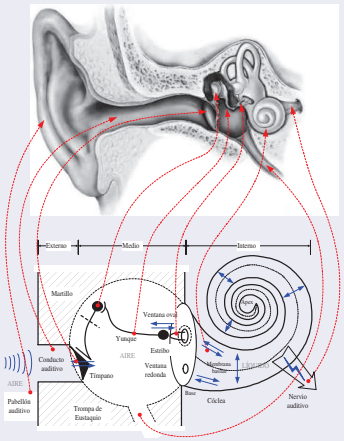
$$\begin{bmatrix} r_m(0) & r_m(1) & r_m(2) & \cdots & r_m(p-1) \\ r_m(1) & r_m(0) & r_m(1) & \cdots & r_m(p-2) \\ r_m(2) & r_m(1) & r_m(0) & \cdots & r_m(p-3) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_m(p-1) & r_m(p-2) & r_m(p-3) & \cdots & r_m(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} r_m(1) \\ r_m(2) \\ r_m(3) \\ \vdots \\ r_m(p) \end{bmatrix}$$

Un método eficiente es el de Levinson-Durbin que aprovecha el hecho de que la matriz dada es Toeplitz.

Contenidos

1 Introducción
2 Aparato fonador
3 Percepción de la voz
4 Análisis Cepstral
5 Estimación de la F_0

Audición



Frecuencia y Pitch

A menudo confundidos en la literatura, el pitch no es igual a la frecuencia fundamental. La frecuencia, intensidad y las propiedades espectrales de un sonido interactúan en formas muy complejas para dar una percepción de pitch que puede ser un reflejo muy pobre de la F_0. El pitch percibido cambia con la intensidad.
El pitch se refiere a un atributo perceptual del sonido, generalmente en una escala continua. La frecuencia es un atributo físico de las señales. La *American Standards Association* define al pitch como "aquel atributo de la sensación auditiva por la cual los sonidos pueden ordenarse según una escala musical".

Frecuencia y Pitch

Mel

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

$$1000 \text{ Hz} = 1000 \text{ mel}$$

.

Frecuencia y Pitch

Fant (1973) (la más utilizada)

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

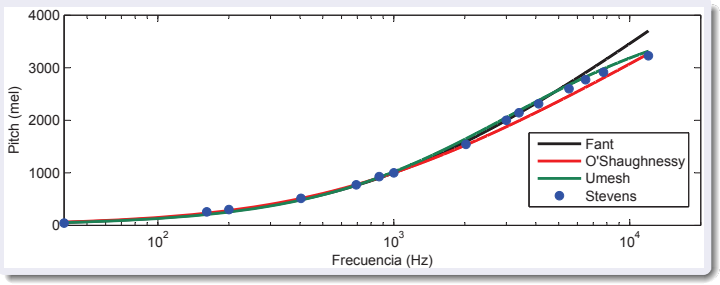
O'Shaugnessy (1987)

$$F_{mel} = 2595 \log\left(1 + \frac{F_{Hz}}{700}\right)$$

Umesh (1999)

$$F_{mel} = \frac{F_{Hz}}{0{,}00024 F_{Hz} + 0{,}741}$$

Frecuencia y Pitch



Contenidos

1 Introducción
2 Aparato fonador
3 Percepción de la voz
4 Análisis Cepstral
5 Estimación de la F_0

Procesamiento Homomórfico

Sistemas homomórficos

Los sistemas homomórficos obedecen a un "principio de superposición generalizado".
Decimos que un sistema $y[n] = h[n] * x[n]$ es homomórfico para la convolución si obedece a un principio de superposición donde la adición se reemplaza por la convolución:

$$\begin{aligned} H\{x[n]\} &= H\{x_1[n] * x_2[n]\} \\ &= H\{x_1[n]\} * H\{x_2[n]\} \\ &= y_1[n] * y_2[n]. \end{aligned}$$

Cepstrum

Cepstrum Real (Bogert, 1963)

Definición:

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \left| X(e^{j\omega}) \right| e^{j\omega n} d\omega$$

donde $X(e^{j\omega})$ es la transformada discreta de Fourier de la señal $x[n]$.

$$c[n] = \text{IDTFT}(\log |\text{DTFT}(x[n])|)$$

Vocabulario

Spectrum → Cepstrum	Amplitude → Gamnitude
Frequency → Quefrency	Harmonic → Rahmonic
Filter → Lifter	Period → Repiod
Phase → Saphe	

Cepstrum

Cepstrum Complejo

Definición:

$$\gamma[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \left\{ X(e^{j\omega}) \right\} e^{j\omega n} d\omega$$

donde $X(e^{j\omega})$ es la transformada discreta de Fourier de la señal $x[n]$.
Cómo calculamos $\log \left\{ X(e^{j\omega}) \right\}$?

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

.

Cepstrum

Preguntas...

- Es posible revertir el proceso y volver a obtener $x[n]$?
- Cómo haría un liftrado ?
- Qué podría hallar con el liftrado?
- Elegiría para ello el cepstrum real, el complejo o no habría diferencia?

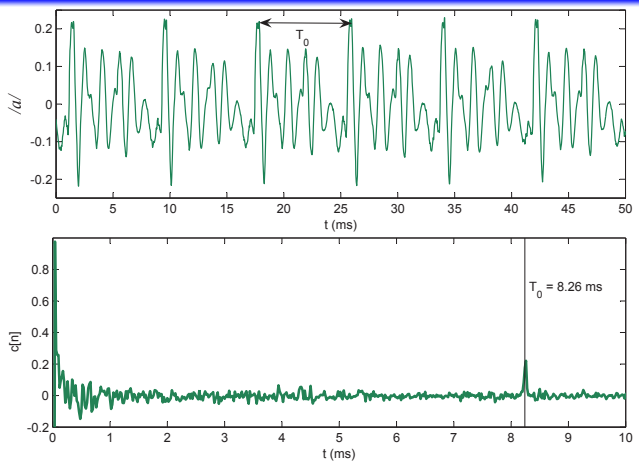
Cepstrum

Aplicación del Cepstrum real: estimación de F_0 (Noll, 1967)

Consideremos un análisis por tramos o ventanas de una señal de voz $x[n]$. Sea $G(\omega)$ la transformada de fourier de la señal de excitación del tracto vocal y $H(\omega)$ la respuesta en frecuencia de dicho tracto. Sean $c_x[n]$, $c_g[n]$, y $c_h[n]$ sus cepstra respectivamente. Entonces: $c_x[n] = c_g[n] + c_h[n]$.

$c_g[n]$ y $c_h[n]$ ocupan partes diferentes del eje de cuefrencias. Podemos separar la parte que varía rápidamente (correspondiente a la excitación del tracto vocal) de la que varía lentamente (la respuesta en frecuencia del tracto).

Cepstrum



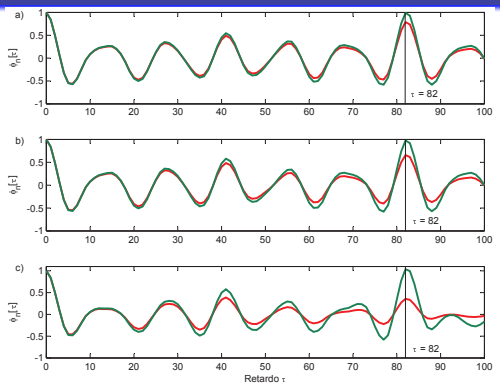
Contenidos

1. Introducción
2. Aparato fonador
3. Percepción de la voz
4. Análisis Cepstral
5. Estimación de la F_0

Método de la autocorrelación

El fundamento de la técnica de extracción de F_0 basada en la autocorrelación es que si la señal es periódica, entonces su función de autocorrelación muestra un pico en el retardo correspondiente al período, valiendo menos en todos los otros retardos (salvo en cero).

Método de la autocorrelación



En rojo se indica la función de autocorrelación sesgada y en verde la función de autocorrelación sin sesgo. Se utilizaron tres longitudes diferentes de señal: (a) 401 datos, (b) 251 datos y (c) 125 datos.

Método de la autocorrelación y clipping

El principal problema de la utilización directa de la autocorrelación es que la frecuencia de la primera formante, que está a menudo cerca o aún por debajo de la frecuencia fundamental, puede interferir con esta detección. Un problema secundario es que la señal de voz es en realidad sólo "cuasi periódica", causando que los picos de la función de autocorrelación sean menos prominentes y así dificultar la detección de los picos.
Para disminuir estos efectos se emplean técnicas de "aplanamiento del espectro", consistentes en aplicar un operador no lineal sobre la señal de voz

Método de la autocorrelación y clipping

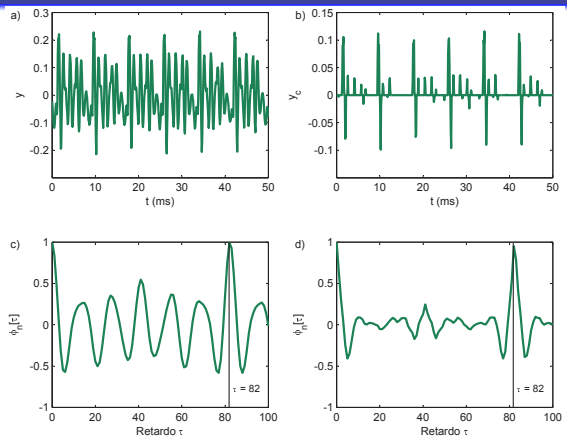
Center Clipping

La técnica más habitual es la llamada *clipping*. La función clipping más común, llamada *center clipping*, está definida de la siguiente manera:

$$C\left(x_s[n]\right) = \begin{cases} x_s[n] - C^+, & x_s[n] > C^+ \\ 0, & C^- \leq x_s[n] \leq C^+ \\ x_s[n] - C^-, & x_s[n] < C^- \end{cases}.$$

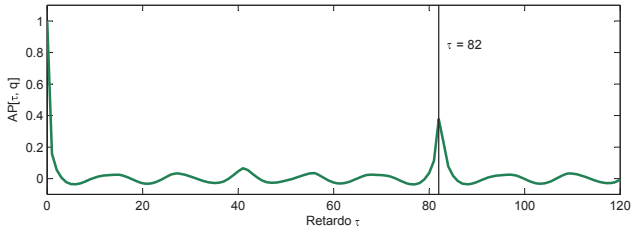
Habitualmente, los límites C^+ y C^- se fijan desde $\pm 30\,\%$ hasta $\pm 50\,\%$ del máximo del valor absoluto de la forma de onda.

Método de la autocorrelación con clipping



Método de la autocorrelación pesada (AP)

$$AP[\tau, q] = \frac{\frac{1}{N-|\tau|}\sum_{n=0}^{N-1} x_s[q+n]x_s[q+n+\tau]}{\epsilon + \sum_{n=q}^{q+N-1} |x_m[n] - x_m[n+\tau]|}.$$



Bibliografía

- J. Deller, J. Proakis, J. Hansen, "Discrete Time Processing of Speech Signals". Macmillan Publishing, NewYork, 1993.
- J. Makhoul, "Linear Prediction: A Tuturial Review," Proc. IEEE, vol 63, no. 4, pp. 561-580, apr. 1975.
- J. W. Piccone, "Signal Modeling Techniques in Speech Recognition", Proceedings of the IEEE, Vol. 81, N°9, pp. 1215-1247, 1993.
- H.L. Rufiner, "Análisis y modelado digital de la voz: Técnicas recientes y aplicaciones", Editorial UNL, 2009. (Capítulo 3).