Capítulo 3

Análisis y representación de señales

"Habrá señales en el sol, en la luna y en las estrellas,
y sobre la tierra, perturbación de las naciones, ..."

(Lucas 21,25)

Contenido

3.1. Introducción

EN la práctica la mayoría de las señales se encuentran en el dominio del tiempo. Esta representación no siempre es la más apropiada cuando se tiene por objetivo su manipulación o clasificación posterior. En muchos casos una gran parte de la información distintiva se encuentra "oculta" en el contenido frecuencial de la señal o en alguna otra forma de representación. Es por ello que un sistema completo de análisis y clasificación de señales está constituido en primer término por una etapa de procesamiento de la señal. El objetivo de esta etapa es el de extraer la información relevante de la señal por medio

de algún tipo de transformación. De esta forma se convierte la señal temporal "cruda" en otra clase de representación para su análisis posterior. Esta representación puede ser de tipo paramétrica, cuando existe un modelo subyacente en términos de cuyos parámetros queda definida la señal. Esto lleva a excelentes resultados si estas suposiciones son válidas, pero obviamente no es de aplicabilidad general. El caso no paramétrico corresponde a la situación donde no existe un modelo *a priori*, y a lo sumo se realizan suposiciones de índole general sobre la naturaleza de la señal y/o del sistema que la generó. Se considera que el hecho de lograr una representación adecuada es de fundamental importancia para la solución de problemas relacionados con el procesamiento de señales. Tanto es así que se llega a decir que una vez encontrada la representación "óptima" el problema está prácticamente resuelto [59]. Cualquiera que sea el tipo de representación empleada se supone que cuanto mejor evidencie las características significativas (y las preserve de posibles distorsiones), los patrones generados serán más fáciles de analizar e identificar por las etapas subsiguientes.

La señal de voz es una de las señales fisiológicas más estudiadas. Existe un amplio rango de posibilidades para poder representarla, dentro del cual se encuentran algunos ejemplos que ya se han utilizado en el Capítulo 2, como por ejemplo la evolución de la energía de corta duración o de la cantidad de cruces por cero. Probablemente, la representación más importante de la voz es el espectro de corta duración. Por lo tanto, los métodos de análisis espectral fueron considerados durante mucho tiempo como el núcleo principal de la etapa de procesamiento de señales.

Como se ha mencionado anteriormente este enfoque más tradicional establece una serie de hipótesis que distan bastante de lo que ocurre en las situaciones reales. Entre éstas figuran hipótesis de linealidad, invarianza temporal, y estadística significativa de segundo orden. Otra hipótesis muy utilizada consiste en asumir la ortogonalidad de los elementos implicados en el análisis de una señal, o que es posible su proyección en espacios de pequeñas dimensiones con un error despreciable[1].

A pesar de la simplicidad y elegancia matemática de estos conceptos y de su éxito inicial, a medida que se pretenden incluir aspectos más complejos de la realidad en las aplicaciones, se encuentran también diferentes limitaciones. Es por ello que más recientemente se han comenzado a considerar enfoques alternativos basados en ideas más generales, que incorporan aspectos relacionados con no-linealidad, no-estacionariedad y estadística de orden superior.

[1] Generalmente, en el enfoque clásico, la nueva representación lograda posee una cantidad de información considerablemente menor, aunque todavía significativa. Sin embargo esta característica no siempre es deseable.

Son precisamente estos enfoques no-convencionales a los que se les ha dado un mayor énfasis en el desarrollo de este trabajo. Los enfoques no convencionales suelen ser más complejos y costosos que el enfoque clásico, y de ninguna manera pretenden reemplazarlo, simplemente tienden a aportar soluciones alternativas en los casos en que se llega más allá de los límites de su aplicabilidad.

En este capítulo se presentarán los tipos de análisis más comunes disponibles para lograr diferentes representaciones de la información contenida en una señal. Esta presentación no será exhaustiva, revisando sólo aquellos análisis que revistan algún interés para el desarrollo del problema planteado en este trabajo. De todas formas el enfoque será más bien general, especificando cuando así se requiera para el caso de la señal de voz. El material está principalmente orientado al análisis en tiempo continuo, aunque se realizan las aclaraciones pertinentes respecto de las versiones de tiempo discreto cuando se estiman importantes para las aplicaciones. El capítulo está organizado de la siguiente manera. Para comenzar se expondrán brevemente los aspectos más relevantes de las técnicas clásicas que sentaron las bases para el análisis de la señal de voz durante las últimas tres décadas (Sección 3.2). Las siguientes dos secciones (3.3 y 3.4) se dedicarán a presentar los fundamentos de los enfoques no convencionales, ésto es los aplicables al caso no estacionario y/o no lineal. El análisis mediante la transformada onditas (Sección 3.3.3) será tratado con un poco más de detalle debido a su carácter más reciente y a su conexión con las técnicas de codificación rala que detallaremos en la Sección 3.4.2 (y a las cuales dedicaremos varios capítulos especiales donde expondremos las razones de su posible aplicación al problema considerado en este trabajo). A continuación (Sección 3.5) se presentarán aquellas técnicas que se basan en algunas características perceptuales o modelos de producción de la señal de voz[2], de acuerdo a lo discutido en el Capítulo 2. En la Sección 3.6 se revisarán los aspectos relacionados con las características de robustez al ruido y a las distorsiones de los análisis presentados.

3.2. Análisis lineal invariante en el tiempo

En esta sección se presentarán los fundamentos de la transformada de Fourier que ha dominado el análisis lineal de señales estacionarias. El análisis de Fourier constituye un campo muy amplio con numerosos resultados teóricos y aplicaciones, por lo que en esta sección se presentan sólo los aspectos más significativos en relación con este trabajo. Para una revisión más detallada se puede consultar la extensa bibliografía al respecto

[2]Varias de estas técnicas pueden considerarse como clásicas para el análisis del habla.

(por ejemplo [32] o [129]).

3.2.1. Transformada de Fourier

Una herramienta muy útil para el análisis de señales es la transformada de Fourier de tiempo continuo (FT). Esta transformación ha sido aplicada principalmente a señales estacionarias, es decir, aquellas cuyas propiedades no cambian con el tiempo. Para esta clase de señales, la transformación lineal estacionaria más "natural" es la FT [42].

Definición 3.1 *Sea $x(t) \in L^2(\mathbb{R})$ entonces su transformada de Fourier $X(f)$ existe y puede calcularse mediante:*

$$X(f) = \left\langle x(t),\ e^{j2\pi ft} \right\rangle = \int_{-\infty}^{\infty} x(t)\ e^{-j2\pi ft} dt. \tag{3.1}$$

Los coeficientes de análisis $X(f)$ definen la noción de *frecuencia global* en una señal. Como resultado, el análisis de Fourier funciona adecuadamente si $x(t)$ esta compuesta por un número reducido de componentes estacionarias. Sin embargo, cualquier cambio abrupto en una señal no estacionaria $x(t)$ se esparce sobre todo el eje de frecuencias en $X(f)$. Desde la perspectiva de bases ortogonales discutida anteriormente, ésto se debe a que las exponenciales complejas contra las que se está comparando la señal tienen soporte infinito (podríamos decir que son "eternas")Es por ello que para el correcto análisis de señales no estacionarias se requiere algo más que la transformada de FT.

La razón de la particular aptitud de la FT para tratar señales derivadas de sistemas LTI se basa en el hecho de que las exponenciales complejas que forman la base de Fourier constituyen las autofunciones de estos sistemas. Esto puede demostrarse de la siguiente forma. Supongamos que excitamos un sistema LTI con respuesta al impulso $h(t)$ mediante una exponencial compleja $e^{j2\pi ft}$. Entonces su respuesta se puede calcular mediante la siguiente convolución:

$$y(t) = \int_{-\infty}^{\infty} h(u)\ e^{j2\pi f(t-u)} du.$$

Es posible reescribir esta expresión de la siguiente forma:

$$e^{j2\pi ft} \int_{-\infty}^{\infty} h(u)\ e^{-j2\pi fu} du = H(f)\ e^{j2\pi ft},$$

donde $H(f)$ es la FT de $h(t)$ evaluada en f, y constituye un autovalor, mientras que $e^{j2\pi ft}$ es la autofunción buscada (es decir la misma que se utilizó para excitar al sistema).

Para reconstruir $x(t)$ a partir de sus proyecciones $X(f)$ en términos de las exponenciales complejas de la base tenemos la siguiente fórmula de inversión:

$$x(t) = \int_{-\infty}^{\infty} X(f)\, e^{j2\pi ft} df.$$

3.3. Análisis lineal no estacionario

La mayoría de las señales reales no son estacionarias. Para el caso de las señales de voz, se espera de hecho que "cambien" sus características de forma continua o al menos cada unos pocos milisegundos. Existen distintas maneras de representar señales cuyas características frecuenciales varían con el tiempo. En forma genérica se habla en este caso de diferentes *representaciones tiempo-frecuencia*, sin embargo es posible también representar la variación temporal de características diferentes a la frecuencia. En esta sección se revisaran los casos lineales, mientras que en la sección siguiente se presentarán los no-lineales.

Se ha visto que la FT realiza un análisis en términos de características frecuenciales globales de la señal, suponiendo válida la hipótesis de estacionariedad. Es por ello que se puede decir que esta transformación es prácticamente "ciega" a cualquier cambio de frecuencia instantánea, mientras se mantenga el contenido frecuencial global. En la Figura 3.1 podemos observar dos señales formadas por combinaciones de dos tonos de diferente frecuencia. Se puede apreciar como, a pesar de los cambios en la secuencia de aparición de los tonos, la magnitud de la FT se mantiene inalterada[3]. La aproximación más común para resolver el problema de la no estacionariedad consiste en introducir la dependencia temporal en el análisis de Fourier pero preservando su linealidad. La idea es introducir un parámetro de "frecuencia local" (local en el tiempo o instantánea), de tal forma que la transformada de Fourier local mire a la señal a través de una ventana sobre la cual ésta es aproximadamente estacionaria.

Una forma equivalente de ver al análisis frecuencial dependiente del tiempo es como una modificación de las funciones exponenciales de la base de Fourier, de manera que se concentren más en el tiempo (y como consecuencia menos en la frecuencia). Esta transformación se denomina *transformada de Fourier de corta duración* (STFT). Sin embargo para muchas señales intrínsecamente transitorias ésto no es suficiente para superar las

[3]Es claro que la información acerca del cambio de secuencia se halla contenida en la fase, sin embargo no resulta directa su interpretación. Un experimento similar podría realizarse comparando la magnitud espectral de funciones $\delta(t-\tau)$ con $\tau \in \mathbb{R}$.

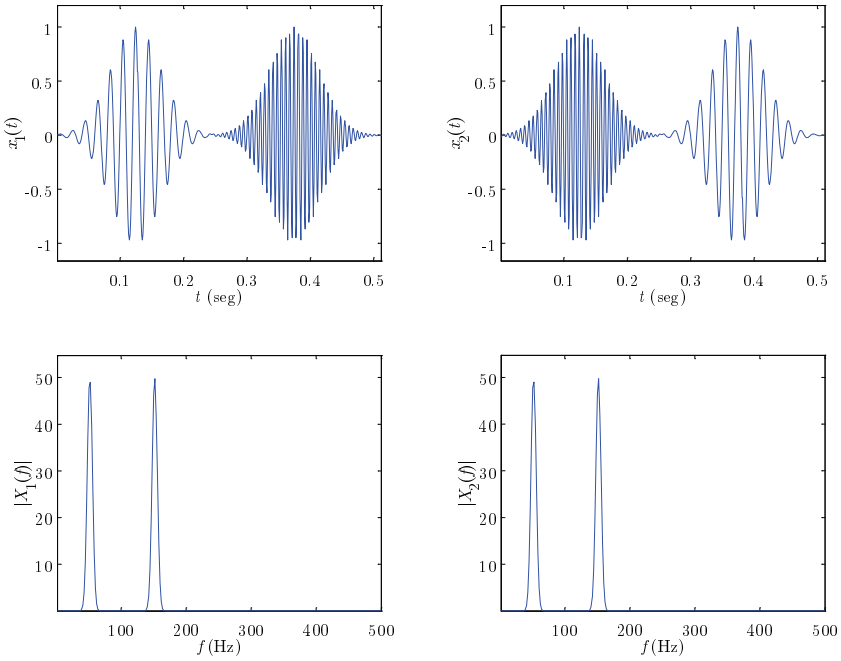

Figura 3.1: *Dos señales $x_1(t)$ y $x_2(t)$ formadas por combinaciones de dos tonos ventaneados de 10 Hz (A) y 30 Hz (B) (arriba) y sus respectivas magnitudes espectrales $|X_1(f)|$ y $|X_2(f)|$ calculadas mediante la FT (abajo). Es posible observar como el espectro no refleja los cambios en la secuencia de aparición de los tonos.*

limitaciones del enfoque original y ha sido necesario buscar enfoques alternativos, como por ejemplo la teoría de las onditas.

El enfoque de *análisis por tramos* o ventanas para el caso del habla se ha extendido también a otro tipo de representaciones como las que se revisarán en la Sección 3.5 (aunque no todos resultan estrictamente lineales). Por ejemplo se ha utilizado ampliamente para los coeficientes LPC, el cepstrum, o incluso la energía [32]. Este "parche" al análisis estacionario se denomina también de *análisis de corta duración*. La idea general consiste en tomar análisis pensados para "larga duración" y "adaptarlos" para ser aplicados al caso de "corta duración". Surge así el problema práctico de trabajar con pequeños trozos o tramos de la señal, obtenidos a partir de una ventana, sobre los cuales se supone que la misma es estacionaria. En el caso de la señal de voz los parámetros del aparato fonador varían en forma continua, sin embargo en la práctica es posible considerarla como estacionaria por tramos tomando ventanas de 10 a 30 mseg de ancho [32]. Las ideas anteriores dan lugar a la siguiente definición .

Definición 3.2 *Se define una señal ventaneada $x_v(t)$ como el producto de la señal $x(t)$ con una ventana desplazada $g(t)$ en el tiempo una cantidad τ:*

$$x_v(t;\tau) = x(t)g(t-\tau), \tag{3.2}$$

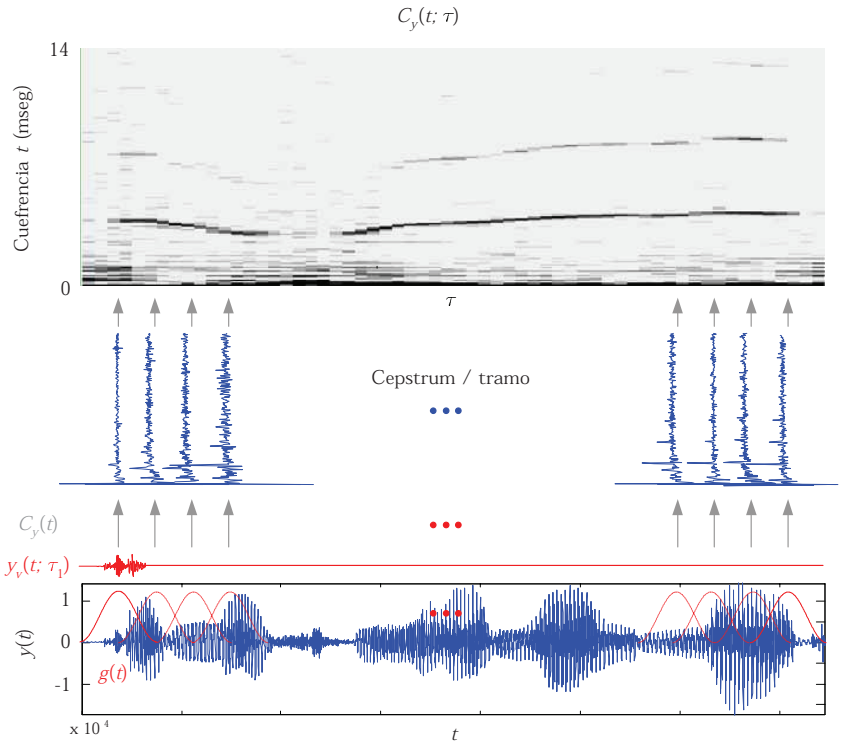

Figura 3.2: *Esquema del análisis por tramos ejemplificado para el caso del cepstrum real de corta duración de una señal de voz. Como se verá en la Sección 3.5 el primer pico del cepstrum está asociado a la frecuencia de entonación de la voz en los fonemas sonoros o F_0, cuya variación puede apreciarse claramente en este análisis.*

donde $g(t)$ posee soporte compacto en un intervalo en el que se considera que $x(t)$ es estacionaria.

La nueva señal $x_v(t)$ es, en la práctica, un trozo de la señal $x(t)$ que ha sido "cortado" por la ventana $g(t)$. Ésto fuerza a la señal a tomar valor cero fuera del intervalo de corta duración $[t, t+\tau]$[4]. Con estas consideraciones, el procesamiento de corta duración es equivalente al procesamiento de larga duración para la señal ventaneada, tomando un trozo diferente para cada desplazamiento τ (Figura 3.2).

3.3.1. Señales analíticas y frecuencia instantánea

La *frecuencia instantánea* [39] ha sido considerada frecuentemente como una forma de introducir la dependencia del tiempo en las representaciones espectrales. Para introducir esta función es necesaria primero la siguiente definición.

[4]En realidad los valores del particular "recorte" de la señal dependen no solo del soporte de la ventana sino también de su morfología.

Definición 3.3 *Sea $x(t)$ una señal con valores en $\mathbb{R}$, se define a la* señal analítica asociada *$x_a(t)$ con valores en $\mathbb{C}$ como:*

$$x_a(t) = x(t) + j\,\mathsf{H}\{x(t)\}\,,$$

donde $\mathsf{H}\{\cdot\}$ es el operador de la transformada de Hilbert.

La interpretación de esta definición resulta sencilla en el dominio frecuencial puesto que $X_a(f)$ posee sólo frecuencias positivas, esto significa que los valores para las frecuencias negativas se han removido y se han duplicado los de las frecuencias positivas, dejando sin cambios la componente de continua:

$$\begin{aligned} X_a(f) &= 0 && \text{si } f < 0, \\ X_a(f) &= X(0) && \text{si } f = 0, \\ X_a(f) &= 2X(f) && \text{si } f > 0\,. \end{aligned}$$

De esta manera se puede obtener una señal analítica a partir de una real forzando a cero su espectro para frecuencias negativas, lo que no altera el contenido de información debido a que para una señal real $X(-f) = X^*(f)$.

A partir de esta señal analítica es posible entonces definir de forma única los siguientes conceptos.

Definición 3.4 *Dada una señal analítica $x_a(t)$, se define la* amplitud instantánea *$a(t)$ y la* frecuencia instantánea *$f(t)$ como:*

$$\begin{aligned} a(t) &= |x_a(t)|, \\ f(t) &= \tfrac{1}{2\pi}\tfrac{d\arg x_a(t)}{dt}\,. \end{aligned}$$

La frecuencia instantánea funciona muy bien para cuando tratamos con señales sencillas, que no posean más de una componente a la vez. Sin embargo, si la señal no es de banda angosta, la frecuencia instantánea promedia diferentes componentes espectrales en el tiempo. Para lograr en estos casos mayor precisión se requiere una representación tiempo-frecuencia de la señal $x(t)$ compuesta de características espectrales dependientes del tiempo. A esta representación la denominaremos $S(t, f)$ por su dependencia de t y f y es posible definir ahora la frecuencia local f de una manera adecuada a través de ella. Esta representación es similar a la notación usada en música, la cual muestra también "frecuencias" (notas) tocadas en distintos instantes de tiempo.

3.3.2. Transformada de Fourier de corta duración

En la sección anterior se presentó el concepto de frecuencia instantánea y se discutieron sus limitaciones. En esta sección presentaremos otra alternativa de aplicación más general basada en la FT. La transformada de Fourier (3.1), fue adaptada por primera vez por Gabor para definir $S(t, f)$ como sigue [43].

Definición 3.5 *Considere una señal $x(t)$ y asuma que es estacionaria si se la observa a través de una ventana $g(t)$ de extensión limitada, centrada en el tiempo τ. Entonces la transformada de Fourier (3.1) de las señales ventaneadas $x(t)\, g^*(t-\tau)$ constituye la transformada de Fourier de corta duración:*

$$S_F(\tau, f) = \int_{-\infty}^{\infty} x(t)\, g^*(t-\tau)\, e^{-j2\pi ft} dt. \tag{3.3}$$

Esta transformación mapea la señal $x(t)$ en una función bidimensional en el plano tiempo-frecuencia (τ, f). El parámetro f en (3.3) es similar a la frecuencia de Fourier y por ello esta transformación hereda varias de las propiedades de la transformada de Fourier. Sin embargo, aquí el análisis también depende de la elección de la ventana $g(t)$. Este punto de vista muestra a la STFT como un proceso de ventaneo de la señal. Una visión alternativa está basada en la interpretación del mismo proceso como un banco de filtros . Para una frecuencia f dada, (3.3) filtra la señal en el tiempo con un filtro pasa-banda cuya respuesta al impulso es la función ventana modulada a esa frecuencia[5]. De esta manera la STFT puede ser vista como un banco de filtros modulado [3, 123].

Finalmente es posible también pensar en la STFT como un método para comparar la señal $x(t)$ con un diccionario de señales $\phi_{\tau,f}(t) = g(t-\tau)\, e^{j2\pi ft}$, bien concentradas en el tiempo o en la frecuencia:

$$S_F(\tau, f) = \left\langle x(t), g(t-\tau)\, e^{j2\pi ft} \right\rangle.$$

De esta manera $\left\langle x(t), \phi_{\gamma=(\tau,f)}(t) \right\rangle$ provee una "porción" de la información de $x(t)$ que corresponde a una región del plano (t, f) cuya localización y características dependen de la dispersión tiempo-frecuencia de $\phi_\gamma(t)$. A esta región del plano tiempo-frecuencia se la conoce como *rectángulo de Heisenberg* de $\phi_\gamma(t)$ y está relacionada con el conocido principio que describiremos a continuación (Ver Figura 3.3).

[5]Para este caso la variación de frecuencia es en realidad continua por lo que podría pensarse en un banco con un número infinito de filtros pasa-banda, cercanos en frecuencia central tanto como se quiera.

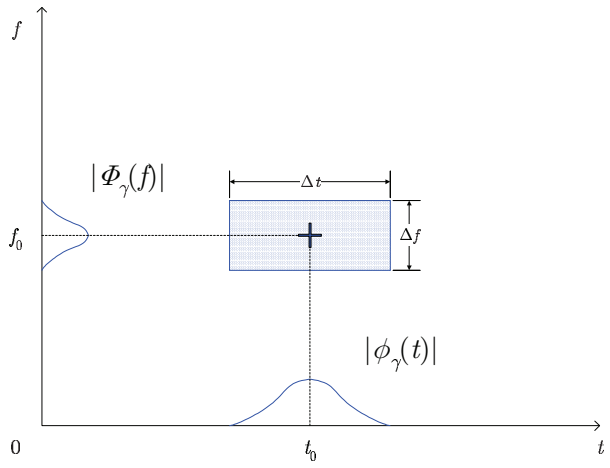

Figura 3.3: *Rectángulo de Heisenberg que representa al átomo $\phi_\gamma(t)$ en el plano tiempo-frecuencia (adaptado de [106]). La localización y las características de este rectángulo dependen de la dispersión tiempo-frecuencia de $\phi_\gamma(t)$, la que está condicionada por el principio de incertidumbre.*

En la Figura 3.4 se pueden apreciar nuevamente dos señales formadas por diferentes secuencias de dos tonos ventaneados junto con sus respectivas magnitudes espectrales calculadas mediante la STFT. Aquí puede observarse como la representación obtenida de esta forma (que resulta difiere de un espectrograma solo por un cuadrado) permite determinar que tono aparece en cada momento para cada caso. Sin embargo se observa cierta incertidumbre respecto al momento exacto donde comienza uno y termina el otro. Ésto se debe al problema de la resolución $t-f$ de la STFT que se discutirá a continuación.

Con las ideas esbozadas hasta aquí es posible deducir algunas relaciones respecto a la resolución de la transformada STFT en el tiempo y en la frecuencia. En este caso será la función ventana la que determinará principalmente las propiedades tiempo-frecuencia del análisis. Dada una función ventana $g(t)$ y su transformada de Fourier $G(f)$, se define el ancho de banda Δf del filtro como:

$$\Delta f^2 \triangleq \frac{\int_{-\infty}^{\infty} f^2 \cdot |G(f)|^2 \, df}{\int_{-\infty}^{\infty} |G(f)|^2 df}. \tag{3.4}$$

Dos sinusoides pueden discriminarse sólo si están más separadas que Δf, por lo que define la resolución en frecuencia de la STFT. De forma similar la dispersión en el tiempo está dada por:

$$\Delta t^2 \triangleq \frac{\int_{-\infty}^{\infty} t^2 \cdot |g(t)|^2 \, dt}{\int_{-\infty}^{\infty} |g(t)|^2 dt}. \tag{3.5}$$

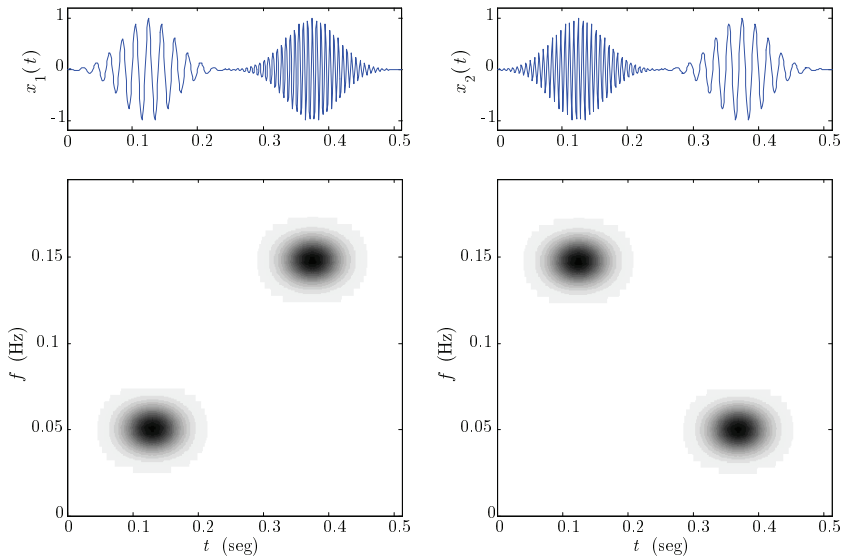

Figura 3.4: *Dos señales $x_1(t)$ y $x_2(t)$ formadas por combinaciones de dos tonos ventaneados de 10 Hz (A) y 30 Hz (B) (arriba) y sus respectivas magnitudes espectrales $|STFT_1(f,t)|$ y $|STFT_2(f,t)|$ calculadas mediante la STFT (abajo). A diferencia de lo que ocurría para el caso de la FT (Figura 3.1), en la representación obtenida a través del espectrograma es posible determinar la secuencia de aparición de ambos tonos.*

Dos pulsos pueden discriminarse sólo si están más lejos que Δt.

Por otra parte, ni la resolución temporal, ni la frecuencial pueden ser arbitrariamente pequeñas, porque su producto debe cumplir la siguiente relación conocida como *principio de incertidumbre de Heisenberg* :

$$\Delta t \cdot \Delta f \geqslant \frac{1}{4\pi}. \tag{3.6}$$

El hecho de fijar la resolución $t - f$ hace que si por ejemplo se quiere analizar una señal compuesta de pequeños transitorios junto con componentes cuasi-estacionarias esta puede ser analizada con buena resolución en tiempo o en frecuencia, pero no ambas.

Una cuestión fundamental con respecto a la STFT es que una vez que se elige una ventana la resolución tiempo-frecuencia queda fija para todo el análisis. Se puede demostrar que el valor óptimo para la relación (3.6) (es decir la igualdad) se da cuando la ventana $g(t)$ es de tipo gaussiana. Para este caso (3.3) se denomina *transformada de Gabor*:

$$g_{Gabor}(t) = e^{\frac{-18t^2}{2}}.$$

La fórmula de reconstrucción para $x(t)$ es la siguiente:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} S_F(\tau, f) g(t-\tau)\, e^{j2\pi f\tau} df\, d\tau. \tag{3.7}$$

Por supuesto que $\left\{g(t-\tau)\, e^{j2\pi f\tau}\right\}_{\tau,f\in\mathbb{R}^2}$ constituye un conjunto sumamente redundante y para asegurar que pueda realizarse la reconstrucción se requiere que $g(t) \in L^2(\mathbb{R})$.

3.3.3. Transformada ondita

En casi veinte años de existencia, el área de las onditas (en inglés *wavelets*) ha llegado a ser de suma importancia para el procesamiento de señales. Esto se debe en gran parte a su manera natural de tratar las señales no-estacionarias. En lugar del análisis tradicional basado en la transformada de Fourier, que examina una señal a una resolución fija, la transformada de onditas posee la característica de hacerlo a distintas escalas (ó resoluciones). Ésto implica un análisis más similar al realizado por los sistemas sensoriales biológicos, en particular análogo al caso del oído según se discutió anteriormente.

El área de las onditas empezó a desarrollarse a mediados de los años 80's con el trabajo de Meyer [109]. Desde entonces ha demostrado ser una herramienta importante para el procesamiento de señales debido a que, desde su concepción original, incorpora de manera más directa elementos de tipo transitorio. Este enfoque permite, por ejemplo, el análisis de discontinuidades, picos o cambios abruptos en la señal. En esta sección mencionamos los principales resultados, para las onditas continuas y discretas en una dimensión, necesarios para nuestro desarrollo posterior. Excelentes referencias son Daubechies [29], Wojtaszczyk [175], y Mallat [105, 104].

Las ideas detrás del enfoque basado en onditas son las siguientes. Para evitar la limitación en resolución de la STFT es posible dejar que Δt y Δf cambien en el plano tiempo-frecuencia de manera de obtener un análisis con resolución variable (o resoluciones múltiples). Una manera de producir ésto y seguir cumpliendo con (3.6) es hacer que la resolución en el tiempo se incremente con la frecuencia central de los filtros de análisis. Más específicamente se impone que:

$$\frac{\Delta f}{f} = c, \tag{3.8}$$

donde c es una constante.

En este caso el banco de filtros de análisis está compuesto por filtros pasa-banda de ancho de banda relativo constante. Otra manera de ver ésto es que la respuesta en frecuencia de los filtros se dispone en escala logarítmica, en lugar de estar regularmente espaciada en el eje de la frecuencia. Este tipo de bancos de filtros se utiliza, por ejemplo, para modelar la respuesta en frecuencia de la cóclea (ver Capítulo 2). Ésto produce una buena resolución temporal a altas frecuencias junto con una buena resolución frecuencial a bajas frecuencias, lo que generalmente funciona muy bien para analizar las señales del

mundo real. Ello se debe a que en muchos casos requerimos conocer más exactamente el momento de los cambios abruptos y las frecuencias de los cambios lentos de la señal.

La *transformada ondita continua* (CWT) sigue las premisas anteriores agregando una simplificación: todas las respuestas al impulso de los bancos de filtros son definidas como versiones escaladas (es decirse expandidas o comprimidas) del mismo prototipo $\psi(t)$:

$$\psi_a(t) = \frac{1}{\sqrt{|a|}}\,\psi\left(\frac{t}{a}\right),$$

donde a constituye un factor de escala. Ésto resulta en la siguiente definición.

Definición 3.6 *Considere una señal $x(t) \in L^2(\mathbb{R})$, y una función $\psi(t) \in L^2(\mathbb{R})$ denominada ondita madre, entonces la transformada ondita continua de $x(t)$ se define de la siguiente forma:*

$$S_w(\tau, a) = \int_{-\infty}^{\infty} x(t)\,\frac{1}{\sqrt{|a|}}\,\psi^*\left(\frac{t-\tau}{a}\right)dt, \tag{3.9}$$

donde $a \in \mathbb{R}$ y se supone además que $\psi(t)$ es suficientemente regular (derivadas continuas hasta cierto orden) y cumple con la siguiente condición de admisibilidad:

$$\int_{-\infty}^{\infty} \psi(t)\,dt = 0. \tag{3.10}$$

La condición de admisibilidad impuesta en la definición hace que $\psi(t)$ oscile en el tiempo como una onda de corta duración y de allí la denominación de ondita. Ésto constituye una función de tipo pasa-banda. Dado que se usa la misma función prototipo $\psi(t)$ (llamada ondita básica o madre) para todos los filtros ninguna escala es privilegiada por lo que el análisis ondita es *autosimilar* a todas las escalas. Además esta simplificación es útil para deducir las propiedades matemáticas de la CWT.

Una de las ventajas de la transformada onditas es que se tiene a disposición una gran cantidad de funciones o familias de onditas con diferentes propiedades. Éste aspecto será revisado con mayor detalle en la Sección 5.3.2. En la Figura 3.5 se puede observar la ondita de Morlet (parte real) a diferentes escalas y localizaciones.

Para establecer una relación con la ventana modulada utilizada en la STFT se puede elegir $\psi(t)$ como sigue:

$$\psi(t) = g(t)\ e^{-2j\pi f_0 t}.$$

Entonces la respuesta en frecuencia de los filtros de análisis satisface (3.8) de la siguiente forma:

$$a = \frac{f_0}{f}.$$

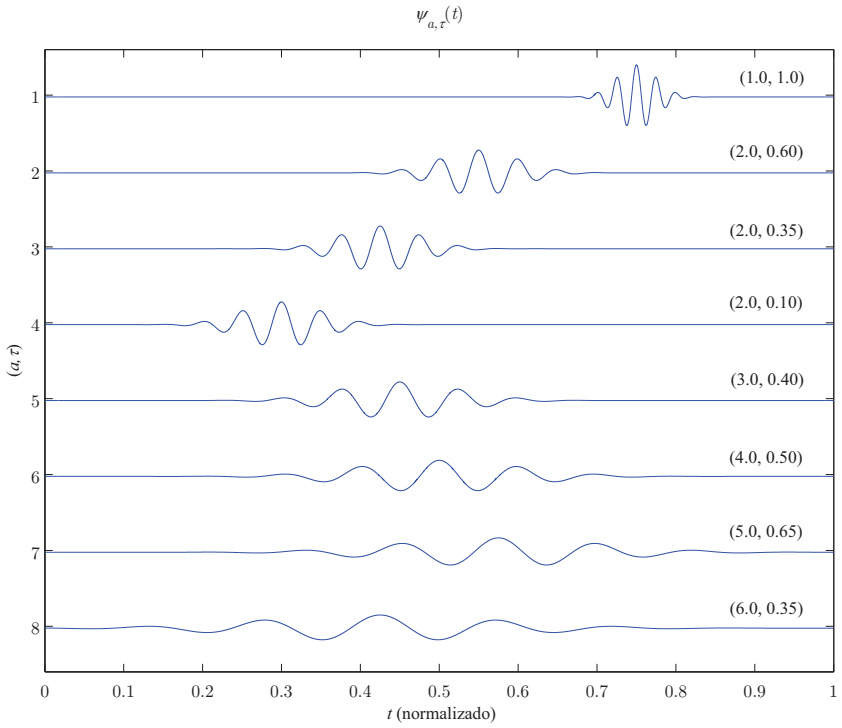

Figura 3.5: *Ejemplos de onditas de Morlet (parte real) a distintas escalas y localizaciones. Las gráficas se realizaron de acuerdo a los parámetros* (a, τ)*, para una ondita* $\psi_{a,\tau}(t)$ *como en (3.11).*

Es importante notar aquí, que la frecuencia local $f = af_0$, tiene poco que ver con la descripta para la STFT y está ahora más bien asociada con el esquema de escalas. Como resultado esta frecuencia local, cuya definición depende de la ondita madre, no está más ligada a la frecuencia de modulación sino a las distintas escalas temporales. Por esta razón se prefiere en general utilizar el término "escala" y no "frecuencia" para la CWT. La escala para el análisis ondita tiene el mismo significado que la escala en los mapas geográficos : grandes escalas corresponden a señales comprimidas ("vistas de lejos") mientras que escalas pequeñas corresponden a señales dilatadas ("vistas de cerca o ampliadas").

Otra manera de introducir la CWT es pensar a las onditas como un diccionario de átomos tiempo-frecuencia. De hecho, estos átomos ya aparecieron en (3.9) y se hacen más evidentes si ahora se reescribe como:

$$S_w(\tau, a) = \langle x(t), \psi_{a,\tau}(t) \rangle = \int_{-\infty}^{+\infty} x(t)\, \psi^*_{a,\tau}(t)\, dt,$$

que mide la similitud entre la señal $x(t)$ y las onditas $\psi_{a,\tau}(t)$, que son versiones escaladas

y trasladadas de la ondita básica o prototipo $\psi(t)$:

$$\psi_{a,\tau}(t) = \frac{1}{\sqrt{|a|}}\,\psi\left(\frac{t-\tau}{a}\right). \tag{3.11}$$

El análisis ondita resulta en un conjunto de coeficientes que nos indican cuan cerca está la señal de una función particular del diccionario. De esta manera esperaríamos que cualquier señal pudiera ser representada como una descomposición en onditas lo que significa que $\{\psi_{a,\tau}(t)\}_{a,\tau\in\mathbb{R}}$ debería comportarse como una base ortogonal [109]. Por supuesto que éste no siempre es el caso con $\{\psi_{a,\tau}(t)\}_{a,\tau\in\mathbb{R}}$ ya que constituyen un conjunto sumamente redundante, sin embargo aún satisfacen la fórmula de reconstrucción:

$$x(t) = c\int_{-\infty}^{\infty}\int_{>0}^{\infty} S_W(\tau,a)\,\psi_{a,\tau}(t)\frac{da\,d\tau}{a^2}, \tag{3.12}$$

con la condición, ya discutida, de que $\psi(t)$ sea de energía finita y pasa-banda. Esta condición es más restrictiva que la impuesta para la STFT que sólo requiere que la ventana tenga energía finita.

3.4. Análisis no lineal y/o no estacionario

En la Sección anterior se revisaron diferentes soluciones lineales para el problema de las representaciones tiempo-frecuencia, principalmente STFT y WT. El enfoque lineal posee algunas restricciones que pueden limitar su utilidad en algunas aplicaciones. Como alternativa existen varios métodos que se apartan de la linealidad en alguno de sus pasos para obtener una representación de la señal. En este caso es posible armar el siguiente cuadro taxonómico de la representaciones $t-f$ no lineales que se presentarán en esta sección:

1. Distribuciones (bilineales o cuadráticas):
   - *a*) Directas o regulares: Wigner-Ville.
   - *b*) Convolucionadas o clase de Cohen:
     1) Choi-Williams.
     2) Espectrograma.
     3) Escalograma.
2. No-lineales (no lineales no cuadráticas):

a) Series de distribución $t - f$

b) Métodos de aproximación:

1) Búsquedas:

a' Búsqueda de bases (en inglés *basis pursuit*, BP)

b' Búsqueda por coincidencia (en inglés *matching pursuit*, MP)

2) Elección adaptativa de la base: Mejor base ortogonal (BOB).

3.4.1. Distribuciones $t - f$ cuadráticas

En esta sección se revisaran algunos métodos que permiten obtener una representación de la señal en términos de la *distribución tiempo-frecuencia* de su energía. En base a ello, y con las normalizaciones necesarias, es posible interpretar estas distribuciones o densidades en el sentido estadístico como medidas de la probabilidad de encontrar energía de la señal considerada en determinada región del plano $t - f$.

Wigner-Ville

Tanto la STFT como la WT se calculan correlacionando la señal con familias de átomos tiempo-frecuencia, según se discutió anteriormente. Por lo tanto su resolución $t-f$ está limitada por la de los átomos correspondientes. Idealmente se querría definir una densidad de energía sin ninguna perdida de resolución. La *distribución de Wigner-Ville* (WVD) posee propiedades muy interesantes en este sentido. Ésta se obtiene comparando la información de la señal con su propia información en otros instantes y frecuencias. Ésto podría verse también como la utilización de una ventana de análisis formada por una versión desplazada de la misma señal. De allí la siguiente definición [106].

Definición 3.7 *Considere una señal $x(t) \in L^2(\mathbb{R})$ entonces la distribución de Wigner-Ville de $x(t)$ se define de la siguiente forma [106]:*

$$P_{WV}(t,f) = \int_{-\infty}^{\infty} x\left(t+\frac{\tau}{2}\right) x^*\left(t+\frac{\tau}{2}\right)\ e^{-j2\pi f\tau} d\tau.$$

La WVD consiste en una función con valores reales que permite la localización de las estructuras tiempo-frecuencia de la señal. Si la energía de $x(t)$ está bien localizada en el tiempo alrededor de t_0 y en la frecuencia alrededor de f_0 entonces $P_v(f,t)$ posee su energía centrada en (t_0, f_0), con una dispersión igual a la de $x(t)$ en el tiempo y en

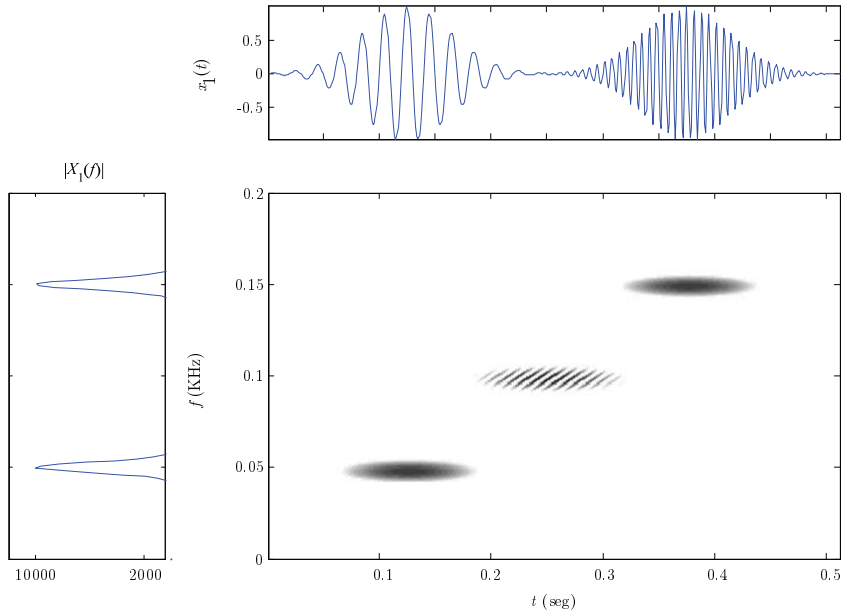

Figura 3.6: *Distribución de Wigner Ville (centro) de dos tonos (arriba) y su correspondiente espectro (izquierda). Es posible observar una mejora en la localización frecuencial con respecto al espectrograma pero a costa de la aparición de los términos cruzados (comparar con la Figura 3.4).*

la frecuencia. La WVD posee algunos inconvenientes, como la existencia de términos de interferencia y la no positividad (Ver Figura 3.6). En la Figura 3.7 se pueden apreciar estos efectos en el análisis de un trozo de voz, así también como la mejora en resolución espectral comparada con el espectrograma de banda angosta.

Clase de Cohen

Para atenuar los términos cruzados de la WVD se requiere realizar una promediación $t-f$, lo que resulta otra vez en una perdida de resolución. Cuando esta promediación se realiza a través de la convolución de la WVD mediante un núcleo adecuado se obtiene una familia general de distribuciones $t-f$ que se denomina *clase de Cohen* [126]:

$$P_{C_\theta}(t,f) = P_{WV}(t,f) * \theta(t,f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \theta(t-t', f-f') P_{WV}(t',f') dt' df'.$$

donde $\theta(t-t', f-f')$ es un núcleo de convolución.

Se puede demostrar que el espectrograma, el escalograma y todas las distribuciones $t-f$ cuadráticas pueden escribirse de esta forma.

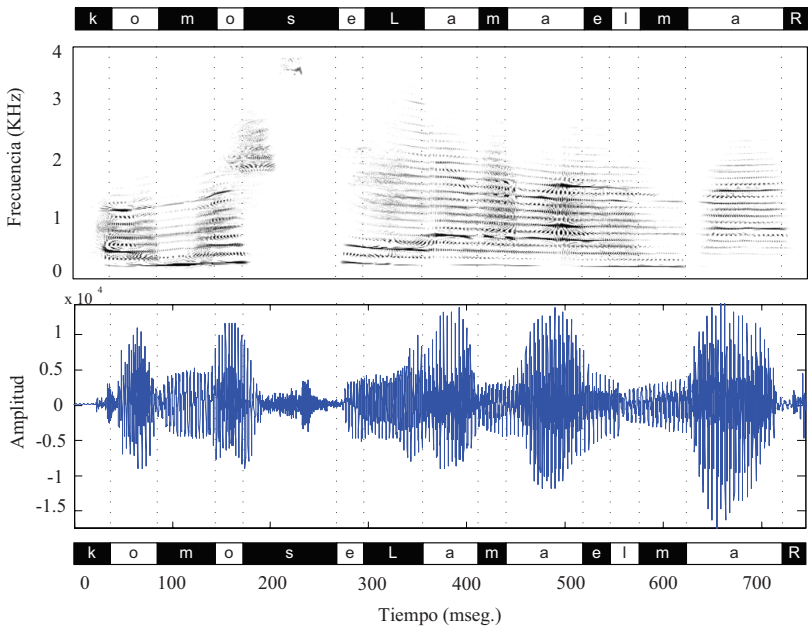

Figura 3.7: *Distribución de WV de un trozo de voz. A pesar de que la localización frecuencial mejora notablemente con respecto al espectrograma (sin sacrificar resolución temporal), puede apreciarse también la aparición de términos cruzados que "oscurecen" el análisis mediante la introducción de elementos no presentes en la señal original.*

Distribución de Choi-William La distribución de Choi-Willian consiste en convolucionar la WVD con un núcleo exponencial bidimensional (cuasi cónico) [126]:

$$\theta(t,f) = \ e^{-\mu^2 t^2 f^2}.$$

donde μ fija la "dispersión" del núcleo.

Esta distribución no conserva todas las propiedades de WVD, pero disminuye los términos cruzados de manera importante (Ver Figura 3.8).

Espectrograma

A partir de (3.3) es posible definir una densidad de energía que se denomina espectrograma:

$$P_F(\tau,f) = |S_F(\tau,f)|^2 = \left| \int_{-\infty}^{\infty} x(t) \cdot g^*(t-\tau) \cdot \ e^{-2j\pi f t} dt \right|^2 . \tag{3.13}$$

El espectrograma mide la energía de $x(t)$ en la vecindad de (τ, f) especificada por el rectángulo de Heisenberg de $g(t-\tau)\ e^{-2j\pi f t}$. Esta densidad de energía ya no constituye un análisis de tipo lineal, si no más bien bilineal (o lineal con respecto a la energía de $x(t)$). En la Sección 3.4 se generalizará el uso de este tipo de representaciones a través de la denominada distribución de Wigner-Ville.

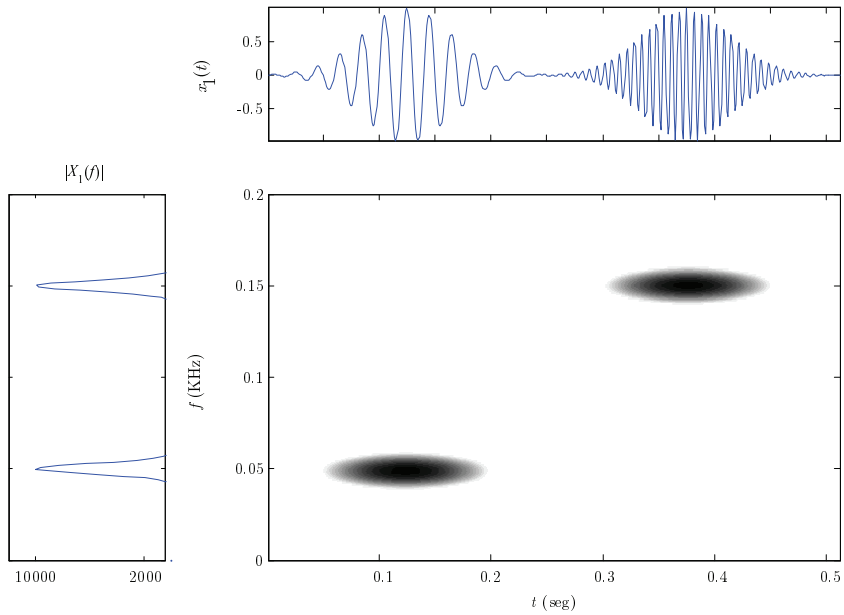

Figura 3.8: *Distribución de Choi-William de dos tonos donde pueden apreciarse la casi desaparición de los términos cruzados. Para esta distribución, que pertenece a la clase de Cohen, el núcleo empleado es exponencial.*

El problema de la resolución $t - f$ de la STFT se traslada directamente al espectrograma. Ésto constituye un problema frecuente para el análisis de señales de voz y es lo que ha llevado a la utilización conjunta de dos tipos de espectrogramas para analizar las distintas características de la voz (Ver Figura 3.9). En los espectrogramas de banda angosta la ventana temporal es relativamente larga, con lo que se logra una muy buena resolución en frecuencia pero una no tan buena localización de los eventos en el tiempo. Ésto último es especialmente útil para la detección de formantes. En los espectrogramas de banda ancha la situación es exactamente la inversa y permiten extraer mejor parámetros como el período de entonación.

Escalograma

La WT permite definir una densidad de energía tiempo-frecuencia $P_W(\tau, f)$ que mide la energía de $x(t)$ en el rectángulo de Heisenberg de cada ondita $\psi_{a,\tau}(t)$ centrada en $(f = \eta/a)$ [106]:

$$P_W(\tau, f) = |S_W(\tau, a)|^2 = |S_W(\tau, \eta/f)|^2 . \tag{3.14}$$

El escalograma "hereda" las mismas propiedades de la WT con respecto a la variación de la resolución en el plano $t-f$ (Ver Figura 3.10). En la Figura 3.11 se puede apreciar el escalograma de un trozo de señal de voz comparado con el correspondiente espectrograma de banda angosta.

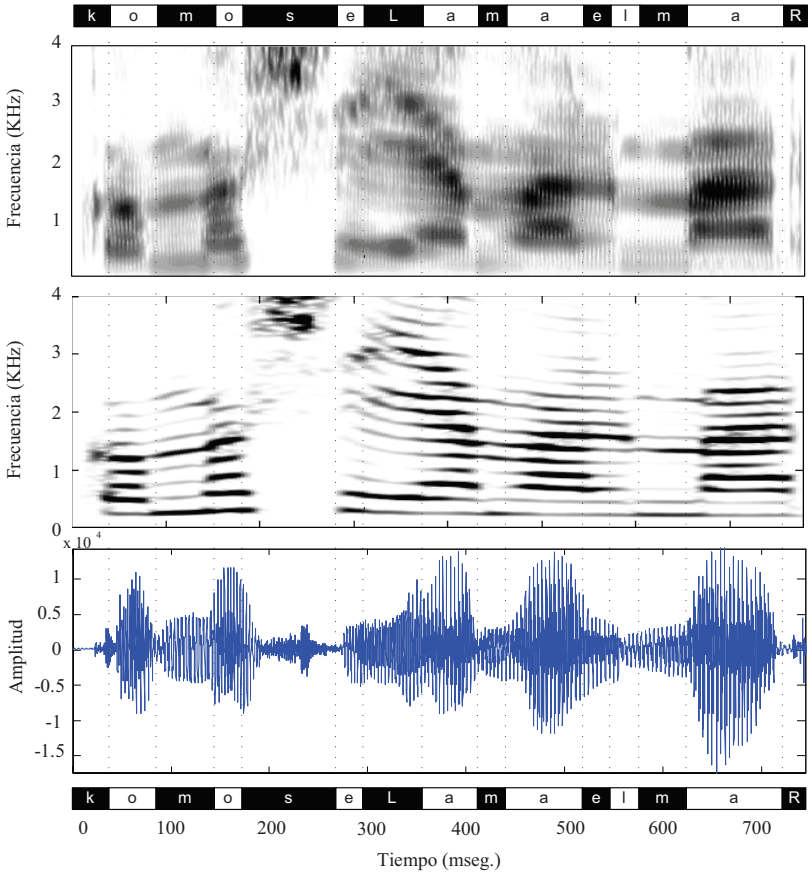

Figura 3.9: *Ejemplo de espectrograma de banda ancha (arriba), angosta (centro) y sonograma correspondiente (abajo). En los espectrogramas de banda angosta se logra una muy buena resolución en frecuencia, lo que resulta especialmente útil para la detección de formantes a partir de las "líneas" horizontales. En los espectrogramas de banda ancha la situación es inversa, lo que permite medir fácilmente eventos temporales como por ejemplo el período de entonación a partir de las "estrías" verticales.*

3.4.2. Representaciones $t - f$ no lineales

Series de distribución tiempo-frecuencia

Es posible descomponer la WVD en una serie de funciones tipo Gabor bidimensionales de la forma [126]:

$$P_{VS}(t,f) = \sum_{i,k,p,q} d_{i,k,p,q} H_{i,k,p,q}(t,f),$$

donde $d_{i,k,p,q} \in \mathbb{C}$ son los pesos de los distintos átomos de Gabor tiempo-frecuencia (bidimensionales):

$$H_{i,k,p,q}(t,f) = \; e^{-\alpha(t-iT)^2 - \frac{1}{\alpha}(2\pi f - kF)^2} \; e^{j(pT2\pi f + qFt - qFpT)},$$

donde T y F son los pasos de muestreo en tiempo y frecuencia, p y q reflejan la tasa de oscilación en tiempo y frecuencia respectivamente.

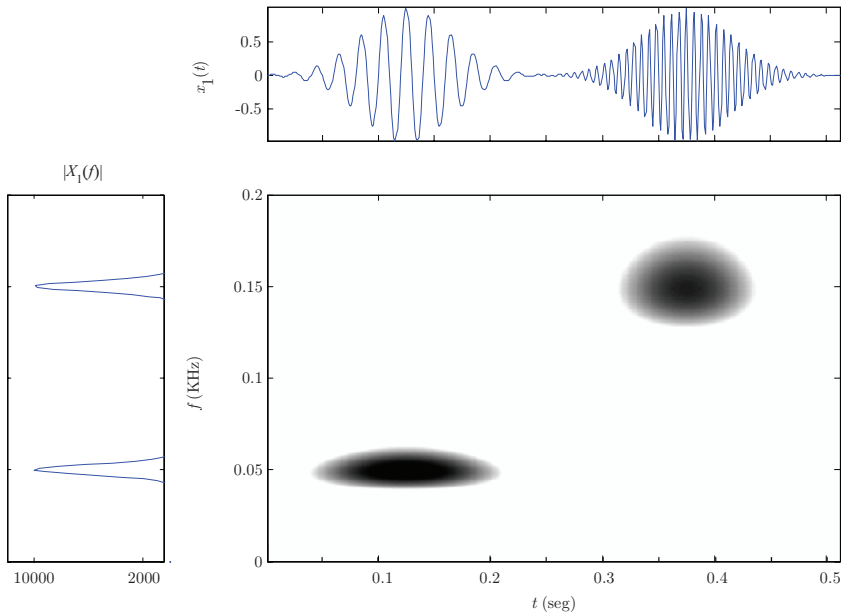

Figura 3.10: *Escalograma (centro, calculado con la ondita de Morlet) de una señal formada por dos tonos ventaneados (arriba, igual a la señal de la parte izquierda de la Figura 3.4) y su correspondiente espectro (izquierda). En el escalograma es posible observar fácilmente el cambio de resolución en las diferentes zonas del plano $t - f$.*

En este caso los términos de interferencia son generalmente los de mayor orden. Entonces es posible eliminar estos términos y dejar sólo aquellos que tienen información más importante. Esto constituye en realidad un proceso no-lineal pero de esta forma es posible conservar la mayoría de las propiedades de la WVD.

Representación mediante aproximaciones no lineales

La idea detrás de los *métodos de aproximación* es en algún sentido similar a la del análisis de señales. En ambos casos es posible ver a una señal como formada por varias componentes de interés. Sin embargo, en el primer caso se presta mayor atención a la evolución del error de la aproximación de esta señal a medida que cambia el número de componentes considerado [106]. Por supuesto que la aproximación puede ser completamente lineal si para ello se utiliza la definición de combinación lineal [135] y un subconjunto de elementos seleccionados de antemano a partir de una base ortogonal[6]. Sin embargo, aunque la base sea ortogonal, es posible lograr una aproximación no-lineal a una función $x(t) \in \mathcal{H}$ si se seleccionan de esta base M elementos en forma adaptativa, esto es dependiendo de la señal.

El interés aquí consiste en explorar casos aún más generales, en los cuales se trabaja

[6]Por ejemplo si selecciono los primeros N elementos de la base.

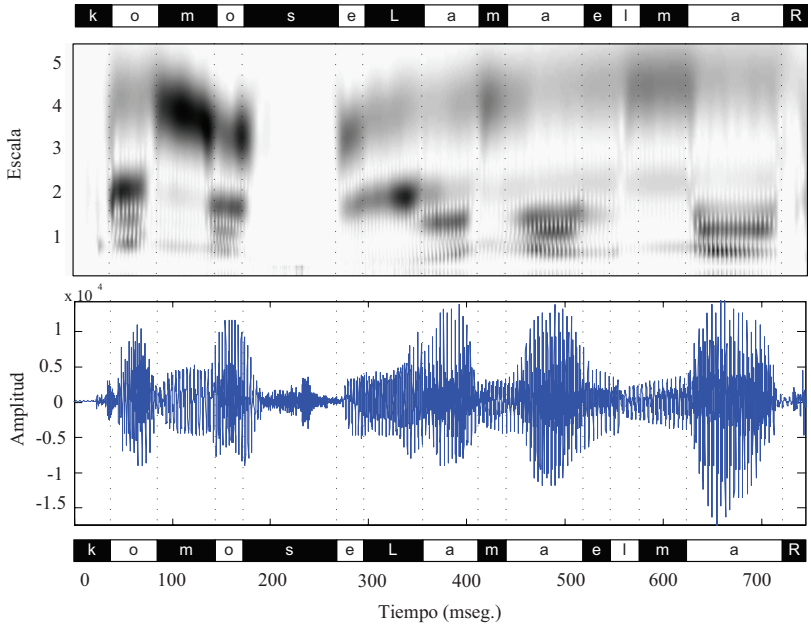

Figura 3.11: *Escalograma calculado utilizando la ondita de Morlet de un trozo de voz. En el escalograma se puede observar un cambio en las estrías verticales de las escalas bajas (que tienen que ver con la F_0) hacia patrones de líneas horizontales en las medias y altas (que corresponden a F_1 y F_2). Por ello se podría decir que, de acuerdo con la escala, esta representación juega un rol mixto que permite para apreciar detalles que antes se evidenciaban con el espectrograma de banda ancha o el de banda angosta por separado (Ver Figura 3.9).*

con diccionarios y donde la relación entre la señal y los coeficientes que la representan tampoco es lineal. Cuando los M elementos que se utilizan para realizar la aproximación de una señal $x(t)$ dependen de la señal en sí misma, ésto puede expresarse como:

$$x(t) = \sum_{i \in \Gamma_{M_x}} c_i \phi_i(t). \tag{3.15}$$

donde $\Phi = \{\phi_i\}_{i \in \Gamma}$ es el diccionario utilizado y $\Gamma_{M_x} \subset \Gamma$ es el subconjunto de elementos del diccionario cuya selección depende de $x(t)$.

Aunque la descomposición en términos de un conjunto bien conocido y comprendido de elementos de un diccionario puede resultar interesante en algunas aplicaciones, no es necesariamente la única forma de realizar este tipo de análisis. A veces no es directamente la "forma" de los elementos del diccionario la que resulta importante, sino más bien las propiedades derivadas de sus dependencias recíprocas y su relación a los datos originales. Muchas de estas transformaciones encuentran el diccionario a partir de los datos, utilizando algunas restricciones adecuadas. Normalmente estas transformaciones son de la naturaleza estadística y están estrechamente relacionadas al análisis estadístico de datos[7].

[7] A veces se denomina a este tipo de métodos como *análisis mediante diccionarios dependientes de*

En relación con este trabajo es importante resaltar aquellos métodos que permiten aproximar una señal en términos de una pequeña cantidad de elementos significativos. Éste es el caso de una *representación rala* [118] que ya se ha introducido en la Sección 4.2. Se dedicará todo el Capítulo 6 para presentar con mayor detalle los fundamentos y las ventajas generales de este tipo de representaciones. Entre estas ventajas se pueden citar: robustez intrínseca al ruido aditivo, mayor separabilidad, óptima generalización, eficiencia en la codificación de la información de la señal y mejor resolución de eventos.

3.5. Análisis específicos para el habla

En esta sección se presentarán aquellos análisis concebidos específicamente para el caso de la señal de voz que se han desarrollado a partir del estudio de las características perceptuales del oído o de un modelo de producción del habla (Ver Capítulo 2). Éstos últimos se basan en suponer a este modelo como lineal aunque con algunas consideraciones adicionales que se describirán a continuación. En general se emplean en las representaciones finales conceptos derivados de ambos esquemas (percepción-producción). Estos análisis "especiales" se pueden considerar como convencionales en el área de análisis del habla.

3.5.1. Coeficientes de predicción lineal

Una de las técnicas paramétricas de análisis del habla más potentes es el método de *análisis predictivo lineal* (LPC) [128]. Este método se convirtió en la técnica predominante para estimar los parámetros del habla básicos, por ejemplo la frecuencia fundamental, las formantes, el espectro, funciones del área del tracto vocal y para representar el habla para transmisiones de baja velocidad o almacenamiento. La importancia de este método está en su habilidad de proveer estimaciones extremadamente precisas de los parámetros del habla, y en su relativa velocidad de cálculo.

Se trata de una técnica intrínsecamente de tiempo discreto. La idea básica detrás del LPC es que las muestras actuales de la señal de voz pueden ser aproximadas por una combinación lineal de sus muestras anteriores. O sea que la señal de voz $y[n]$ puede aproximarse mediante la salida $\hat{y}[n]$ de un sistema lineal de tiempo discreto frente a una excitación o entrada $x[n]$, lo que resulta compatible con un modelo lineal discreto *auto-regresivo* (AR) de producción de la voz como el de la Figura 3.12.

los datos o adaptativos.

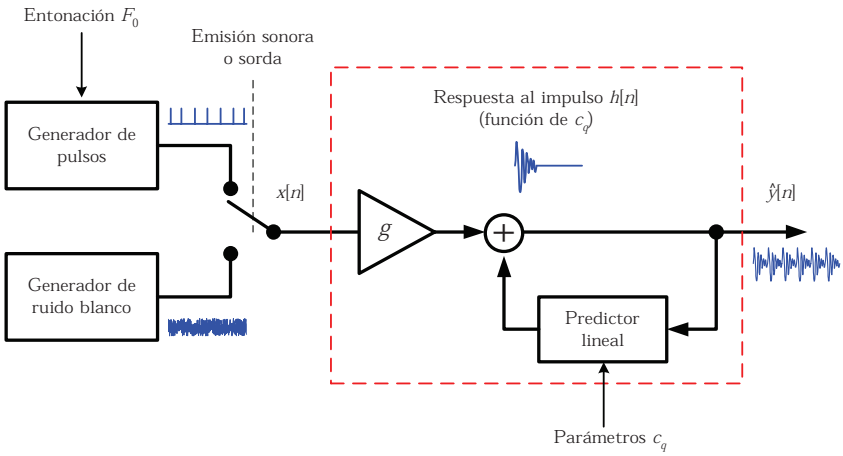

Figura 3.12: *Diagrama para el modelo AR del aparato fonador, donde la señal de voz $y[n]$ se aproxima mediante la salida $\hat{y}[n]$ de un sistema lineal de tiempo discreto frente a una excitación o entrada $x[n]$. Esta señal de excitación puede ser un tren de pulsos o ruido blanco dependiendo de si el fonema considerado es sonoro o sordo respectivamente.*

Definición 3.8 *Se denominan coeficientes de predicción lineal $c_q \in \mathbb{R}$ a aquellos que satisfacen la siguiente ecuación:*

$$\hat{y}[n] = -\sum_{q=1}^{Q} c_q \, y[n-q] + g \, x[n],$$

donde $y[n]$ es la versión de tiempo discreto de la señal, $\hat{y}[n]$ su versión estimada, c_q son los coeficientes de predicción que pesan las muestras sucesivas (y dan cuenta de la relación entre ellas), y $g \in \mathbb{R}$ es la ganancia de la excitación $x[n]$. Para este caso es posible, mediante la minimización del valor esperado[8] de la suma de las diferencias cuadradas entre las muestras reales del habla y las predichas linealmente, determinar un único conjunto de coeficientes de predicción c_q:

$$\frac{\partial \; \mathcal{E}[(y[n] - \hat{y}[n])^2]}{\partial c_q} = 0. \tag{3.16}$$

De esta forma, para tramos relativamente estacionarios, el habla puede ser modelada mediante un sistema lineal que puede ser excitado por pulsos cuasi periódicos (durante habla sonora), o ruido aleatorio (durante habla sorda) (Ver Figura 3.12). Los métodos de predicción lineal proveen una forma precisa, confiable y robusta para la estimación de los parámetros que caracterizan este sistema lineal.

Aplicado al procesamiento del habla, el término predicción lineal se refiere a una variedad de formulaciones esencialmente equivalentes de la modelización de la señal de

[8] Aquí se ha supuesto que x y y son *v.a.*.

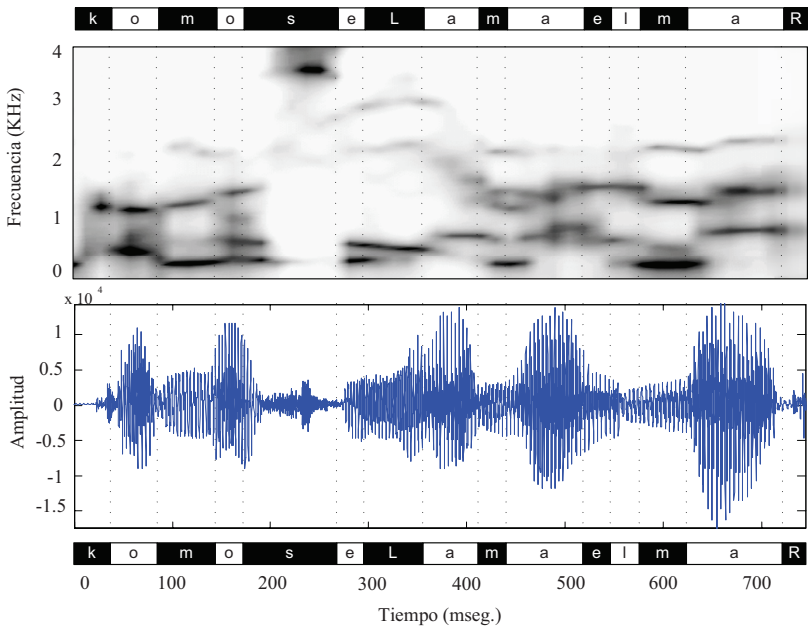

Figura 3.13: *Espectrograma "suavizado" estimado a partir de los coeficientes LPC de un trozo de voz. Es posible observar como en este suavizado se pierden algunas de las componentes espectrales más finas pero se conservan rasgos importantes como las resonancias asociadas a las frecuencias formantes.*

voz. Las diferencias entre estas formulaciones son comúnmente de enfoque o tienen que ver con los detalles de los cálculos usados para obtener los coeficientes de predicción.

Basado en esta teoría, y en sus implicaciones, se ha desarrollado una gran variedad de aplicaciones del análisis LPC al procesamiento del habla. Se han diseñado esquemas para la estimación de todos los parámetros básicos del habla mediante el análisis LPC. Finalmente, estas técnicas han sido usadas en muchos sistemas de análisis y procesamiento del habla para tareas como verificación e identificación de hablantes, ASR, clasificación, derreverberación, entre otras [32]. En la Figura 3.13 se puede observar un espectrograma "suavizado" estimado a partir de los coeficientes LPC de un sistema AR de orden 16. Obsérvese como la información relativa a la frecuencia glótica se pierde mediante este suavizado.

3.5.2. Análisis cepstral

Un análisis comúnmente empleado para la señal de voz es el denominado *cepstrum*. De acuerdo al modelo de producción de la voz que hemos presentado en la Sección anterior, ésta corresponde a la salida de un sistema lineal ante una excitación de entrada. Ésto quiere decir que la señal de voz está compuesta por una señal de excitación convolucionada con la respuesta al impulso del modelo del tracto vocal (Ver Sección 2.2.1).

Ésto resulta similar al planteo anterior, salvo por el hecho de que ahora el modelo es de tiempo continuo:

$$y(t) = x(t) * h(t). \tag{3.17}$$

En general se tiene acceso sólo a la salida $y(t)$ de este sistema, pero frecuentemente es deseable eliminar una de las componentes $x(t)$ o $h(t)$, de tal forma de poder examinar la restante. La eliminación de una de estas dos señales es, en general, un problema difícil. Sin embargo, existen métodos para resolver este tipo de problemas cuando las señales están combinadas mediante la convolución como en este caso. Uno de estos métodos es el *análisis cepstral*.

Para esbozar sus fundamentos conceptuales se puede realizar el siguiente razonamiento. Si se realiza la FT de (3.17), entonces la ecuación en el dominio de la frecuencia es ahora un producto:

$$Y(f) = X(f)H(f). \tag{3.18}$$

Si luego se toman logaritmos en ambos miembros de (3.18), este producto se convierte en una suma. Finalmente es posible volver a un dominio "temporal" (que se denomina *cuefrencia*) si se calcula la FT inversa de este último paso.

De esta forma se ha convertido una operación convolutiva en una simple adición, mediante el cálculo del cepstrum de $y(t)$. De aquí se desprende la siguiente definición.

Definición 3.9 *Se define al cepstrum $C_y(t)$ de la señal $y(t)$ como:*

$$C_y(t) = \mathsf{F}^{-1}\left\{\log\left(\mathsf{F}\left\{y(t)\right\}\right)\right\},$$

donde $\mathsf{F}\{\cdot\}$ es el operador de la FT. Se supone que $y(t)$ es generada por un sistema LTI.

El cepstrum representa una transformación sobre la señal de voz con dos propiedades importantes sobre sus componentes: éstas se combinan linealmente y pueden además aparecer separadas en el cepstrum. Para que esta última propiedad se cumpla es necesario que existan diferencias entre las velocidades de cambio del espectro de $X(f)$ y $H(f)$, de manera que sus componentes cepstrales aparezcan en cuefrencias diferentes. Éste es precisamente el caso de las señales de voz, especialmente para los fonemas sonoros, donde el espectro de la excitación $X(f)$ se asemeja a un tren de pulsos decreciente, mientras que la respuesta en frecuencia del tracto vocal $H(f)$ es casi-continua con sólo algunos picos (ver Figura 3.14). En la Figura 3.15 se puede observar el cepstrum real correspondiente a un fonema sonoro donde se resalta la separación producida entre las bajas y altas cuefrencias, lo que permite descomponer a la señal en la respuesta del tracto vocal y la excitación.

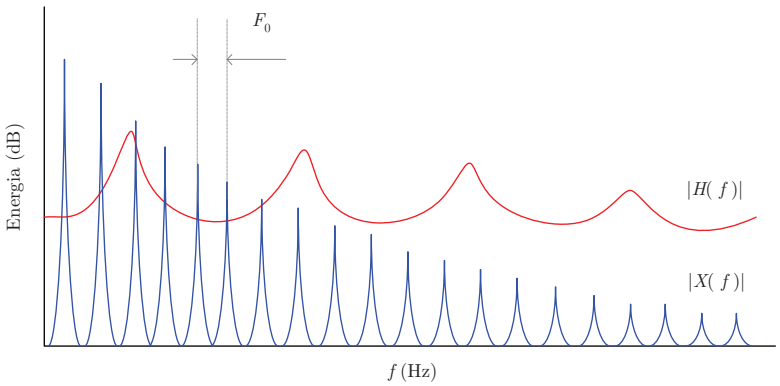

Figura 3.14: *Magnitud espectral de la excitación $X(f)$ y de la respuesta en frecuencia del tracto vocal $H(f)$ "simulado" para el caso de los fonemas sonoros. El espectro de la excitación $X(f)$ se ha representado mediante un tren de pulsos decrecientes, mientras que la respuesta en frecuencia del tracto vocal $H(f)$ mediante una función continua con varios picos correspondientes a las frecuencias formantes.*

El análisis cepstral es un caso especial de una clase general de métodos conocidos como *procesamiento homomórfico*. El cepstrum derivado del procesamiento homomórfico es comúnmente llamado *cepstrum complejo* (CC), mientras que el *cepstrum real* (RC) es generalmente más utilizado para el habla [32]. La definición del RC es equivalente a la parte real del CC sobre la región en la cual éste está definido. La diferencia básica entre el RC y el CC, es que el primero descarta información acerca de la fase de la señal, mientras que el CC la retiene. Sin embargo, en la práctica, el CC es difícil de usar, por lo cual, es empleado ampliamente el CR. Una de las más importantes aplicaciones del análisis cepstral en el procesamiento de la voz es la representación de un modelo LP a partir de parámetros cepstrales. En este caso, la señal parametrizada es de fase mínima, una condición bajo la cual el RC y el CC son esencialmente equivalentes.

Debido a que la discriminación de las frecuencias en nuestro oído no es lineal (Ver Sección 2.4) cuando se procesan señales de voz generalmente se utilizan bancos de filtros para las denominadas *bandas críticas*. Se han propuesto varios tipos de filtros para las bandas críticas, siendo una de las configuraciones más usadas el de ventana triangular, en la escala psicoacústica de mel. La relación entre la escala lineal en Hz y la escala de mel se muestra en la Figura 3.16. El mapeo es aproximadamente lineal por debajo de 1 KHz y logarítmico por encima, lo cual lleva a una aproximación comúnmente utilizada [32]:

$$f_{mel} = \frac{1000}{\log_2}\left[1 + \frac{f_{Hz}}{1000}\right],$$

en la cual f_{mel} (f_{Hz}) es la frecuencia percibida (real) en mels (Hz).

Las técnicas anteriores de extracción de características trabajan sobre el espectro de

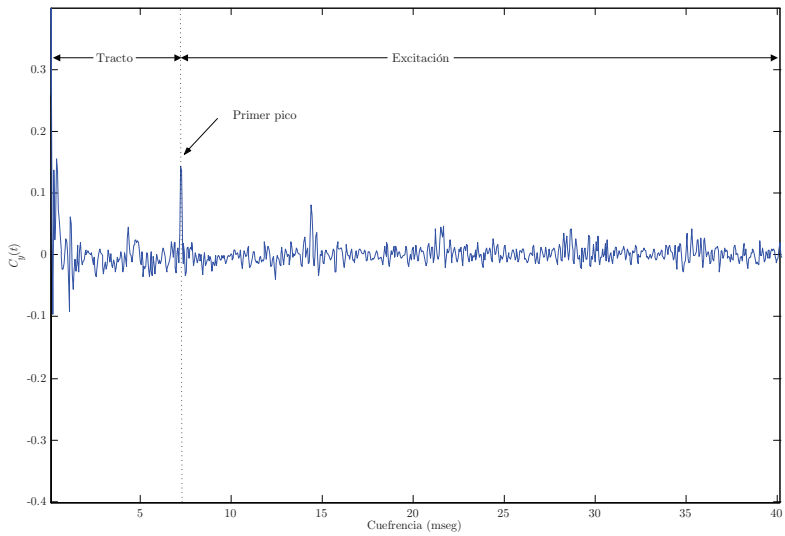

Figura 3.15: *Cepstrum real correspondiente a un trozo de una vocal /e/ sostenida de un hablante masculino. Se puede apreciar que la parte de bajas cuefrencias (antes del primer pico) corresponde a la componente de la respuesta del tracto vocal, mientras que las altas cuefrencias corresponden a la componente de la excitación.*

potencia y los coeficientes cepstrales de la señal dando una representación denominada *coeficientes cepstrales en escala de mel* (MFCC).

3.5.3. Análisis predictivo lineal perceptual

El *análisis predictivo lineal perceptual* (PLP) fue introducido por Hermansky [61] con el objetivo de alterar el espectro para minimizar las diferencias entre hablantes, pero preservando la información importante. Aunque no se darán mayores detalles aquí es posible decir que este enfoque combina nuevamente la aplicación de varias aproximaciones ingenieriles a determinadas características de la audición humana, entre las que se cuentan:

1. Resolución frecuencial no lineal en las bandas críticas, (en forma similar al mel cepstrum, pero en la escala denominada de *Bark*).

2. Asimetría de los filtros auditivos.

3. Desigual sensibilidad a diferentes frecuencias.

4. Relación no-lineal entre la intensidad física del sonido y la sensación correspondiente.

5. Integración más ancha que la de las bandas críticas.

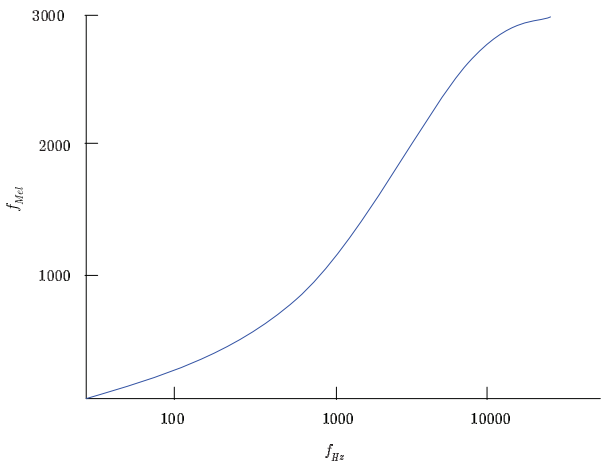

Figura 3.16: *Relación entre la escala frecuencial lineal en Hz y la escala frecuencial de mel. Esta escala esta dada por la relación entre la altura tonal percibida y la frecuencia "real" obtenida a partir de experimentos de proporcionalidad entre sensaciones.*

Posteriormente se agregaron una serie de filtros temporales para fenómenos de variación lenta, que mejoraron el comportamiento de este enfoque frente a diferentes distorsiones. Esta técnica se denominó *transformación espectral relativa PLP* (RASTA-PLP) [63]. En la Figura 3.17 es posible ver la aplicación de este análisis sobre una emisión, comparándolo con el espectrograma tradicional. En esta Figura es posible observar la pérdida de alguna información en el análisis RASTA-PLP, relacionada principalmente con la identidad del hablante, como por ejemplo la relativa a la frecuencia glótica y la entonación.

3.5.4. Modelos auditivos

Como ya se mencionó es posible aprovechar los conocimientos acerca de la anatomía y fisiología del sistema auditivo para elaborar un modelo de oído que rescate las pistas acústicas más significativas para el análisis. Para una discusión actualizada acerca de la utilización de este tipo de conocimiento en un sistema de ASR ver el artículo de Hermansky [62] (véase también [163] y [125]).

Generalmente este enfoque requiere un mayor tiempo de cálculo, aunque se han reportado modelos bastante "exactos" que se han optimizado en este sentido [31]. Mayoritariamente estos modelos contemplan hasta las denominadas representaciones auditivas tempranas (Ver Capítulo 2) con las siguientes consideraciones:

1. El meato auditivo no afecta substancialmente a la señal sonora y es por ello que se considera con transferencia igual a la unidad.

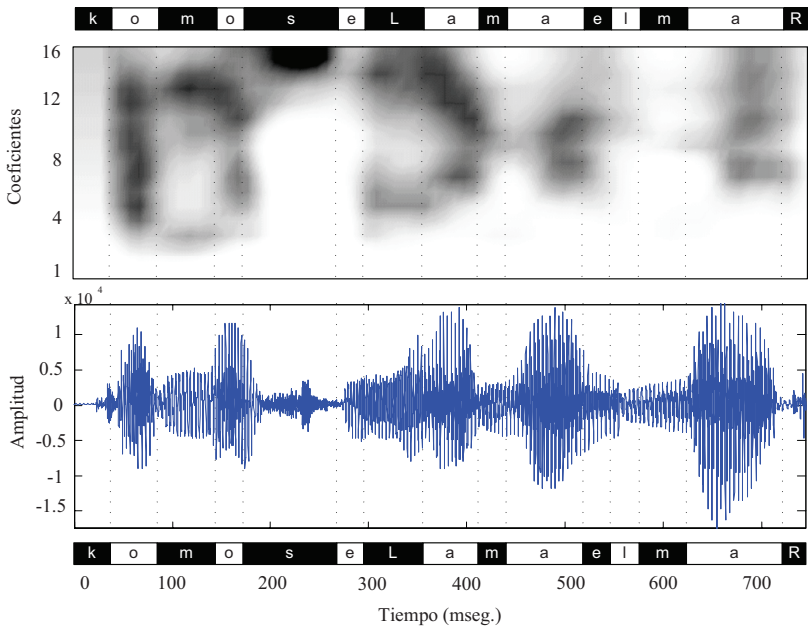

Figura 3.17: *Análisis RASTA-PLP de un trozo de voz. Para facilitar la comparación se ha realizado una interpolación bidimensional del análisis RASTA-PLP que es de naturaleza discreta. Aquí se observa la pérdida de alguna información, que aparecía claramente en el espectrograma, como por ejemplo la relativa a la frecuencia glótica y la entonación. Sin embargo esta información se asocia generalmente a características propias del hablante, y no tanto a las características de los fonemas presentes en la emisión.*

2. La cadena de huesecillos junto con los músculos correspondientes se suele asimilar a un amplificador de ganancia controlada.

3. La membrana Basilar se asimila a un banco de filtros de bandas críticas (esta etapa se considera muy importante).

4. La codificación eléctrica llevada a cabo en las células ciliadas se incorpora como una "rectificación".

5. Los nervios y los núcleos se asimilan a un mecanismo sencillo de inhibición lateral.

Con respecto al procesamiento en la corteza, se trata de un análisis de nivel superior que, por lo tanto, no forma parte de los modelos clásicos utilizados en la etapa de extracción de características o análisis sino más bien de las etapas siguientes. Sin embargo de acuerdo a los descubrimientos recientes acerca de la importancia del procesamiento llevado a cabo a nivel cortical sería deseable incluir también al menos algunos de estos aspectos[9]. El análisis mostrado en la Figura 2.19 ha sido realizado mediante un modelo auditivo [153].

[9]Por ejemplo la obtención de una representación rala e independiente, la que se ha demostrado como una característica presente en la representación cortical a través de modelos [97] y pruebas in vivo [30].

3.6. Aspectos relacionados con la robustez

La mayoría de los análisis presentados no tiene en cuenta el problema del ruido o las distorsiones de manera intrínseca. Se dice entonces que las representaciones logradas no son robustas. Sin embargo en varios casos se han incluido posteriormente algunos cambios para mejorar este aspecto. Por ejemplo, el espectro de potencia y el cepstrum no siempre son aconsejables para el reconocimiento de patrones dado que la amplitud y la forma cambian con un simple cambio de micrófono. Una alternativa simple que provee una mayor robustez en la representación de los patrones la constituye el *delta cepstrum* (ΔC) [32]. La noción aquí es que la percepción del sonido depende de la diferencia espectral. El ΔC calcula la diferencia cepstral entre el segmento de voz actual y el anterior lo que constituye una aproximación a la derivada temporal del cepstrum. Algunos sistemas usan solamente el ΔC como vector patrón mientras otros usan tanto el cepstrum como el ΔC, e inclusive la segunda derivada ($\Delta\Delta$C). Este análisis tiene la ventaja de incorporar la información temporal y posee propiedades interesantes de robustez al cambio de canal (si éste es lineal). Por otro lado sufre la desventaja de atenuar información importante en el rango de 1 a 10 Hz. Otro análisis al que se le han incorporado algunos aspectos que mejoran la robustez a ciertas distorsiones es el basado en RASTA-PLP.

Si la robustez no se incluye explícitamente en la representación, entonces es necesario aplicar algún método de limpieza o filtrado, previo a su clasificación o manipulación posterior, o de otro modo antes de realizar el análisis. Entre estas técnicas de limpieza se pueden contar las clásicas basadas en sustracción espectral o filtrado óptimo tradicional, o algunas extensiones más recientes como el *filtrado óptimo probabilístico* (POF, *Probabilistic Optimum Filtering*) o el *filtrado no lineal mediante redes neuronales.*

Debido a características especiales de los coeficientes derivados del análisis mediante onditas es posible implementar diferentes estrategias de limpieza de ruido. En forma similar, y como ya se ha discutido, una de las ventajas de las representaciones ralas es que permiten la inclusión del tratamiento del ruido de manera bastante directa. Otra posibilidad para mejorar la robustez de la representación es agregar información adicional aunque sea redundante, lo cual es compatible con algunas de los procedimientos utilizados por el sistema auditivo (Ver Capítulo 2). En este sentido se ha demostrado por ejemplo que la adición de la información contenida en los cambios de complejidad temporal de la señal de voz mejora el desempeño en ruido de los sistemas de ASR [143].

3.7. Comentarios de cierre del capítulo

En este capítulo se ha presentado un panorama organizado de una serie de técnicas que permiten la representación de señales generales, y en particular de la señal de voz. Cada una de estas representaciones resalta diferentes aspectos de la señal, utilizando por ejemplo planteos alternativos para realizar un análisis $t - f$, o inclusive utilizando aspectos relacionados con características propias de la señal de voz.

El enfoque ha estado centrado principalmente en lo que se conoce clásicamente como análisis de señales de tiempo continuo. En el Capítulo 5 se presentan un conjunto de técnicas para lograr representaciones atómicas de las señales basadas en diccionarios discretos. Sin embargo es importante remarcar que es posible lograr representaciones útiles originadas desde otras perspectivas, como ser la de modelización de señales, aspectos que abordarán en el próximo capítulo.Por ejemplo, a diferencia de la mayoría de las técnicas desarrolladas en el presente capítulo, la técnica de ICA nace principalmente como un modelo estadístico. Sin embargo, esta técnica puede utilizarse también para realizar un análisis que resalte las características significativas de los datos.

Frente a esta variedad de representaciones posibles surge nuevamente la pregunta acerca de cómo encontrar una representación óptima para una aplicación determinada. Por ejemplo, para el caso de clasificación de la señal de voz en fonemas, ¿es mejor utilizar Fourier u onditas?. Responder esta pregunta no resulta tan sencillo como parece y se retomará su discusión desde una perspectiva de modelización en el capítulo siguiente. Es posible decir hasta aquí que la comprensión de los aspectos deseables para lograr una representación “ideal”, junto con un conocimiento de las técnicas disponibles y las características principales del sistema de comunicación humano permiten orientar la búsqueda de una respuesta.