UNA VISIÓN PRÁCTICA EN EL USO DE LA TRANSFORMADA DE FOURIER COMO HERRAMIENTA PARA EL ANÁLISIS ESPECTRAL DE LA VOZ

[1] Jesús Bernal, [2] Pedro Gómez y [1] Jesús Bobadilla

[1]
Departamento de Informática Aplicada
Universidad Politécnica de Madrid
Ctra. De Valencia Km. 7, 28031 Madrid
Tfn: +34.913367860, Fax: +34.913367527
e-mail: jbernal@eui.upm.es, jbobi@eui.upm.es

[2]
Departamento de Arquitectura y Tecnología de Sistemas Informáticos
Universidad Politécnica de Madrid
Campus de Montegancedo, s/n, Boadilla del Monte, 28660 Madrid
Tfn: +34.913367384, Fax: +34.913367412
e-mail: pedro@pino.datsi.fi.upm.es

RESUMEN

Prácticamente todas las herramientas de análisis espectral utilizadas en la actualidad se realizan a través de ordenador. Ello nos proporciona la flexibilidad de cambiar el valor de algunos parámetros en el cálculos de la Transformada de Fourier. Pensamos que es importante comprender el efecto que tiene cada uno para establecer valores que optimicen los resultados que buscamos.

Este artículo nos realiza un estudio de los parámetros que intervienen en el cálculo de la Transformada de Fourier, aportando ilustraciones prácticas de los efectos provocados en los espectros. Los parámetros que se estudian son: frecuencia de muestreo, tamaño de la ventana, tipo de ventana, números de ceros por ventana y forma de representación. Realiza unas recomendaciones basadas en una combinación de estudios teóricos y empíricos.

Finalmente realiza una referencia a diferentes métodos que se han utilizado para la mejorar del espectro, describiendo los tipos de filtros utilizados.

ABSTRACT

The majority of the actual spectra analysis tools use computers, then it is possible to change the Fourier Transform main parameters.

This article studies the parameters used in the Fourier Transform, including graphics showing the effects obtained varying the main variables. The studied parameters are: sampling frequency, windows size, type windows, number of zeros in the window and the representation form.

Finally, the article references different methods used in order to enhance the spectra, describing the covered filters.

1. INTRODUCCIÓN

Con el presente artículo queremos mostrar una visión práctica en la utilización de la Transformada de Fourier (TF) como una herramienta útil para el estudio de la fonética acústica.

La Transformada de Fourier [Bri88] es una herramienta de análisis muy utilizada en el campo científico (acústica, ingeniería biomédica, métodos numéricos, procesamiento de señal, radar, electromagnetismo, comunicaciones, etc.). Transforma una señal representada en el dominio del tiempo al dominio de la frecuencia pero sin alterar su contenido de información, sólo es una forma diferente de representarla. La potencia del análisis de Fourier radica en que nos permite descomponer una señal compleja en un conjunto de componentes de frecuencia única; sin embargo, no nos indica el instante en que han ocurrido. Por ello, esta descomposición es útil para señales estacionarias: las componentes de las frecuencia que forman la señal compleja no cambian a lo largo del tiempo.

Aunque se han realizado estudios fonéticos utilizando el análisis espectral mediante la distribución de WIGNER [Ran95], lo cierto es que la mayoría de los fonólogos utilizan los espectrogramas para la realización de sus investigaciones.

Para señales no estacionarias nos vemos obligados a tomar tramos o ventanas en donde se pueda considerar estacionaria y así poder aplicar la Transformada de Fourier. Para realizar el análisis completo debemos tomar una secuencia de ventanas para observar la evolución de la frecuencias de la señal original. Nos podemos plantear una pregunta fundamental: ¿Cuál es tamaño ideal de una ventana?

La TF debe aplicarse de $-\infty$ a $+\infty$; para tomar tramos debemos multiplicar la señal por una ventana temporal que nos aísle la parte requerida. Este hecho nos provoca una distorsión en el espectro obtenido, ya que el resultado es la convolución de la transformada de la señal con la transformada de la ventana. Nos podemos plantear una segunda pregunta: ¿Cuál es el mejor tipo de ventana?

Todos los cálculos se realizarán mediante ordenador; para ello debemos trabajar con modelos discretos y finitos. Nos podemos plantear una tercera pregunta: ¿Cuál es el número idóneo de valores para realizar la TF discreta?

Descriptores: Transformada de Fourier, Espectro, Espectrograma, Fonética Acústica.

2. LA TRANSFORMADA DE FOURIER

Se parte de la base de que toda señal genérica, por compleja que sea se puede descomponer en una suma de funciones periódicas simples de distinta frecuencia. En definitiva, la Transformada de Fourier visualiza los coeficientes de las funciones sinusoidales que forman la señal original.

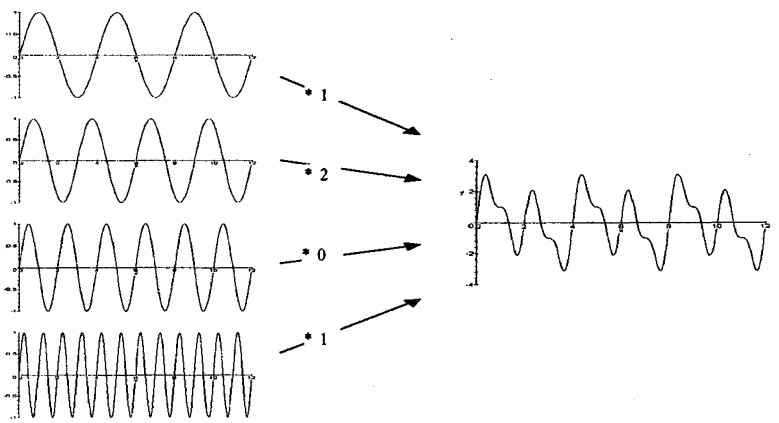


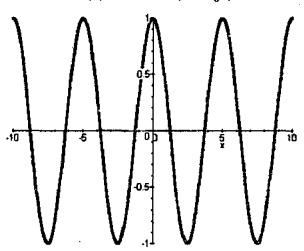
Figura 1. Una señal genérica se forma por un sumatorio de señales sinusoidales.

Si aplicáramos la transformada a la señal genérica de la figura anterior nos daría como resultado una proporción de los coeficientes que hemos utilizado para generarla.

La Transformada de Fourier se define como: $X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft}dt$. En general, $X(f)$ es una función compleja: $X(f) = R(f) + jI(f) = |X(f)|e^{j\varphi(f)}$. A partir de la señal en el dominio de la frecuencia se puede recuperar la señal en el dominio en el tiempo aplicando la Transformada inversa de Fourier: $X(f) = \int_{-\infty}^{\infty} x(t)e^{j2\pi ft}dt$.

A continuación vamos a presentar la Transformada de Fourier de cuatro funciones básicas:

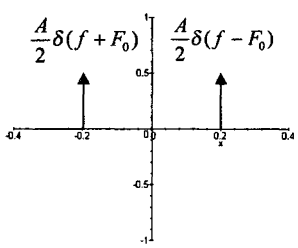
$x(t) = A\cos(2\pi F_0 t)$



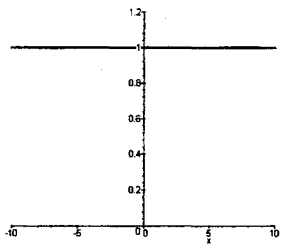
TF

$\Leftrightarrow$

$X(f) = R(f)$



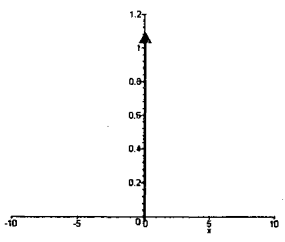
$x(t) = k$



TF

$\Leftrightarrow$

$X(f) = k\delta(f)$



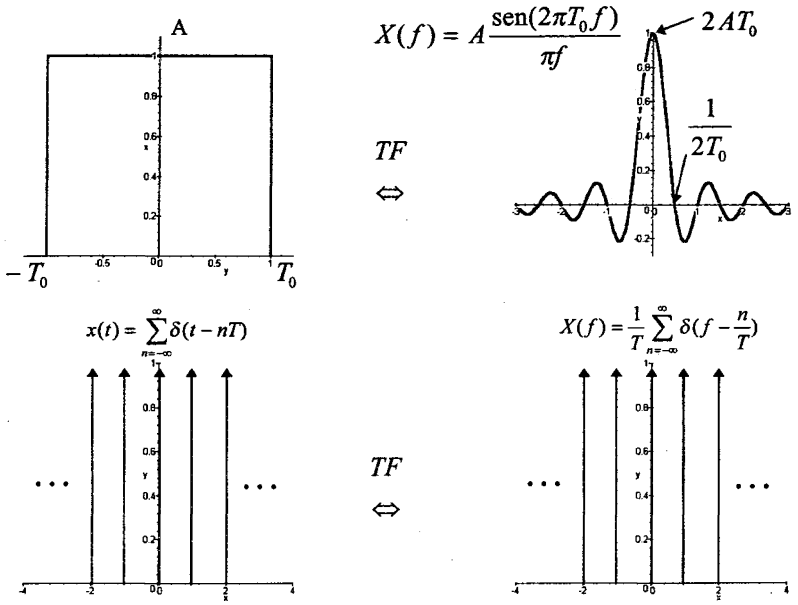


Figura 2. Transformada de Fourier de cuatro funciones básicas.

Para la función coseno, la Transformada sólo tiene parte real y para la función seno parte imaginaria. Todas las funciones representadas sólo tiene parte real.

Propiedades de la Transformada de Fourier:

- Linealidad. Si las funciones $x(t)$ y $y(t)$ tienen como transformada a $X(f)$ y $Y(f)$, respectivamente, entonces la suma de ambas tiene como transformada a $X(f)+Y(f)$.
- Simetría. Si $X(f)$ es la transformada de $x(t)$, la transformada de $X(t)$ es $x(-f)$.
- Escalado en el tiempo y en la frecuencia. Si realizamos un escalado de la variable *t* mediante la constante *k* la transformada da: $\frac{1}{|k|}x(\frac{t}{k}) \Leftrightarrow X(kf)$.

Igualmente, si realizamos un escalado de la frecuencia mediante la constante k la transformada inversa nos da: $x(kt) \Leftrightarrow \frac{1}{k} X\left(\frac{f}{k}\right)$.

- Desplazamiento en tiempo y en frecuencia. Si desplazamos el tiempo según la constante t0, la transformada nos queda: $X(t-t_0) \Leftrightarrow x(f)e^{-j2\pi ft_0}$.
  Si desplazamos la frecuencia con la constante f_0, la transformada inversa nos queda: $x(t)e^{-j2\pi tf_0} \Leftrightarrow X(f-f_0)$.
  Multiplicación frente a convolución. Definiendo la operación de convolución entre dos funciones como: $x(t)*y(t) = \int_{-\infty}^{\infty} x(\tau)y(t-\tau)d\tau$, se cumple que si multiplicamos dos funciones en el tiempo y calculamos la transformada, equivale a realizar la operación de convolución (*: operador de convolución) entre las transformadas de ambas funciones; la operación de convolución en el tiempo equivale a la multiplicación en frecuencias.
  $x(t)y(t) \Leftrightarrow X(f)*Y(f)$
  $x(t)*y(t) \Leftrightarrow X(f)Y(f)$

3. MUESTREO DE ONDAS

Como todas las operaciones se realizarán por ordenador, no podemos trabajar con funciones continuas; por ello, lo primero que debemos realizar es un muestreo de la señal de voz.

En definitiva, para muestrear la señal $x(t)$ debemos multiplicarla por un tren de deltas: $x(t)\delta(t-T) = \sum_{-\infty}^{\infty} x(nT)\delta(t-nT) = x[nT]$, siendo el período de muestreo de T. $x[nT]$ representa una secuencia infinita de impulsos equidistantes, cada uno de los cuales tiene una amplitud que corresponde con el valor de $x(t)$ en el tiempo correspondiente al impulso.

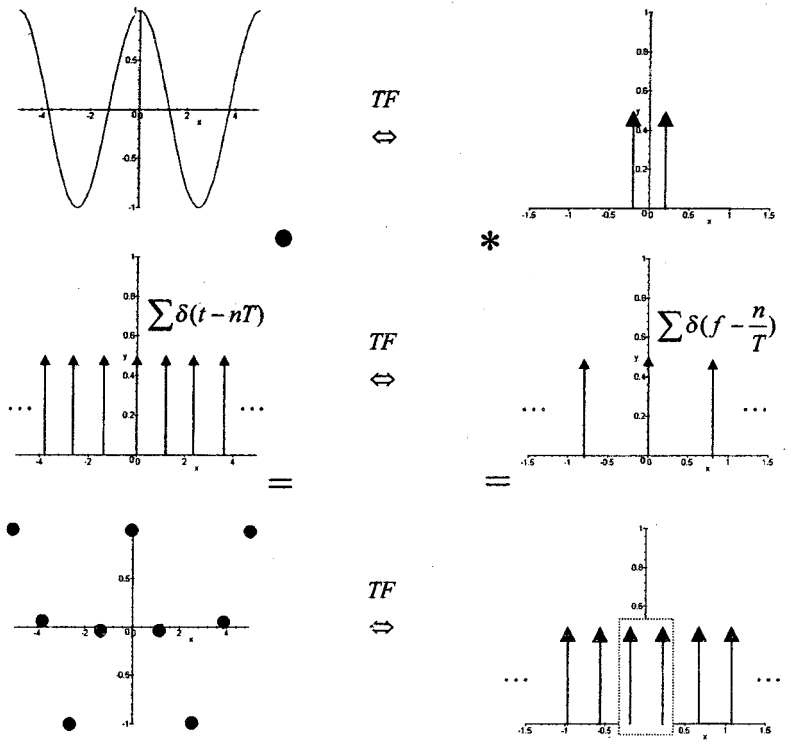


Figura 3. Proceso de muestreo de una señal.

Como ya mencionamos, muestrear una señal es equivalente matemáticamente a multiplicar por un tren de deltas; además, una multiplicación en el dominio del tiempo equivale a una convolución en el dominio de la frecuencia. La transformada de un tren de deltas cuya distancia entre deltas es T es otro tren de deltas cuya separación es $\frac{1}{T}$. Por otra parte, la convolución de una función cualquiera por un tren de deltas da como resultado una copia de la función en la posición de cada delta. Nos queda que la transformada de una señal muestreada es un tren de funciones, cada una de las cuales representa la transformada de la señal sin muestrear, separadas por sus centros en la frecuencia de muestreo.

Para conservar la transformada de la función original se debe muestrear a una frecuencia en la que no se produzcan solapamientos, siendo ésta a más del doble del ancho de banda de la señal o frecuencia de corte; se conoce a este requisito como criterio de NYQUIST o teorema del muestreo de SHANNON. Observese en la figura que si no se sigue dicha teoría se juntaría el espectro de la señal original con su copia, produciéndose una distorsión de la misma.

Para estudiar el espectro de una función muestreada sólo tendremos en cuenta el rango de frecuencias que sea menor a la frecuencia máxima o de corte.

4. TF DISCRETA. VISIÓN PRÁCTICA

La Transformada de Fourier es una herramienta muy útil cuando se trabaja con modelos matemáticos, pero si queremos trabajar con señales reales físicas y operando mediante ordenador debemos trabajar con modelos finitos y discretos.

Lo primero que debemos hacer es muestrear la señal de voz a analizar. Es importante saber la frecuencia de muestreo que debemos aplicar y para ello debemos saber el ancho de banda de la onda original. Teniendo en cuenta lo descrito en el apartado 3, un muestreo de 11.025 Hz puede ser suficiente para representar la voz (se capturaría hasta frecuencias de 5.512 Hz). Se toma esta frecuencia por ser estándar en ficheros con formatos WAV.

Una vez muestreada debemos convertirla en finita. Para ello limitaremos el número de puntos que se toman. Matemáticamente es multiplicar la señal por una ventana temporal; el efecto que provocamos es convolucionar el espectro de la señal muestreada con el espectro de la ventana. Por ello conviene elegir un tipo de ventana que produzca una menor distorsión. Aunque en el modelo hayamos aplicado una ventana, seguimos teniendo infinitas muestras que valen cero, obteniendo un espectro continuo; como última decisión debemos limitar el número de puntos (muestras mas ceros) que tomamos, lo que provocará un espectro discreto.

La Transformada de Fourier discreta se define como: $G\left(\frac{n}{NT}\right) = \sum_{k=0}^{N-1} g(kT)e^{\frac{-j2\pi nk}{N}}, n = 0..N-1$, donde $\frac{n}{NT}$ es la frecuencia de estudio, $g(kT)$ es el valor de cada muestra, T es el período de muestreo de la señal original y N es el número de puntos que se toman (incluyendo los ceros).

Parte Real, parte Imaginaria y Módulo

Al calcular el espectro de una señal, los resultados obtenidos a través de la TF son valores complejos. Aquí nos debemos plantear una cuestión: ¿Qué partes vamos a representar en el espectro?.

Si partimos de una señal original de única frecuencia, al calcular el espectro nos interesaría ver energía en la frecuencia que la compone. Al calcular la TF de una ventana obtenemos unos resultados con parte real y parte imaginaria; al cambiar de ventana obtenemos otros resultados. Según la propiedad del desplazamiento temporal, la diferencia entre ambos resultados se encuentra en que se obtiene distinta fase, pero el módulo permanece constante. Por ello, debemos a calcular exclusivamente el módulo, para que los resultados sean independientes de la posición de la ventana respecto a la señal a analizar.

Número de muestras por ventana

Al ser la voz una señal no estacionaria, debemos tomar ventanas relativamente pequeñas para que dentro de cada una se pueda considerar cuasiestacionaria. Por otra parte, si tomamos ventanas pequeñas la resolución en frecuencias resulta pobre; notar que en la TF discreta se calcula la amplitud en las frecuencias n/NT, siendo el rango $n = 0..N-1$, por lo tanto, se estudian tantas frecuencias como sea el valor de N. Por lo tanto, cuanto mayor sea el número de muestras por ventana mejor es la resolución del espectro. Aquí nos encontramos con un doble compromiso que debemos equilibrar.

En cuanto al calculo de $n = 0..N-1$ es suficiente con la mitad, ya que la segunda parte resulta ser la simétrica de la primera. Supongamos que la onda $g(t)$ tiene una frecuencia máxima de f_{max}, y muestreamos al doble de su frecuencia máxima, $T = \frac{1}{2f_{max}}$; si calculamos las frecuencias desde $n = 0..N-1$, se obtiene el espectro que buscamos y el simétrico. Por ello, debe calcularse hasta $n = \frac{N}{2} - 1$ si N es par, y hasta $n = parte\ \ entera\left(\frac{N}{2}\right)$ si n es impar. El cálculo de la siguiente mitad sería redundante. La frecuencia máxima de estudio sería $\frac{\frac{N}{2}-1}{NT}$ (para N par). El efecto de aumentar de tamaño la ventana hace que el número de frecuencias que se estudian sea mayor, a costa de disminuir la distancia entre ellas y obteniendo una mejor precisión, pero no cambia el límite de la frecuencia máxima: para un tamaño de ventana infinita sería de $\frac{1}{2T}$, que es f_{max}.

En la figura siguiente tenemos tres ejemplos demostrativos del espectro obtenido de una misma señal original de frecuencia única (2.000 Hz) y ventanas con un número diferente de muestras.

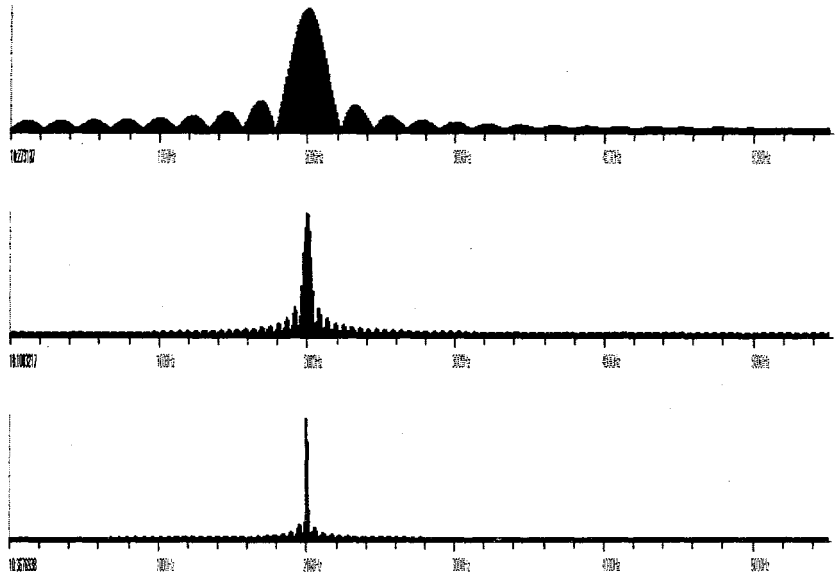


Figura 4. En la gráfica superior se han tomado 50 puntos, en la intermedia 200 y en la inferior 800. Los tres casos corresponden al módulo del espectro de una función sinusoidal pura de 2.000 Hz de frecuencia, muestreada a 11.025 Hz.

La razón matemática es que cuanto más ancha sea la ventana, más estrecha resulta la función Sinc, que es la transformada de la ventana, produciendo una menor distorsión al espectro de la función original. El límite sería un pulso de ancho infinito cuyo espectro es una delta, y así no se produciría ninguna distorsión (gráfica inferior izquierda de la Figura 2).

A veces se habla de ancho de banda al aplicar la TF. Se entiende por ello la frecuencia de muestreo divido por el tamaño de la ventana de muestras. Así, cuando hablamos de la TF en banda ancha (300 Hz) nos referimos a que si muestreamos a 11.025 Hz se toma un tamaño de ventana de 37 puntos.

Tipos de ventana

Hasta ahora hemos hablado de ventanas rectangulares. Se define como: $h(t) = \begin{cases} 1 \rightarrow |t| \leq \dfrac{T_0}{2} \\ 0 \rightarrow |t| > \dfrac{T_0}{2} \end{cases}$

Recordemos que el hecho de utilizar ventanas hace que se convolucione la transformada de la señal con la transformada de la ventana. Por lo tanto, debemos elegir aquella ventana que produzca menor distorsión.

Existen otros tipos de ventana cuyas transformadas pueden producir menos distorsión. Entre ellas tenemos la Hamming, Hanning y la Parzen. Se definen como:

Hamming: $h(t) = \begin{cases} 0'54 + 0'46\cos\left(\dfrac{2\pi t}{T_0}\right) \rightarrow |t| \leq \dfrac{T_0}{2} \\ 0 \rightarrow |t| > \dfrac{T_0}{2} \end{cases}$

Hanning: $h(t) = \begin{cases} \dfrac{1}{2}\left[1 + \cos\left(\dfrac{2\pi t}{T_0}\right)\right] \rightarrow |t| \leq \dfrac{T_0}{2} \\ 0 \rightarrow |t| > \dfrac{T_0}{2} \end{cases}$

Parzen: $h(t) = \begin{cases} 1 - 24\left(\dfrac{t}{T_0}\right)^2 + 48\left|\dfrac{t}{T_0}\right|^3 \rightarrow |t| < \dfrac{T_0}{4} \\ 2\left[1 - \dfrac{2|t|}{T_0}\right]^3 \rightarrow \dfrac{T_0}{4} < |t| < \dfrac{T_0}{2} \\ 0 \rightarrow |t| \geq \dfrac{T_0}{2} \end{cases}$

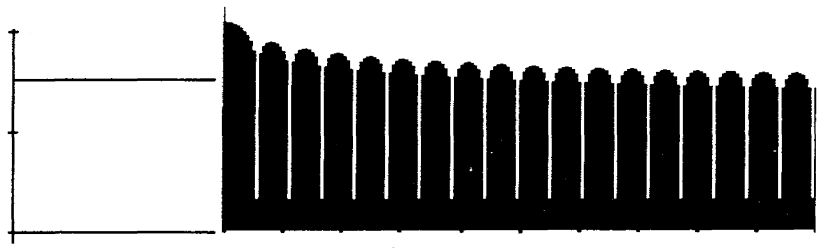


Figura 5. Ventana rectangular, TF en db.

Figura 6. Ventana Hamming, TF en db.

Figura 7. Ventana Hanning, TF en db.

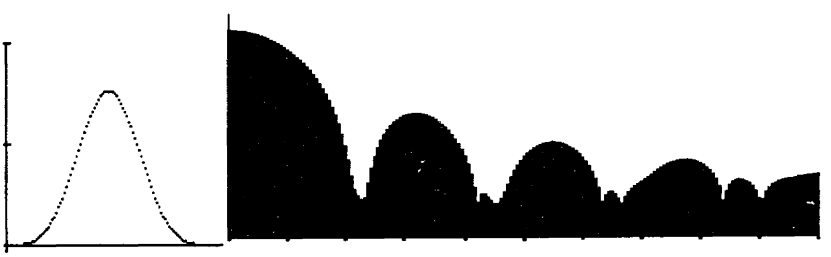


Figura 8. Ventana Parzen, tf en db.

Entre los distintos tipos de ventanas expuestas es la ventana Hamming la más utilizada.

Número de puntos de una ventana

Otra cuestión interesante es el número de ceros que se toman para el cálculo de la TF. Este aspecto es diferente al número de muestras de la ventana, detalle mencionado anteriormente. Por las características de la voz, debíamos tomar un número reducido de muestras; llamemos a este número M. Para cuestiones de cálculo, trabajaremos con la expresión $g(kT)h(kT)$, siendo $h(kT)$ la ventana utilizada. Ahora nos queda por definir la variable *N*, que es el número de puntos con que aplicamos la TF discreta. Podría tener un rango de $M..M'$ (siendo M' finito); para cualquier valor de N finito obtendremos un muestreo del espectro continuo, con una distancia entre muestras de $\frac{1}{NT}$. Normalmente se toma un $N > M$ para dar una mejor visión del espectro.

En la figura posterior tenemos un ejemplo del efecto de añadir ceros al cálculo de la Transformada de Fourier.

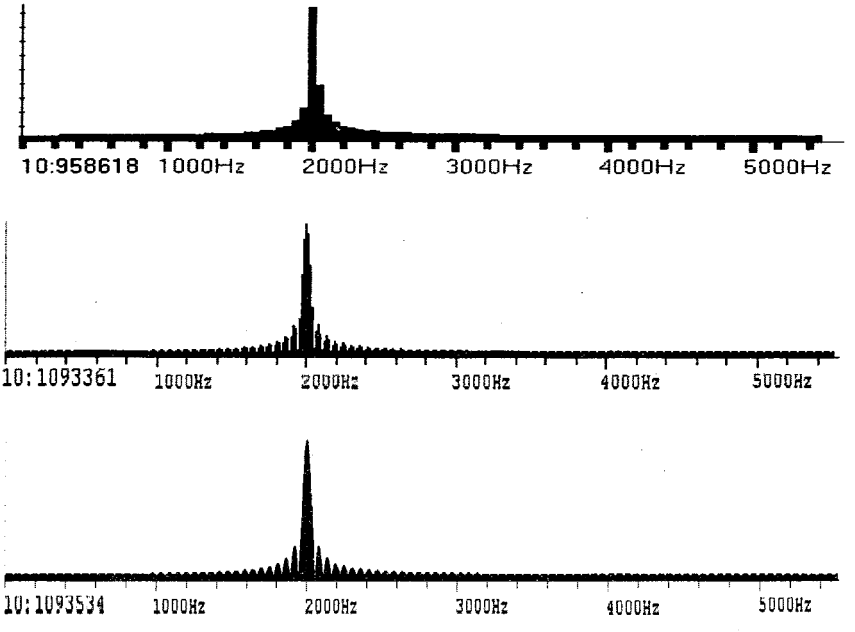


Figura 9. Corresponde a una función sinusoidal de 2.000 Hz de frecuencia, mustreada a 11.025 Hz y con un ancho de ventana rectangular de 200 muestras. En la gráfica superior N=200 puntos, en la intermedia N=800 y en la inferior N=3200.

Como observamos en la figura, aunque al añadir ceros aumentamos el rango de frecuencias visualizadas no mejoramos la resolución del espectro, aspecto que si se observa en la Figura 4.

Un caso particular es cuando tomamos un tamaño de ventana que sea un múltiplo del período de la señal original. En tal caso obtenemos un espectro ideal ya que se muestrea el espectro continuo en los puntos donde la función Sinc resultante de la ventana vale cero, con lo que la distorsión queda oculta. En la gráfica siguiente se ha tomado un número de ceros suficiente para apreciar el espectro completo.

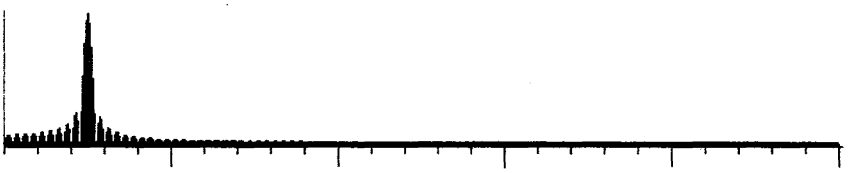


Figura 10. Esta representa un espectro bien definido de la señal original. Es una función sinusoidal de 500 Hz de frecuencia y muestreada a 11.025 Hz. Un período completo contiene 20 muestras.

En la Figura 11 se han tomado tres casos; en la gráfica superior coincide el número de muestras con el período de la señal original; en la intermedia se desfasa en un cuarto de período, y en la inferior en medio período.

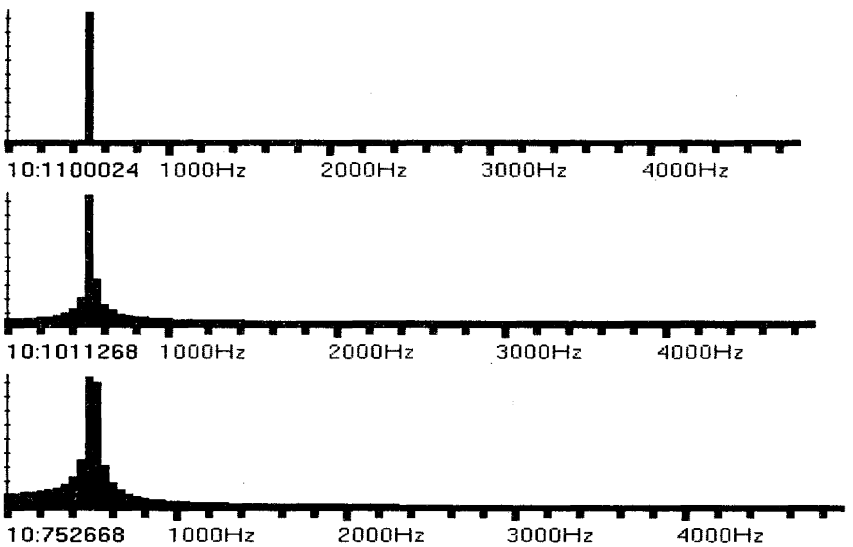


Figura 11. En la gráfica superior se ha tomado un múltiplo del período de la señal original (200 muestras), en la intermedia se ha tomado un cuarto de período más (205 muestras) y en la inferior medio período de más (210 muestras), este último sería el peor de los casos.

La gráfica superior sería la más deseable, pero para ello deberíamos conocer con precisión el periodo de la señal original.

5. ESPECTROS VARIANTES EN EL TIEMPO

Hasta ahora hemos visto el espectro de una sola ventana. Pero para observar la evolución de la señal de voz debemos visualizar una secuencia de espectros, uno por cada ventana.

Existen dos alternativas básicas:

1. Utilizar una tercera dimensión que represente el tiempo, presentando gráficos en tres dimensiones (la tercera dimensión sería la secuencia de ventanas).
2. Utilizar el color como tercera dimensión, presentando gráficos en color y en dos dimensiones.

Consideramos que la segunda solución es más fácil de interpretar y será la que tomemos. En la figura siguiente tenemos un ejemplo general de cómo se va formando el espectro variante en el tiempo.

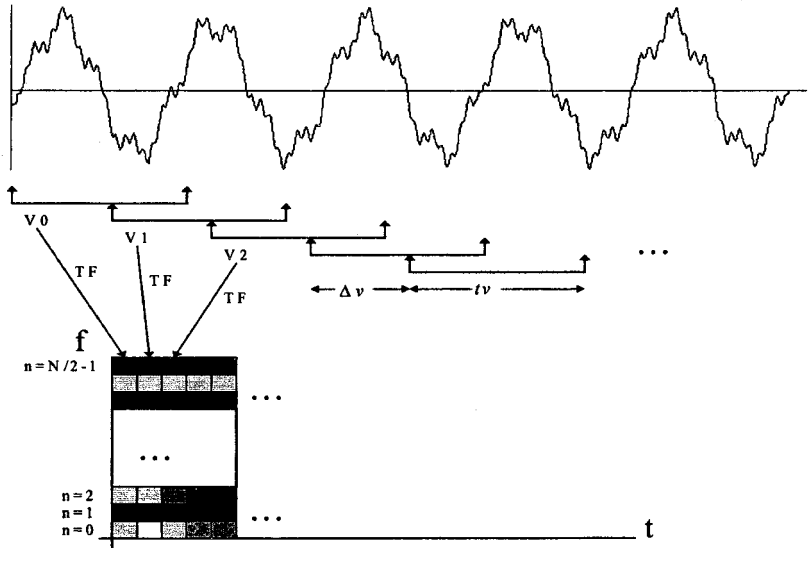


Figura 12. Esquema de cómo se forma el espectro variante en el tiempo.

Al aplicar la TF a una ventana nos da como resultado un conjunto de valores que corresponden a las amplitudes de las distintas frecuencias. Es el módulo de esta amplitud el que se codifica en colores. Como la diferencia de las amplitudes es muy diferente, normalmente se aplica una escala logarítmica para la codificación. Además, se establece un nivel mínimo de señal, debajo de la cual se considera ruido y se elimina, y nivel máximo para la codificación de colores.

Las ventanas pueden estar solapadas, contiguas o separadas. Todo dependen de la intensidad de información que queramos visualizar; pero no hay que olvidar el coste de computo que supone.

6. REALZAMIENTO DEL ESPECTROGRAMA

Aunque se han realizado estudios fonéticos utilizando el análisis espectral mediante la distribución de WIGNER [Ran95], lo cierto es que la mayoría de los fonólogos utilizan los espectrogramas para la realización de sus investigaciones.

Posteriormente al espectrógrafo han surgido equipos que mejoran la obtención del espectrograma [Mar87]. El analizador de BRÜEL & KJAER obtiene el espectro por el cálculo de la Transformada rápida de Fourier, permitiendo la programación de algunos de los parámetros (tamaño de ventana, tipo de ventana...). Existen otras herramientas y equipos, pero en definitiva se reducen a acelerar el proceso y facilitar cierto control para la realización del cómputo.

Por su parte R. BRISTON-JOHNSON [Bri95] realiza un estudio comparativo de la distorsión que produce el uso de distintos tipos de ventanas.

Además se ha estudiado un gran conjunto de representaciones espectrales (FOURIER, WIGNER, CHOI-WILLIAMS...) estableciendo las propiedades de cada una, comparándolas para establecer sus ventajas e inconvenientes [Ril89] [Jon92].

Prácticamente se repite en todas las referencias la problemática del espectrograma: en banda ancha se dispone de una resolución en frecuencia pobre, y en banda estrecha una resolución temporal pobre.

P. BASILE (capítulo 13 de [Coo93]) plantea aplicar una escala logarítmica en frecuencias; con ello se consigue una mejor representación de las frecuencias bajas y una peor en las frecuencias altas. El problema no se resuelve ya que se sigue teniendo una representación pobre temporal en las frecuencias bajas. Por otra parte pensamos que es adecuado aplicar la escala logarítmica, ya que es así como percibimos acústicamente las distintas frecuencias.

En cualquiera de los casos, el nivel de ruido que aparece es muy elevado, con una calidad de presentación de las características acústicas deficiente. Es notoria la importancia que en fonética y fonología tiene la experimentación, ya que es un método válido para incrementar la información de cuestiones lingüísticas [Sol84].

Encontramos una necesidad de investigar sobre dicho tema, buscando como objetivo general obtener un espectro más legible.

En vista de las referencias encontradas, podemos agrupar en tres las técnicas utilizadas para mejorar el espectro:

1. Combinación del espectro en banda ancha y estrecha.
2. Utilizar un filtro espacial de la imagen completa.
3. Reasignamiento de la energía.

Por combinación de espectro se entiende el trabajo de S. CHEUNG [Che91]. Consiste en utilizar banda ancha y banda estrecha de forma simultánea para obtener una buena resolución en tiempo y en frecuencias.

Se define X_c como espectro combinado, X_e como espectro de banda estrecha y X_a como espectro de banda ancha. La combinación queda:

$$|X_c(n,k)| = \sqrt{|X_e(n,k)||X_a(n,k)|}\,.$$

Se utiliza una ventana tipo Hamming.

Presenta un ejemplo con la frase: "These shoes were black and brown", pero en nuestra opinión produce una mejora escasa; Y. SHIN [Shi97] por su parte también opina que los resultados no son satisfactorios.

T. G. THOMAS [Tho94] no aporta nada nuevo a la idea de S. CHEUNG, plantea calcular un espectro combinado con una formulación similar.

A continuación vamos a describir el concepto básico de la aplicación de filtros espaciales [Gon87] [Mar93]. Sea $X(f,t)$ el espectro resultante de la señal temporal $x(t)$. Veamos a $X(f,t)$ como una imagen en dos dimensiones, donde el tiempo es un eje, la frecuencia otro, y el valor de la intensidad se codifica en intensidad del pixel de la imagen. La relación entre las variables (*n*, *v*) y (*f*, *t*) depende de la frecuencia de muestreo, tamaño de la ventana aplicado y el desplazamiento temporal de las ventanas. La variable *n* hace referencia a la posición del pixel en el eje y, representando una frecuencia determinada y la variable *v* hace referencia a la posición en el eje X, representando tiempo.

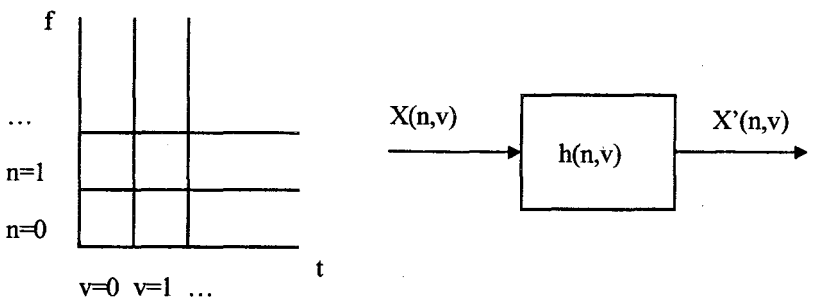


Figura 13. Imagen de dos dimensiones.

La función de filtrado se describe en la parte derecha de la Figura 13, siendo $h(n,v)$ la respuesta del filtro a la función impulso. Debemos convolucionar ambas funciones para obtener la salida:

$$X'(n,v) = X(n,v) * h(n,v) = \sum_{i,j} X(i,j)h(n-i,v-j).$$

Se aplica el filtro espacial cuando $h(n,v)$ tiene un rango reducido; por ejemplo $h(2,2)$.

El uso de filtros espaciales se describe en la referencia de Y. SHIN [Shi97]. Propone el filtro de la Figura 14. Utiliza una escala de grises de 256 niveles y una ventana Hamming.

1/9 *

-1	-1	-1
-1	20	-1
-1	-1	-1

Figura 14. Pesos del filtro.

Presenta el ejemplo con la frase "red and blue shoes". Se aprecia que la imagen mejora, pero consideramos que sigue teniendo excesivo ruido de fondo.

Más interesante nos parece la aportación de V. R. CHARI [Cha95]. Plantea un método adaptativo para mejorar el espectro de Fourier. Se basa en que el cambio de frecuencias y amplitud de los formantes se produce de forma lenta (en rangos de tiempo entre 2ms y 8ms). Describe un algoritmo que enfatiza los formantes, no los extrae. El procedimiento que describe está formado por tres etapas:

1. Aísla las zonas que corresponden a sonido de voz.
2. Determina las regiones donde existen cambios bruscos. Se pretende aplicar el algoritmo a las regiones que no contengan cambios bruscos.
3. Se aplica el algoritmo de suavizado que se describe a continuación.

Se trabaja con el espectro como un esquema bidimensional, donde el eje *x* es el tiempo y el eje *y* las frecuencias. Por cada punto se define un rectángulo $(w'(r))$ de longitud R, y se calcula

$$d_{n,k}(\theta)=\frac{1}{R}\sum_{r=-\infty}^{\infty}a(n+entero(r\cos(\theta)),k+entero(r\,\mathrm{sen}(\theta)))w'(r)$$

donde $a(n,k)=|X(nL,k)|^2$. La ventana rota 45° con respecto del eje de abscisas. Se busca el ángulo que hace mínima la expresión $|a(n,k)-d_{n,k}(\theta)|$, que corresponde a la máxima correlación. Se define el nuevo valor como $a'(n,k)=d_{n,k}(\theta_0)$, siendo θ_0 el ángulo anteriormente calculado. Este nuevo valor se almacena en otra tabla para que no afecte a cálculos posteriores. El método demuestra cierta eficacia.

El método de reasignamiento fue propuesto por K. KODERA [Kod78]. Describe cuatro métodos (MS, ME, MWM y MMWM), llegando a la conclusión de que los mejores resultados se obtienen con el método MMWM; consiste en asignar el valor del espectrograma al centro de gravedad del valor de la energía de la ventana, en lugar del centro de la ventana como el método MWM. Los nuevos puntos de reasignamiento vienen definidos por:

$$t'=t-\frac{1}{2\pi}\frac{\partial\phi(t,f)}{\partial f}$$

$$f'=\frac{1}{2\pi}\frac{\partial\phi(t,f)}{\partial t}$$

siendo $\phi(t,f)$ la fase de la Transformada de Fourier resultante.

F. PLANTE [Pla95] lo aplica sólo para el reasignamiento de frecuencias utilizando una implementación más rápida propuesta por F.AUGER [Aug94]:

$$t'(t,\omega)=t-R\left\{\frac{STFT_{th}(t,\omega)STFT^{*}{}_{h}(t,\omega)}{|STFT_{h}(t,\omega)|^2}\right\}$$

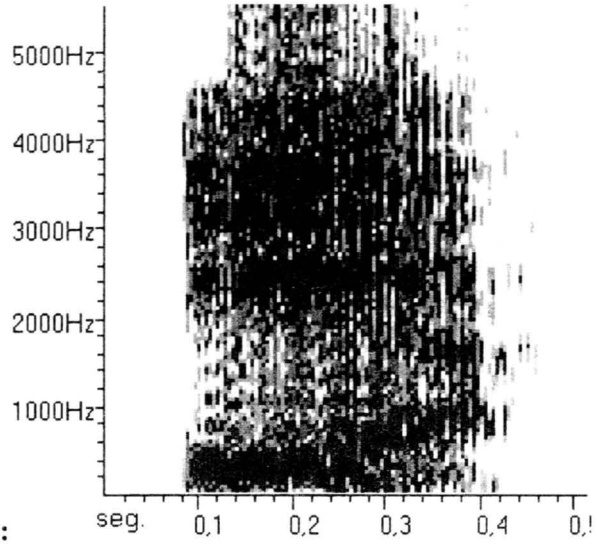
$$\omega'(t,\omega)=\omega+I\left\{\frac{STFT_{Dh}(t,\omega)STFT^{*}{}_{h}(t,\omega)}{|STFT_{h}(t,\omega)|^2}\right\},$$

siendo $\tau h = t \cdot h(t)$, y $Dh = \frac{\partial h}{\partial t}$. $STFT_h$ es la Transformada de Fourier utilizando la ventana *h*, τh representa la ventana multiplicada por *t* y *Dh* la derivada respecto a *t*.

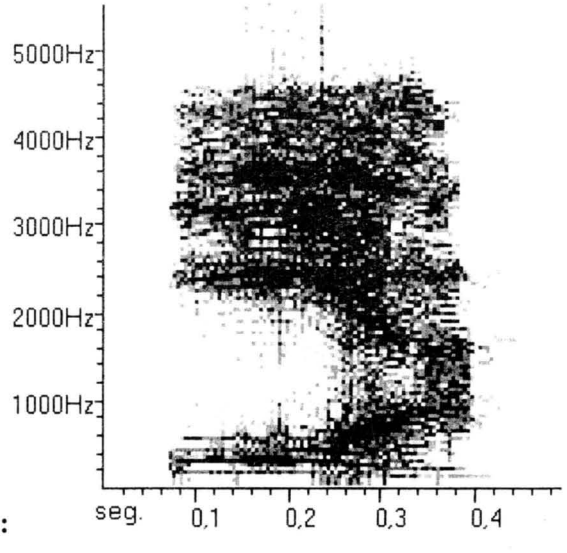
En los resultados obtenidos aparecen dos efectos: por una parte, realza los formantes y, por otra, permite discriminar entre dos formantes que estaban fusionados, aunque consideramos que el ruido que permanece es muy elevado.

La primera cuestión que nos debemos plantear es la elección del tamaño de la ventana. Hasta ahora siempre se ha hablado de banda ancha (300Hz) y banda estrecha (45Hz); algunos autores proponen una combinación de ambas. Pensamos que estos valores extremos son herencia del espectrógrafo, pero en principio podríamos tomar cualquier tamaño. Para obtener siempre el mismo número de frecuencias, rellenaremos con ceros hasta un total de 256 puntos (normalmente se toman 256 o 512 puntos).

Se ha realizado un conjunto de pruebas cambiando el tamaño de la ventana desde su valor más pequeño (300Hz) hasta su valor más grande (45Hz). En la figura siguiente tenemos tres espectros representativos. Para que los resultados se puedan comparar con mayor facilidad se ha normalizado la energía y el número total de ventanas temporales calculadas.


5000Hz
4000Hz
3000Hz
2000Hz
1000Hz
seg.
0,1
0,2
0,3
0,4

(A):


5000Hz
4000Hz
3000Hz
2000Hz
1000Hz
seg.
0,1
0,2
0,3
0,4

(B):

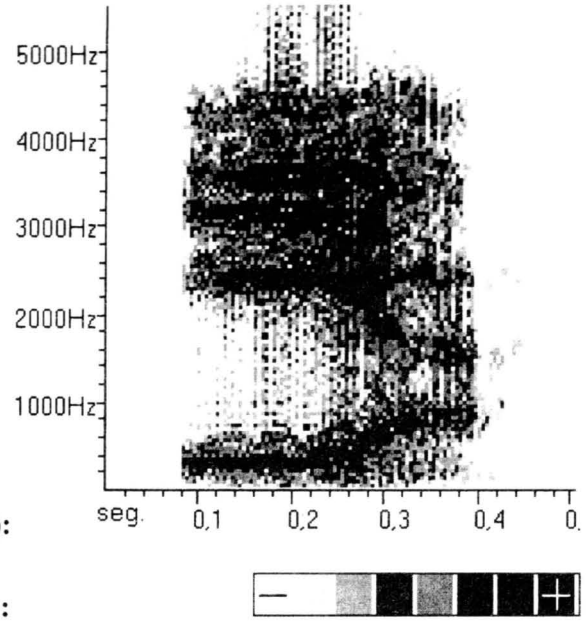


Figura 15. Comparación de espectros de distinto tamaño de ventana. Corresponde a la secuencia 'ia'. Figura A: 37 muestras, figura B: 245 muestras, figura C: 90 muestras. Figura D: paleta de colores utilizada.

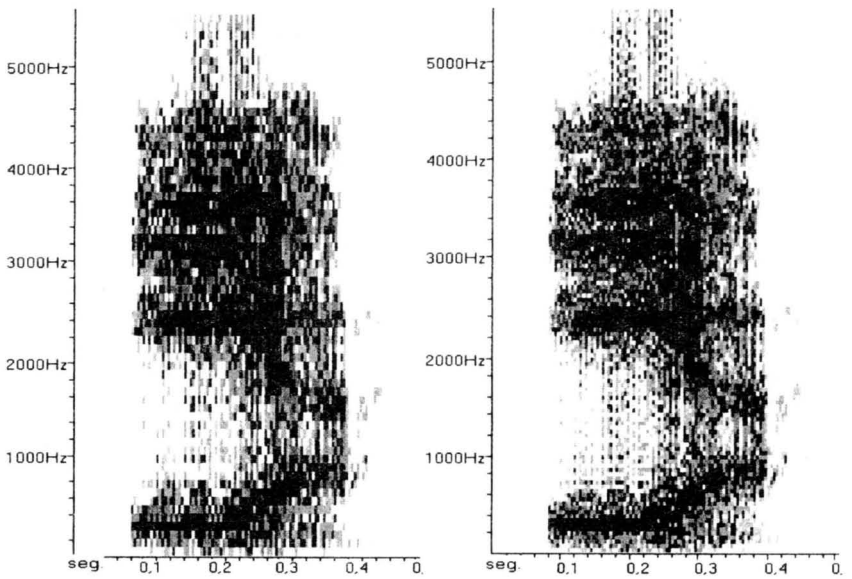
La sensación visual del espectro en banda ancha es que existe un rayado vertical, y en banda estrecha que existe un rayado horizontal, mientras que para el espectro de ancho de banda de 123Hz (90 muestras de señal) parece equilibrado.

Nuestra conclusión es que el tamaño de ventana que da unos resultados más exactos es el correspondiente al período del fundamental (*pitch*), conclusión que coincide con la opinión de R. SMITS [Smi94].

La cuestión que se plantea a continuación es si utilizamos ventanas variables que se adapten continuamente a la frecuencia fundamental. Ello generaría el problema de extraer el pitch de forma automática y provocaría inexactitudes por los errores de la extracción.

Después de analizar el estado de los detectores del fundamental se comprobó que ésta era una cuestión compleja. Se realizaron un conjunto de pruebas, comprobando que utilizando un tamaño constante similar al valor de la frecuencia del fundamental media daba unos resultados lo suficientemente buenos para no introducir una etapa de captura automática del mismo.

Respecto al número de ceros, se considera que con rellenar hasta 256 da una calidad suficiente. Se podría utilizar 512 o 1.024, pero es una cuestión de definición de pantalla del ordenador, más que una cuestión de claridad del espectro.



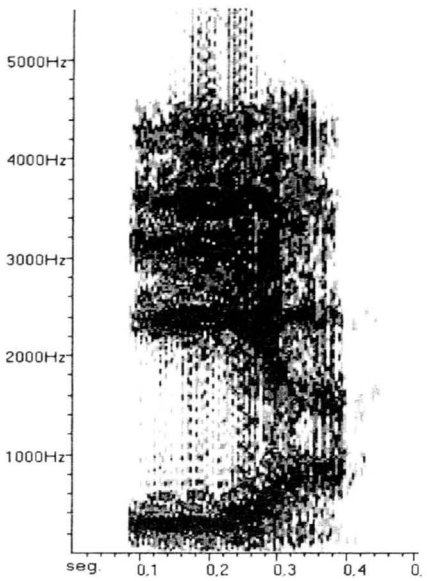


Figura 16. Espectro de la secuencia "ia". Tamaño de ventana 90 muestras. Frecuencia de muestreo 11.025 Hz. Izquierda: 128 puntos, centro: 256 puntos, derecha 512 puntos.

7. CONCLUSIONES

Hemos descrito de una forma general y práctica la Transformada de Fourier para su uso correcto mediante la implementación con ordenador, mencionando el efecto que nos produce los distintos parámetros que intervienen en el cálculo de los espectros variantes en el tiempo.

Prácticamente en todas las referencias se habla de espectros en banda ancha (300Hz) y banda estrecha(40Hz); algunos autores proponen la combinación de ambas. Pensamos que son herencia de los espectrógrafos y recomendamos un tamaño equivalente al periodo del fundamental. Dependería de la voz en cuestión, pero sería válido considerar, después de un conjunto de pruebas empíricas, que 110Hz

podría ser una media válida. Ello nos lleva a ventanas de 100 muestras para un periodo de muestreo de 11025Hz. El número de ceros que recomendamos es rellenar hasta 256. Elegimos esta cifra por se fácilmente computable mediante la Transformada Rápida de Fourier.

La ventana tipo Hamming es la que mejores resultados nos ha dado y es la utilizada en general por todas la publicaciones.

El aspecto general de los espectros es muy ruidoso, y los métodos utilizados para realzarlos no consiguen el nivel de calidad deseable. Consideramos que este es un tema importante y que existen muchos aspectos por investigar. El espectro es una herramienta muy importante para el estudio de la fonética acústica y por lo tanto es un campo de interés para una mayor investigación.

8. REFERENCIAS

[Aug94] F. Auger & P. Flandrin, "Then Why and How of Time-Frequency Reassignment", *IEEE Symp. On Time-Frequency and Time-Scale Analysis*, octubre 1994, pp. 197-200.

[Bri88] E. O. Brigham, *The Fast Fourier Transform and its Applications*, Prentice-Hall, Gran Bretaña, 1988.

[Bri95] R. Briston-Johnson, "A detailed Analysis of a time-domain format-corrected pitch-shifting algorithm", *J. Audio Eng. Soc.*, vol. 43(5), mayo 1995, pp. 340-352.

[Coh89] L. Cohen, "Time-Frequency Distributions - A Review", *Proc. IEEE*, vol. 77(7), julio 1989, pp. 941-981.

[Coo93] M. Cooke, S. Beet & M. Crawford, *Visual Representations of Speech Signals*, Wiley, Inglaterra, 1993.

[Cha95] V. R. Chari & C. Y. Espy-Wilson, "Adaptative Enhancement of Fourier Spectra", *IEEE Trans. Speech and Audio Processing*, vol. 3(1), enero 1995, pp. 35-39.

[Che91] S. Cheung & J. S. Lim, "Combined Multi-Resolution (Wideband/Narrowband) Spectogram", *Proc. ICASSP'91*, 1991, pp. 457-460.

[Shi97] Y. Shin, H. Choi & Ch. Kim, "A New Method For Enhanced Spectrogram of Speech", *Proc. ICSP'97*, agosto 1997, pp. 623-628.

[Gon87] R. C. Gonzalez & P. Wintz, *Digital Image Processing*, Addison-Wesley, EE.UU., 1987.

[Mar87] J. Martí Roca, "FFT como herramienta de análisis en fonética", *Estudios de fonética experimental,* mayo 1987.

[Pla95] F. Plante & W. A. Ainsworth, "Formant Tracking Using Reassigned Spectrum", *Proc. EUROSPEECH'95*, septiembre 1995, pp. 741-744.

[Rab93] L. R. Rabiner & B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, New Jersey, 1993.

[Ran95] M. Rangoussi & A. Delopoulos, "Recognition of unvoiced stops from their time-frequency representation", *Proc. ICASSP'95*, vol. 1, 1995, pp. 792-795.

[Smi94] R. Smits, "Accuracy of quasistationary analysis of highly dynamic speech signals", *JASA*, vol. 96(6), diciembre 1994, pp. 3401-3415.

[Sol84] M. J. Solé Sabater, "La experimentación en fonética y fonología", *Estudios de fonética experimental*, vol. 1, pp. 1-70.

[Tho94] T. G. Thomas, P. C. Pandey & S. D. Agashe, "A PC-Based Multi-resolution Spectrograph", *Inst. Electronics & Telecom. Engrs.*, vol. 40(2 & 3), marzo-junio 1994, pp.105-108.