Análisis y representación de la voz mediante técnicas no convencionales

–Speech analysis and representation by means of non-conventional techniques–

por

Hugo Leonardo Rufiner

Bioingeniero, Universidad Nacional de Entre Ríos, Argentina, 1993
Maestro en Ing. Biomédica, Univ. Autónoma Metropolitana, México, 1996

Disertación doctoral enviada como requerimiento
parcial para obtener el grado de

Doctor

FACULTAD DE INGENIERÍA

Comité a cargo:

Ing. Luis F. Rocha , Director (UBA)
Dr. John Goddard Close, Codirector (UAM)

Ing. Jorge Gurlekian, Evaluador (CONICET)
Dr. Angel Plastino, Evaluador (UNLP, CONICET)
Dr. Ruben Spies, Evaluador (UNL, CONICET)
Dr. Ing. Silvano Zanutto, Evaluador (UBA, CONICET)

2005

La defensa de la tesis de Hugo Leonardo Rufiner está aprobada:

_____

Director Fecha

_____

Codirector Fecha

_____

_____

_____

_____

Universidad de Buenos Aires, Facultad de Ingeniería

2005

Análisis y representación de la voz mediante técnicas no convencionales

–Speech analysis and representation by means of non-conventional techniques–

por

Hugo Leonardo Rufiner
Doctor de la

Universidad de Buenos Aires
Facultad de Ingeniería

Ing. Luis F. Rocha , Director (UBA)
Dr. John Goddard Close, Codirector (UAM)

Resumen: La señal de voz se encuentra entre las señales naturales más estudiadas. En el campo del análisis, representación y modelado del habla se han realizado avances importantes. Sin embargo, después de más de 30 años de investigación, existen problemas que aún no han logrado resolverse satisfactoriamente. Por ello el desempeño actual de los sistemas artificiales está todavía lejos del de las personas para tareas similares. En los últimos años, debido a las limitaciones de las técnicas disponibles, varias investigaciones han seguido un enfoque diferente al tradicional para el procesamiento de señales. En el enfoque "ideal" la suposición implícita consistía en que las señales provenían de sistemas lineales invariantes en el tiempo y con estadística significativa de hasta segundo orden. El cambio de enfoque hacia uno más general originó la aparición de técnicas como las basadas en la teoría de onditas, la estadística de alto orden, el análisis de componentes independientes y la obtención de representaciones ralas de una señal. A partir de allí es posible pensar en nuevas propuestas de solución para los problemas planteados. Recientemente se han encontrado importantes conexiones entre la manera en la que el cerebro procesa las señales sensoriales y algunos de los principios que sustentan estos nuevos enfoques. Entre estos principios se pueden destacar la existencia de muy pocos elementos activos para lograr la representación de cualquier señal y la independencia estadística entre estos elementos. Como resultado emergen capacidades útiles como la super-resolución y robustez al ruido. En este trabajo se presentan los conceptos fundamentales detrás de este nuevo enfoque y se desarrollan técnicas específicas para el análisis y representación de la señal de voz. Además,

se comparan las características de las representaciones logradas con aquellas resultantes del enfoque convencional, incluyendo ejemplos de aplicación en el contexto de sistemas artificiales de clasificación de fonemas, reconocimiento del habla y limpieza de ruido.

Abstract: Speech signals are amongst the most studied natural signals. In the field of analysis, representation and modeling of speech important advances have been accomplished. However, after more than 30 years of research, there are still problems which have not been satisfactorily solved. Because of this, current performance of artificial systems is far behind that of human performance for certain speech related tasks. Due to the limitations of the available techniques, in recent years several researchers have followed a different approach to the traditional one for signal processing. In the traditional approach the implicit assumption has been that the signals come from linear time invariant systems, with significant statistics of only up to second order. The change of emphasis towards a more general one, has involved the use of techniques such as wavelet theory, higher order statistics, independent component analysis and finding sparse representations of a signal. From these ideas, different types of solutions can be proposed for a given problem. Recently, important connections have been noted between some of the principles that support these new approaches and the way in which the brain seems to process sensory signals. Among these principles it is possible to highlight the facts that the representation of a signal requires very few active elements, and there is a statistical independence among these elements. As a result of this useful properties emerge such as super-resolution and robustness to noise. In this work the fundamental concepts behind this new approach are presented and specific techniques are developed for the analysis and representation of speech signals. Furthermore, the properties of the representations obtained are compared with those found using traditional methods when applied to problems of phoneme classification, speech recognition and denoising.

Agradecimientos

Como es usual en tareas que requieren tanto tiempo y esfuerzo como ésta, mucha es la gente que nos ayuda en alguna parte del camino y resulta difícil incluir aquí a todos ellos. Quisiera empezar mi lista de agradecimientos por alguien a quien solemos dejar muchas veces en el último lugar: Dios. Él ha guiado como verdadero Padre todos los aspectos de mi vida y sin duda también este emprendimiento. También quisiera agradecer a mi Madre del Cielo, María, que con mucha paciencia ha ido abriendo mis oídos para escuchar las verdaderas "señales" de la "voz" de Dios en la realidad y me ha conducido hasta la "palabra" viva que es su Hijo. Quiero también agradecer a mi madre Silvia por su ejemplo siempre presente de que para lograr algo bueno se requiere perseverancia y dedicación. Mi padre Hugo y Nelly me brindaron su cariño y hospitalidad, además de su apoyo "logístico" en mis numerosos viajes a Buenos Aires. Mis hermanos Danilo, Giselle y María Sol han contribuido en esta empresa de muchas formas, particularmente mediante el aliento para continuar adelante. Quiero agradecer de una manera muy especial a mi esposa y compañera Stella por todo su apoyo y excepcional comprensión que hicieron posible mi dedicación a esta tarea "más allá de lo posible". A mis hijos Juan Ignacio, Santiago y Julieta por su ternura y alegría que me permitieron seguir adelante a pesar de las dificultades de todos los días. A mi director Luis Rocha por haberme iniciado y formado en el procesamiento de las señales y del habla y conducido en esta parte del camino. A mi codirector John Goddard por las largas y enriquecedoras discusiones que en numerosas oportunidades me brindó para que "pensáramos" esta tesis, guiándome siempre a pesar de los esfuerzos extras demandados por las distancias. A los integrantes de la Comisión de Seguimiento y a los miembros del Jurado evaluador por su disposición y recomendaciones. A mis compañeros de la cátedra de Bioingeniería I y del Laboratorio de Cibernética de la Facultad de Ingeniería de la UNER por su respaldo constante, especialmente por "cubrirme" durante mis numerosas estancias en la Universidad Autónoma Metropolitana y la Facultad de Ingeniería de la UBA. Aquí también debo mencionar a todos mis compañeros de la UAM que hicieron muy fructíferas y llevaderas estas estancias. En el último tramo de este esfuerzo quiero destacar el apoyo recibido de parte de mis compañeros de las Secretarias de la Facultad de Ingeniería y en particular el apoyo del Sr. Decano César Osella. Quisiera también agradecer especialmente a Diego Milone por las extensas conversaciones relacionadas con esta tesis y todas las sugerencias

aportadas. Maria Eugenia Torres fue responsable de la revisión minuciosa y paciente de todo el material, además de brindarme consejo y ánimo constante para finalizar esta etapa. Finalmente quisiera dar gracias a toda la comunidad de la FIUNER por su apoyo silencioso para que pudiera realizar esta tarea.

A Stella

Juan Ignacio, Santiago y Julieta

Índice general

Índice de figuras

Glosario

Notación

a, i	Variables escalares.
a_i	Componente o elemento i-ésimo.
$\mathbf{x}, \boldsymbol{\phi}$	Vectores columna.
$\mathbf{X}, \boldsymbol{\Phi}$	Matrices.
$\boldsymbol{\Phi}^{-1}$, $\tilde{\boldsymbol{\Phi}}$	Matriz inversa, inversa generalizada.
$\mathbf{x}^{\mathrm{T}}, \boldsymbol{\Phi}^{\mathrm{T}}$	Transpuesta de un vector o matriz.
$\nabla_{\mathbf{x}}$	Operador gradiente en las coordenadas $\mathbf{x}$.
$\Delta\mathbf{x}$	Incremento de $\mathbf{x}$.
$\mathbb{N}, \mathbb{Z}, \mathbb{R}, \mathbb{C}$	Conjuntos numéricos (naturales, enteros, reales, complejos).
$N, M, N \times M$	Dimensiones.
$\mathbb{Z}^N, \mathbb{R}^N, \mathbb{C}^N$	Espacios de vectores de dimensión N.
$\mathcal{H}$	Espacio de Hilbert.
$\mathcal{X}, \mathcal{A}$	Conjuntos en general (espacios, alfabetos, etc.).
$\lvert\mathcal{X}\rvert$	Cardinalidad, cantidad de elementos del conjunto $\mathcal{X}$.
$x(t)$, $x[n]$	Señal de tiempo continuo, señal de tiempo discreto.
a^*	Complejo conjugado de $a \in \mathbb{C}$
$\langle x, y \rangle$	Producto interno de x y y.
$x * y$	Convolución de x y y.
$\lVert x \rVert^q$, ℓ_q	Norma q de x (con $0 \leq q \leq \infty$).
e^x	Exponencial de x.
$L^2(\mathbb{R}), L^2(\mathbb{C})$	Espacios de señales continuas de energía finita.
$U \oplus V$	Suma directa de dos espacios vectoriales.
$n \bmod N$	Resto de la división entera de n módulo N.
$\approx$	Aproximadamente igual a.
$\triangleq$	Igual por definición.
$e.v.$	Espacio vectorial
$v.a.$	Variable aleatoria
$t - f$	Tiempo-frecuencia (plano)
/a/, /f/	Fonemas.
F_0	Frecuencia fundamental de los fonemas sonoros.
F_1, F_2, F_3	Frecuencias formantes de los fonemas sonoros.

$\overline{X}$	Variable aleatoria que toma valores en x.
$P_{\overline{X}}(x)$ ó $P(x)$	Función de distribución de probabilidad de $\overline{X}$.
$p_{\overline{X}}(x)$ ó $p(x)$	Función de densidad de probabilidad de $\overline{X}$.
$P_{\overline{X},\overline{Y}}(x,y)$ ó $P(x,y)$	Probabilidad conjunta de $\overline{X}$ y $\overline{Y}$.
$P_{\overline{X},\overline{Y}}(x\|y)$ ó $P(x\|y)$	Probabilidad condicional de $\overline{X}$ dado $\overline{Y}$.
$\mathcal{E}\left[\,\overline{X}\,\right]_{P(x)}$	Valor esperado de $\overline{X}$ tomado sobre $P(x)$.
$\mathcal{K}(\overline{X})$	Curtosis de $\overline{X}$.
$\mathcal{L}(\overline{X})$	Verosimilitud de $\overline{X}$.
$\mathcal{H}(\overline{X})$	Entropía de Shannon de $\overline{X}$.
$\mathcal{H}_q(\overline{X})$	Entropía de Tsallis o $q-$entropía de $\overline{X}$.
$h(\overline{X})$	Entropía diferencial de $\overline{X}$.
$j(\overline{X})$	Negentropía de $\overline{X}$.
$\mathcal{D}_{KL}(p_1\|p_2)$	Divergencia de Kullback-Leibler entre $p_1(x)$ y $p_2(x)$.
$\mathcal{I}(\overline{X};\overline{Y})$	Información mutua entre $\overline{X}$ y $\overline{Y}$.
$\mathcal{M}$, $\mathcal{M}^{-1}$	Modelo directo o generativo, modelo inverso.
$\#bits$	Cantidad de bits.
$\mathsf{H}\{\cdot\}$, $\mathsf{F}\{\cdot\}$, $\mathsf{Z}\{\cdot\}$	Operador de la transformada de Hilbert, Fourier, Z.
$X(f)$	Tranformada de Fourier de $x(t)$.
$h(t)$, $H(f)$	Respuesta al impulso, función transferencia.
$x_v(t)$	Versión ventaneada de $x(t)$.
$x_a(t)$	Señal analítica asociada a $x(t)$.
$C_y(t)$	Cepstrum de $y(t)$.
ΔC, $\Delta\Delta C$	Delta cepstrum, doble delta cepstrum.
$S_F(\tau,f)$	Transformada de Fourier de corta duración.
$S_w(\tau,a)$	Transformada ondita continua.
$P_{WV}(t,f)$	Distribución de Wigner-Ville.
$P_{C_\theta}(t,f)$	Clase de Cohen.
$P_F(\tau,f)$	Espectrograma.
$P_W(\tau,a)$	Escalograma.
$P_{VS}(t,f)$	Serie de distribución tiempo-frecuencia.
Δf, Δt	Ancho de banda, dispersión temporal.
$\delta(t), \delta[n]$	Distribución delta de Dirac, delta de Kroenecker.
$\max\limits_x f(x)$	Valor máximo de $f(x)$.
$\min\limits_x f(x)$	Valor mínimo de $f(x)$.
$\arg\max\limits_x f(x)$	Valor de x que maximiza $f(x)$.
$\arg\min\limits_x f(x)$	Valor de x que minimiza $f(x)$.
$\max\limits_x f(x)$ sujeto a $y=g(x)$	Problema de maximización de $f(x)$ con restricciones.
$\min\limits_x f(x)$ sujeto a $y=g(x)$	Problema de minimización de $f(x)$ con restricciones.
$\hat{x}, \widehat{x}$	Valor estimado u óptimo de x.
$\mathcal{O}(M)$	Orden de la cantidad de operaciones requeridas.

Acrónimos[1]

ASR	Reconocimiento automático del habla.
HMM	Modelos ocultos de Markov.
ANN	Red neuronal artificial.
SOM	Mapa auto-organizativo.
SVMs	Máquinas de soporte vectorial.
TDNN	Red neuronal con retardos temporales.
RP	Reconocimiento de palabras (%).
PP	Precisión en reconocimiento de palabras (%).
RF	Frases reconocidas completas (%).
TRN,TST	Entrenamiento, prueba.
SNR	Relación señal–ruido.
DAT	Cinta de audio digital.
MSE	Error cuadrático medio.
RT	Teoría de la regularización.
POF	Filtrado óptimo probabilístico.
MAP	Probabilidad a posteriori máxima.
LTI	Lineales invariantes en el tiempo (Sistemas).
AR	Auto-regresivo (modelo o sistema lineal discreto).
CC	Cepstrum complejo.
RC	Cepstrum real.
MFCC	Coeficientes cesptrales en escala de mel.
TC	Transformada coseno.
FT	Transformada de Fourier.
DFT	Transformada discreta de Fourier.
FFT	Transformada rápida de Fourier.
STFT	Transformada de Fourier de corta duración.
STDFT	Transformada discreta de Fourier de corta duración.
LPC	Coeficientes de predicción lineal.
PLP	Análisis predictivo lineal perceptual.
RASTA-PLP	Transformación espectral relativa PLP.
WVD	Distribución de Wigner-Ville.
MRA	Análisis multi-resolución.
CWT	Transformada ondita continua.
DWT	Transformada ondita discreta.
DDWT	Transformada ondita discreta diádica.
FWT	Transformada rápida ondita.
SCWT	Transformada ondita continua muestreada.
SWT	Transformada ondita semicontinua.
WPT	Transformada paquetes de onditas.
FWPT	Transformada rápida de paquetes de onditas.
POWPT	Transformada paquetes de onditas orientadas perceptualmente.
LCB	Bases de cosenos locales.
CPT	Transformada de paquetes de cosenos.

[1]Los acrónimos corresponden en general a la versión inglesa utilizada en la bibliografía del área.

MI	Información mutua.
MDL	Descripción de mínima longitud.
PCA	Análisis de componentes principales.
FA	Análisis de factores.
PP	Búsqueda de proyecciones.
ICA	Análisis de componentes independientes.
NOCICA	ICA sobrecompleto y con ruido.
LP-ICA	ICA por predicción lineal.
SVD	Descomposición en valores singulares.
LDB	Base discriminante local.
LSDB	Base menos dependiente estadísticamente.
BSB	Base de mejor dispersión.
BOB	Mejor base ortogonal.
BP	Búsqueda de bases.
MP	Búsqueda por coincidencia.
MOF	Método de marcos.
MF	Filtro de coincidencia.
GBP	Búsqueda de bases generalizada.
BPD	Limpieza de ruido por búsqueda de bases.
WDN	Limpieza de ruido mediante onditas.
HDN	Limpieza de ruido heurística.
FC	Frecuencia característica.
VOT	Tiempo de ataque de la sonoridad.
PEA	Potenciales evocados auditivos.
AI	Corteza auditiva primaria.
AII	Corteza auditiva secundaria.
TORC	Combinaciones de ondas móviles temporalmente ortogonales.
STRF	Campos receptivos espectro-temporales.

Prefacio

La emulación de la comunicación humana en forma artificial ha sido una meta largamente perseguida. Alcanzarla permitiría interactuar con las máquinas de una manera más sencilla y completamente distinta a la actual. A pesar de la aparente sencillez con la que los humanos realizan esta tarea, y de más de tres décadas de investigación, su resolución completa por medios artificiales ha permanecido fuera del alcance. Este trabajo se presenta como requisito para obtener el doctorado en ingeniería y constituye un aporte en la dirección de esta meta. Mediante un enfoque interdisciplinario se explora la idea acerca de cómo definir y buscar una representación "óptima" de la señal de voz mediante técnicas no convencionales a los fines de su posterior manipulación mediante un sistema artificial.

Para llegar al punto actual se ha recorrido un camino de varios años. Durante la tesis de grado se encaró el estudio inicial de los sistemas de *reconocimiento automático del habla* [168]. Se diseñó y probó un sistema que utilizaba la *transformada de Fourier* para la etapa de procesamiento, *redes neuronales* para la acústico-fonética y *sistemas expertos* para el análisis del lenguaje. Sin embargo, no se integraron completamente las etapas, y el enfoque tenía problemas para ser escalado a situaciones más complejas que la planteada originalmente.

Posteriormente se utilizó un enfoque más biológico para la etapa de procesamiento basado en *modelos de oído*, y se inició también el estudio de la aplicación de los *modelos ocultos de Markov* al problema del reconocimiento [166]. Aquí comenzó a surgir la idea de comparar las técnicas clásicas, basadas generalmente en estimadores espectrales, con un análisis más similar al que realiza el sistema auditivo.

En la tesis de Maestría se inició la comparación entre el análisis basado en la denominada *transformada ondita* y el basado en la clásica transformada de Fourier [167]. Se exploró principalmente el caso de la *transformada ondita discreta diádica*, en contraste

con la *transformada de Fourier de corta duración*. A pesar de sus atractivas cualidades para el procesamiento de señales transitorias, la comparación resultó desfavorable para este tipo particular de transformación basada en onditas. Allí se advirtió la multiplicidad de facetas que presentaban las representaciones relacionadas con onditas, y la dificultad de brindar una respuesta definitiva acerca de cuál resultaba mejor en este contexto.

Comenzando el trabajo para la tesis de doctorado se continuó la exploración de la aplicación de onditas, tratando de superar algunos de los problemas detectados previamente, resultando en mejoras significativas [55, 202, 170, 200, 172, 201, 189]. A pesar de que se abordaron también otros caminos el trabajo está sin duda influenciado por las ideas de la teoría de onditas, por lo que se le dedicará una parte importante del mismo. Dentro de los caminos alternativos explorados aparecen principalmente las representaciones ralas y/o independientes [171, 174, 173], y la inclusión de información adicional en la representación como la relacionada con los cambios de dinámica del aparato fonador [175]. La idea de lograr una representación "óptima" de la señal de voz ha guiado el trabajo desde el principio, motivo por el cual el análisis de señales ocupa también una porción importante.

Se puede decir que este trabajo aporta una visión alternativa al problema de análisis y representación de una señal partiendo de un *modelo* más complejo que el tradicional. En este sentido sigue una tendencia presente en la ciencia moderna que implica prácticamente un cambio de paradigma. Se ha cambiado el enfoque de la *simplicidad*, que intentaba analizar la realidad a partir de su descomposición en fenómenos sencillos, por el de la *complejidad* que intenta incluir muchos más aspectos de la realidad desde la concepción inicial del modelo [139].

Organización de la tesis

El documento de la tesis se ha organizado de la siguiente forma. En el Capítulo 1 se presenta la introducción al trabajo, destacando la motivación inicial y los conceptos fundamentales que guían el desarrollo de la tesis. Estas ideas se ilustran con algunas notas históricas relevantes.

Una revisión de los antecedentes y el marco conceptual de varias de las técnicas utilizadas se desarrolla en el Capítulo 2, junto con las diferentes formas de medir la

eficacia de una representación.

Las bases fisiológicas de la comunicación humana se presentan en el Capítulo 3. Se exponen aquí las características principales de la señal de voz, del aparato fonador y del sistema auditivo. El énfasis está puesto en los aspectos neurosensoriales de este último que permiten identificar las pistas acústicas del habla.

En el Capítulo 4 se exponen los fundamentos matemáticos del *análisis de señales*. Se discuten las diferentes formas de representación de una señal, desde las tradicionales hasta las más recientes o *no convencionales*, incluyendo aquellas más orientadas al caso de la señal de voz. El enfoque está orientado principalmente a señales de tiempo continuo.

En el Capítulo 5 se desarrollan diferentes alternativas para obtener la representación de señales de tiempo discreto a partir de *diccionarios*, destacando los casos ortogonales basados en Fourier, onditas, paquetes de onditas y paquetes de cosenos.

Los fundamentos de las *representaciones ralas y/o independientes* obtenidas a partir de diccionarios se presentan en el Capítulo 6, mencionando sus ventajas y desventajas, las conexiones con otros enfoques y algunos ejemplos de aplicación.

En el Capítulo 7 se desarrollan diferentes métodos basados en técnicas no convencionales y se estudia su aplicación al análisis y la representación de la señal de voz. Se comparan las alternativas propuestas con las clásicas, a partir de aspectos cualitativos, medidas de eficacia y resultados de aplicación a problemas de clasificación de fonemas, reconocimiento del habla y limpieza de ruido.

Finalmente en el capítulo 8 se discuten las conclusiones generales y particulares, junto con los posibles trabajos futuros. También se agregan los Apéndices A y B que presentan los aspectos referentes a la implementación práctica de los experimentos desarrollados y los datos utilizados.

Aspectos institucionales

El desarrollo de este trabajo se apoyó en la infraestructura y los recursos aportados por diferentes instituciones. Entre estas instituciones se pueden contar a la Facultad de Ingeniería de la UNER, particularmente el *Laboratorio de Cibernética*. Aquí se realizó gran parte de las actividades con recursos provenientes de subsidios de la UNER,

del CONICET y de la ANPCyT[2]. De acuerdo a lo convenido originalmente, los cursos y una parte importante del trabajo se han llevado a cabo en el *Instituto de Ingeniería Biomédica* de la Facultad de Ingeniería de la UBA, bajo la dirección del Ing. Luis Rocha[3]. Finalmente se han realizado varias estancias en el *Laboratorio de Voz* de la Universidad Autónoma Metropolitana Iztapalapa, en la ciudad de México, bajo la supervisión del codirector de la tesis el Dr. John Goddard[4]. El Dr. Goddard ha realizado también visitas a la FI-UNER y a la FI-UBA.

[2]Proyectos I+D UNER: #6036 "Sistema de reconocimiento automático del habla" y #6062 "Desarrollo de un laboratorio de voz". UNER: Beca para cursado de carreras de 4º Nivel, Res. CD Nº 233/97. CONICET: Beca interna de perfeccionamiento en la Investigación. Proyecto ANPCyT: PICT cod. 11-12700, tipo A "Técnicas no convencionales aplicadas a la reducción de ruido en audífonos digitales".

[3]Varios viajes fueron financiados también mediante fondos de FIUBA, Expte Nº 965.271/98.

[4]Financiado mediante fondos CONACYT (México), Proyecto #31929-A: "Análisis de la señal de voz en español", y mediante fondos provenientes de convenios específicos de colaboración UAM-UNER.

Capítulo 1

Introducción

"Al principio era el Verbo, y el Verbo estaba en Dios y el Verbo era Dios. El estaba al principio en Dios. Todas las cosas fueron hechas por Él, y sin Él no se hizo nada de cuanto ha sido hecho."

(Juan 1,1)

Contenido

1.1. Motivación

EL análisis de *señales* ha cobrado una importancia superlativa dentro de la teoría de las comunicaciones, y casi en todas las aplicaciones actuales de la misma. Estas aplicaciones tecnológicas han invadido prácticamente nuestra vida diaria. Los conceptos de *señal*, *sistema* e *información* están íntimamente relacionados. Las señales transportan información del sistema que las produjo, contenida o codificada en un patrón de variaciones de alguna magnitud física. Desde el punto de vista matemático las señales pueden tratarse como funciones. Esto permite su manipulación y transformación de manera precisa dentro del marco de las teorías correspondientes.

La señal de voz, objeto de la investigación desarrollada en esta tesis, es una de las señales "naturales" más estudiadas. En el campo del modelado y análisis de la señal de voz se han hecho importantes avances como los *coeficientes de predicción lineal* (LPC) [160], o los *coeficientes cepstrales en escala de mel* (MFCC) [37]. Sin embargo existen

todavía problemas que no han podido resolverse satisfactoriamente. Esta señal es producida por el *aparato fonador* humano a través de un complicado mecanismo en el que intervienen varios órganos. Los mismos son comandados por el cerebro para modificar las propiedades acústicas del tracto vocal y de los estímulos sonoros implicados. De esta manera se producen patrones característicos de variación de la presión sonora que constituyen la base de la comunicación humana. La señal así generada contiene información codificada a diferentes niveles, incluyendo aspectos fonéticos, léxicos, prosódicos y gramaticales, entre otros. El sistema auditivo logra descifrar el mensaje "escondido" en estos patrones de variación sonora producidos por el aparato fonador de una manera asombrosamente robusta y eficaz, en forma casi independiente de factores como la identidad del hablante o el ruido de fondo.

A pesar de los avances realizados, los dispositivos artificiales que han tratado de emular estos aspectos distan mucho de poseer actualmente capacidades similares. El desempeño de estos sistemas se degrada enormemente con el ruido de fondo, variaciones entre diferentes hablantes, e inclusive cambios en la forma de hablar de un único sujeto (como puede ser en la velocidad de pronunciación). Por ejemplo, en la Figura 1.1 se presenta el porcentaje de reconocimiento de palabras para un sistema de *reconocimiento automático del habla* (ASR) del estado del arte frente a distintos tipos y cantidades de ruido aditivo. Este sistema y está basado en *modelos ocultos de Markov* (HMM) y utiliza como representación de la señal de voz a los MFCC. La degradación resulta ampliamente superior a la experimentada con sujetos normo-oyentes en condiciones similares[1]. Para tener una idea cualitativa de lo que significa esta degradación en la Tabla 1.1 se muestra como se traducen estos porcentajes de reconocimiento, para las distintas condiciones, en términos del resultado final de reconocimiento de una oración determinada. Aquí resulta claro que en varias circunstancias, inclusive para niveles de ruido relativamente bajos para los humanos, se pierde el sentido de las frases total o parcialmente.

Generalmente los errores se deben a que los modelos utilizados para el análisis y reconocimiento se mantienen lo más sencillos posible a costa de sacrificar su ajuste a las situaciones reales. La mayoría de las soluciones encaradas recientemente implican algún tipo de "parche" a estos modelos de manera que contemplen las situaciones no previstas originalmente, aunque sea parcialmente. En este sentido se puede decir que, en cuanto a su desempeño, los sistemas actuales basados en las técnicas convencionales han llegado a una especie de "mínimo local", lo cual obliga a repensar algunas de sus bases y fundamentos [13]. Por lo tanto hay todavía un camino importante por recorrer para poder solucionar los problemas planteados. Ésto conduce a la necesidad de explorar otras alternativas, como podrían ser el uso de *técnicas no convencionales*[2]. Hacen falta nuevas ideas en el campo, o bien repensar algunas abandonadas tal vez prematuramente. Es necesario también revisar algunos detalles acerca de cómo los humanos analizan y re-

[1] En sujetos normo-oyentes expuestos a este material, explicitando previamente las características particulares del mismo, el reconocimiento permanece prácticamente inalterado en este rango de niveles de ruido y para ambos tipos de ruido.

[2] El sentido que se le dará en este trabajo a la expresión "técnica no convencional" es el de aquellas técnicas que no se han utilizado de manera habitual para el análisis y representación de la señal de voz con fines a su posterior reconocimiento o clasificación.

Tabla 1.1: *Comparación entre las oraciones reconocidas y las tasas de reconocimiento para la oración "acgp3104" correspondiente a la base de datos de habla española Albayzin [17] y el mismo sistema de reconocimiento utilizado para confeccionar la Figura 1.1. A pesar de que la inclusión de información léxico-gramatical vuelve al sistema bastante más robusto, es posible observar que con cantidades relativamente pequeñas de ruido la frase reconocida pierde total o parcialmente el sentido. Para más detalles referirse al Apéndice B.*

Ruido	SNR	Resultado	RP (%)	PP (%)	RF (%)
-	∞	"dime el caudal de todos los rios que desemboquen en el mar mediterraneo"	92.54	92.20	58.82
Blanco	50	"dime el caudal de todos los rios que desembocan en el mar mediterraneo"	92.71	92.46	62.18
	25	"dime el caudal de todos los rios que desembocan en el mar mediterraneo"	90.53	88.94	44.54
	15	"tajo dime el caudal de todos los rios que desembocan en el mar mediterraneo tajo"	77.54	67.64	3.36
	10	"rio el caudal del rio mas rios que desembocan en el mar en el mar"	48.78	43.25	2.52
	5	"dime el caudal del rio mas rios desembocan en el mar menor"	28.50	21.46	0.84
	0	"dime el caudal el nombre de menor en el mar menor"	16.31	8.88	0.00
Murmullo	50	"dime el caudal de todos los rios que desemboquen en el mar mediterraneo"	92.29	91.79	59.66
	25	"dime el caudal de todos los rios que desemboquen en el mar mediterraneo"	91.45	90.19	49.58
	15	"dime el caudal de todos los rios que desemboquen en el mar mediterraneo mar"	79.63	70.58	2.52
	10	"dime el caudal de todos los rios que desembocan en el mar mediterraneo mar"	65.47	55.83	2.52
	5	"mar en el caudal de los rios desembocan en el mar en el mar"	42.08	24.73	0.00
	0	"dime el caudal del mar en el rio desemboca el mar menor de menor"	27.72	7.31	0.00
Original	-	"dime el caudal de todos los rios que desembocan en el mar mediterraneo"	-	-	-

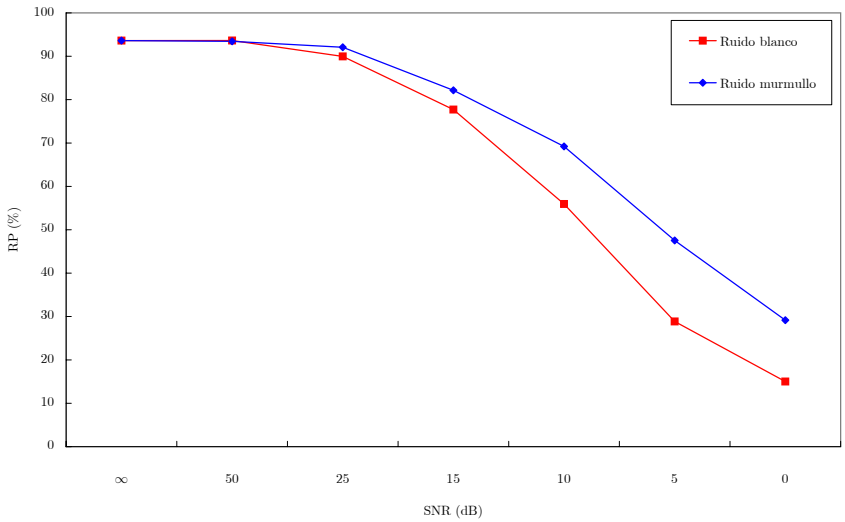

Figura 1.1: *Porcentaje de reconocimiento de palabras (RP) para un sistema del estado del arte (MFCC+HMM) para distintos tipos (blanco y murmullo) y cantidades de ruido (SNR). La porción de la base de datos utilizada para entrenar y probar el sistema contiene material léxica y gramaticalmente restringido, que corresponde a una serie de preguntas acerca de la geografía de España. Este "conocimiento" se ha integrado al reconocedor de manera explícita mediante un modelo del lenguaje adecuado y aumenta la robustez del sistema. Para más detalles referirse al Apéndice B.*

conocen el habla que pueden haberse dejado de lado, sobre todo en aspectos relacionados con la eficiencia y robustez del sistema auditivo [119, 64].

En muchos casos el avance en la Matemática o en sus aplicaciones tecnológicas ha sido influenciado por los conocimientos de los sistemas naturales equivalentes. Anteriormente algunos tipos de procesamiento y análisis de la voz que incluían aspectos *biológicamente inspirados* han resultado provechosos para mejorar el funcionamiento de los sistemas artificiales. Este es el caso de los MFCC ya mencionados, que fueron concebidos incorporando conocimiento acerca del proceso de generación de la señal de voz y algunas características psicoacústicas del sistema auditivo. Otro ejemplo podría ser la estrategia denominada *análisis predictivo lineal perceptual* (PLP) [74, 76]. Sin embargo existen todavía aspectos que, ya sea por cuestiones de simplificación, o por desconocer su función e importancia en el proceso de comunicación humana, no se han incorporado en los enfoques convencionales. Por ejemplo, algunas de las simplificaciones generalmente empleadas provienen de considerar sistemas lineales, estacionarios por tramos y con estadística significativa de segundo orden. En general se trata también de disminuir lo más posible la cantidad de dimensiones de análisis. Otra práctica muy difundida consiste en utilizar bases de funciones ortogonales "sencillas" para realizar el análisis de la señal.

Se sabe que los sistemas biológicos en general no cumplen con las restricciones antes mencionadas y sobrepasan ampliamente las capacidades de los sistemas artificiales. Las personas realizan complicados análisis de señales a través de los sistemas neurosensoriales y extraen información útil acerca de su entorno, en una gran variedad de situaciones adversas y en forma prácticamente "transparente" para ellas.

La comprensión de estos procesos ha ido mejorando paulatinamente junto con el desarrollo simultáneo de las ciencias biológicas y de las exactas. Se han comprendido mejor varios principios de funcionamiento de los sistemas naturales a partir de nuevos desarrollos de las teorías matemáticas existentes. De hecho, a medida que los conocimientos acerca de los sistemas sensoriales biológicos han ido aumentando, se han ido adaptando los "modelos" subyacentes en el análisis de señales (y viceversa)[3]. Por ejemplo, inicialmente se dijo que estos sistemas sensoriales se comportaban como si realizaran un *análisis de Fourier* de las señales provenientes de su entorno [62]. Posteriormente, en los 80's, se indicó que en realidad podía considerarse como si realizaran un análisis similar al de *Gabor* [212], y más recientemente que era más propio considerarlo como un análisis basado en *onditas* [42]. En este mismo sentido, se ha demostrado últimamente que las especiales características de los sistemas sensoriales obedecen a una adaptación para procesar de manera óptima los estímulos de su entorno [145]. Los principios detrás de esta codificación están todavía siendo estudiados, pero ya existen numerosas pistas que permiten orientar la búsqueda en esta dirección. Uno de estos principios consiste en lograr codificar cada una de las señales implicadas en el problema en términos de sólo unas pocas características significativas. Ésto es lo que se denomina una *representación rala*. Su utilización como esquema de codificación eficiente se ha demostrado a nivel de los sistemas sensoriales biológicos. Se puede decir que el propio código neural –basado en trenes de pulsos– también es ralo. Existen varios trabajos que estudian la aplicación de este principio como "modelo" para las representaciones generadas en los campos receptivos de la corteza visual, y más recientemente también en la corteza auditiva [146, 114].

El curso seguido en este trabajo pretende plantear algunas alternativas para acercarse un poco más al desempeño de los sistemas sensoriales naturales. La meta consiste en explorar las bases para el diseño de un sistema que permita extraer, de manera robusta y eficiente, las componentes más significativas de la señal del habla, a los fines de su posterior clasificación o reconocimiento.

Este capítulo se organiza de la siguiente forma. En la Sección 1.2 se presentan los objetivos generales de la tesis. En lo que resta del Capítulo se presenta un panorama general de las diferentes técnicas disponibles para el análisis y representación de la señal de voz que incluye algunas notas de su evolución histórica. Las técnicas convencionales son las primeras en aparecer en la Sección 1.3. El desarrollo comienza con los fundamentos matemáticos del análisis de señales clásico y concluye con las técnicas convencionales específicamente desarrolladas para el habla. A continuación, en la Sección 1.4, se introducen las ideas y fundamentos de las técnicas más recientes o no convencionales. Se concluye con algunos comentarios de cierre del Capítulo en la Sección 1.5.

1.2. Objetivos

Son objetivos de esta tesis:

- Avanzar en la comprensión del proceso de la comunicación oral humana.

[3]Ésto ha culminado recientemente en la acuñación, para el área de aplicaciones informáticas, del término *computación biológicamente inspirada* (en inglés Biologically-Inspired Computing).

- Comprender los procesos implicados en la decodificación del mensaje contenido en la señal de voz, especialmente a nivel neurosensorial.
- Caracterizar estos procesos desde un punto de vista matemático mediante elementos de la teoría de análisis de señales.
- Desarrollar métodos y/o algoritmos de procesamiento de señales que permitan extraer las características significativas de la señal de voz, de manera robusta y eficiente, a los fines de su posterior reconocimiento.
- Validar los métodos desarrollados mediante experimentos con datos artificiales y reales.
- Interpretar los resultados desde una perspectiva de análisis, modelado e identificación de sistemas.

1.3. Técnicas convencionales

En esta sección se presentarán aquellas técnicas que podrían agruparse dentro de las técnicas convencionales o clásicas para el análisis y representación de la señal de voz. No existe un límite de separación clara entre las técnicas convencionales y las que no lo son. Ésto se debe a que los cambios de paradigmas generalmente comienzan a manifestarse de manera gradual a partir de la aparición de nuevas ideas y de su paulatina aplicación. Es recién cuando la técnica se "estabiliza", en cuanto a sus bases y utilización para resolver determinados problemas, cuando puede considerarse como convencional. Sin embargo muchas veces el rango de aplicaciones puede ser muy variado, por lo cual una técnica nueva no necesariamente reemplaza completamente a las anteriores.

Para comenzar con esta Sección es necesario presentar algunas definiciones básicas del análisis de señales, cuyos fundamentos se remontan a las ideas presentadas por Newton y Fourier hace ya varios siglos [41].

1.3.1. Análisis de señales

La palabra análisis proviene de la base griega *analyo* que significa "desatar". Es posible definirla como: "distinción y separación de las partes de un todo hasta llegar a conocer los principios o elementos de éste"[4]. El análisis de una señal consiste en aislar aquellas componentes que poseen una forma compleja para tratar de comprender mejor su naturaleza u origen. En este contexto se designa como *ruido* a cualquier fenómeno que perturba la percepción o interpretación de una señal. Comparte la misma denominación que los efectos acústicos análogos y siempre está presente en la obtención de cualquier señal real que intentemos analizar. Se puede decir entonces que analizar una señal consiste en encontrar y desagregar aquellas partes características o *componentes ocultas* que mejor permitan describirlas, minimizando simultáneamente los efectos del

[4]Diccionario General de la Lengua Española Vox

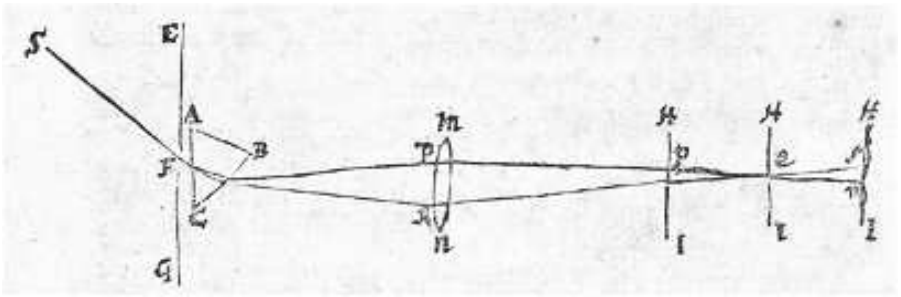

Figura 1.2: *Esquema del telescopio reflector de Newton (reproducido de su artículo [138]). Luego de observar la aberración cromática de su telescopio y de realizar el experimento de separación de las componentes individuales de la luz blanca, en el prisma que se muestra en el esquema, concluyó que ésta no era una única entidad. De aquí surge la idea de espectro, trasladada posteriormente al análisis de Fourier.*

ruido. Las aplicaciones últimas de la representación lograda en términos de estas componentes pueden ser: la simple inspección con fines de estudio, la compresión de la señal, la codificación o transmisión de la misma, o la alimentación de un sistema automático que tome decisiones en base a ella, entre otras.

El análisis de fenómenos físicos posee elementos análogos, debido a que las señales constituyen manifestaciones del mundo físico. Su aparición es bastante anterior a este siglo, casi con el comienzo de la ciencia, y de hecho sentó las bases para el desarrollo de las teorías que hoy sustentan el análisis de señales. En este sentido se puede citar como ejemplo "cercano" el análisis de la luz visible mediante un prisma, que permite descomponerla en sus componentes fundamentales. Estas componentes están "ocultas" en la luz blanca y se manifiestan en su interacción con los objetos del mundo físico. Este fenómeno fue descubierto y estudiado por Newton como uno de sus primeros aportes a la óptica en 1670. Newton diseñó y construyó el primer telescopio reflector (ver Figura 1.2) y concluyó que la luz blanca no era una única entidad después de observar la aberración cromática de su telescopio y de realizar el experimento del prisma en donde pudo observar el espectro (término que proviene de *spectrum*, o fantasma) de los componentes individuales de la luz blanca y recomponerlo con un segundo prisma.

1.3.2. Análisis de Fourier

Aunque Newton no reconoció el concepto de *frecuencia*, el parecido de este espectro con el de Fourier no es casual. Fourier conocía los trabajos de Newton y desarrolló las bases de su análisis cuando estudiaba la conducción del calor en los cuerpos sólidos. En 1807 Fourier difundió el primer esbozo de su *Teoría analítica del calor* [52], en la cual demostró que la conducción del calor en los cuerpos sólidos se podía expresar como una suma infinita de términos trigonométricos cada vez más pequeños. Estos términos constituían las "componentes ocultas" que había podido descubrir en este fenómeno. De esta forma Fourier había encontrado el "prisma" adecuado para analizar a los fenómenos de conducción del calor y como resultado había desarrollado la teoría del análisis armónico

para descomponer funciones periódicas arbitrarias en términos de funciones sinusoidales. A pesar del tremendo aporte que constituiría su teoría, ésta fue fuertemente criticada por notables matemáticos de la época como por ejemplo Laplace.

Debido a la complejidad y variabilidad del habla las ideas del análisis de Fourier tardaron un poco en aplicarse con éxito a esta señal. Los primeros intentos fueron realizados con dispositivos mecánicos como los basados en cuerdas, resonadores o filtros. Estos dispositivos realizaban una descomposición de los sonidos análoga a la propuesta por Fourier mediante principios mecánicos. De hecho, pueden también establecerse ciertas analogías entre el análisis de Fourier y el funcionamiento de la cóclea dentro nuestro oído. Aunque este aspecto se revisará posteriormente con más detalle basta con decir por ahora que la membrana basilar constituye un complejo analizador espectral. También se utilizaron dispositivos de tipo estroboscópico. Con el advenimiento de los medios electrónicos "modernos" comenzaron a publicarse algunos trabajos que intentaron evidenciar las características y componentes fundamentales de esta señal. Como ejemplo se puede citar [206], donde con dispositivos oscilográficos bastante sencillos y un banco de filtros analógicos dispuestos en octavas se logró obtener los resultados de la Figura 1.3 para una vocal del inglés. Por la naturaleza cuasiperiódica de las vocales pronunciadas en forma aislada, son las que más fácilmente se ajustan a un análisis de este tipo.

1.3.3. Bases y transformaciones lineales

La teoría de Fourier para la descripción de funciones sería generalizada posteriormente mediante la idea de *base ortogonal*, es decir un conjunto de funciones con condiciones particulares que permiten descomponer a funciones más generales en términos de ellas (las primeras). Se puede decir, de manera corriente, que las condiciones particulares que debían cumplir el conjunto de funciones de la base eran las siguientes:

1. No presentar "componentes comunes" entre ellas,
2. Ser suficientes como para "generar" cualquier función (de un conjunto determinado).

Se encontraron muchas bases que permitían descomponer funciones de diferentes formas. Algunas de estas bases eran más adecuadas que otras para describir algunos tipos de funciones. Ésto es porque las funciones o elementos de la base estaban precisamente relacionados con las componentes importantes que se pretendía extraer de las señales. Mediante estas bases se podían realizar transformaciones lineales que permitían llevar las señales a espacios alternativos donde las características significativas se manifestaban de forma más evidente.

Lamentablemente la base de funciones exponenciales de Fourier poseía importantes limitaciones para la descripción de señales intrínsecamente transitorias, como por ejemplo las consonantes del habla. Estas exponenciales complejas resultan ser las autofunciones derivadas de los sistemas *lineales invariantes en el tiempo* (LTI) y de allí que constituyan un buen "prisma" para analizar señales derivadas de estos sistemas. Sin embargo al no presentar ningún cambio abrupto durante su evolución, tampoco pueden representar

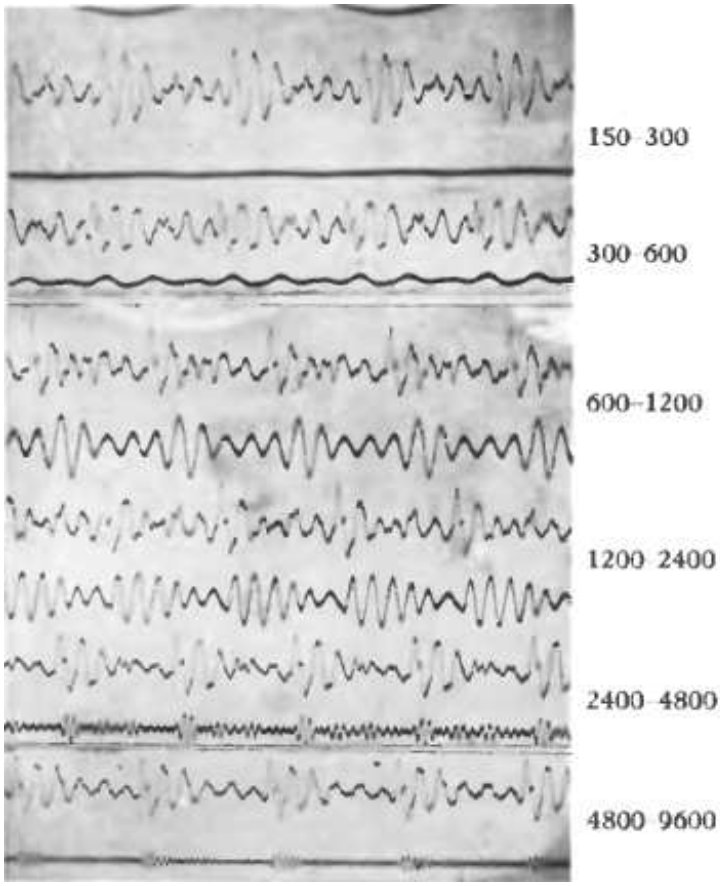

Figura 1.3: *Descomposición de la /a/ (como en* father*) en oscilagramas obtenidos por medio de filtros dispuestos en octavas (reproducido de [206]). A la derecha se detalla el ancho de banda en Hz de cada octava. Obsérvese la apariencia cuasiperiódica de los oscilogramas.*

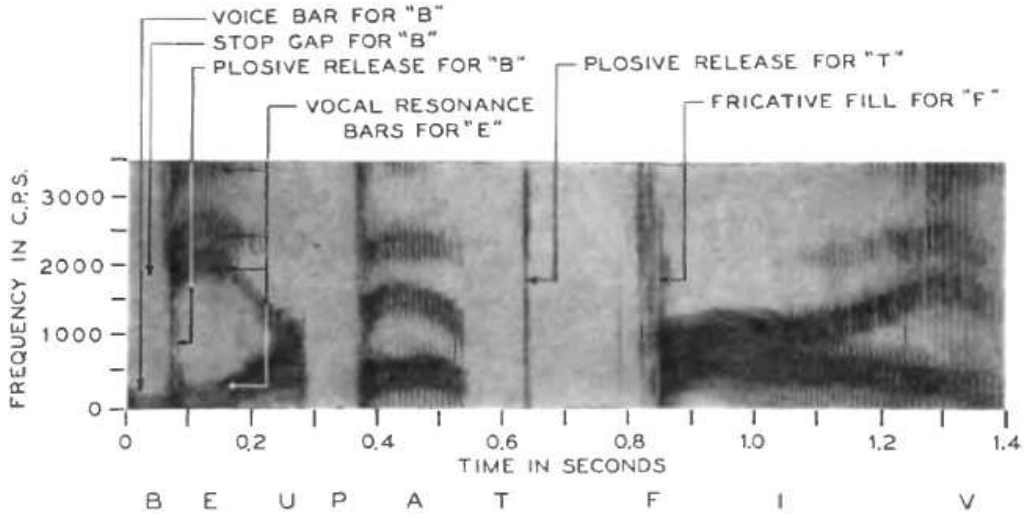

Figura 1.4: *Espectrograma de la frase inglesa "Be up at five" que permite* ver *los rasgos distintivos de la misma (reproducido de [192]). En la parte superior se han remarcado diferentes eventos acústicos relevantes. En esta representación es posible observar claramente la evolución de las formantes principales de la señal de voz.*

adecuadamente dichos cambios. Ésto se manifiesta en el conocido fenómeno de Gibbs donde se requiere un número infinito de componentes para aproximar una discontinuidad, lo que por supuesto no constituye una representación "eficiente".

1.3.4. Análisis de Gabor

Debido a estas limitaciones, en 1946 Gabor extiende las ideas de Fourier con el concepto de *frecuencia local* para el estudio de sonidos musicales [54]. De esta forma propone una nueva base que rescata características tanto en la frecuencia como en el tiempo. La nueva base es idéntica a la de Fourier, salvo por el hecho de la aparición de la modulación con una ventana gausiana deslizante en el tiempo, que da la necesaria localización temporal. La aplicación de otras ventanas temporales da lugar a la *transformada de Fourier de corta duración* (STFT). El concepto involucrado consiste ahora en la representación de las señales en un *plano tiempo-frecuencia* $(t - f)$ a partir de *átomos* o componentes elementales con mínima dispersión en este plano. Ésto también sienta las bases matemáticas detrás de los dispositivos analógicos utilizados para el estudio del habla durante esa misma época. Dichos dispositivos seguían utilizando bancos de filtros para realizar el análisis, pero se habían perfeccionado en la manera de realizar registros y presentar los resultados. Así surgen los espectrogramas, que permiten comenzar a "ver" los rasgos distintivos del habla a partir de la distribución de su energía en el plano $t - f$ [192]. Además, ésto permite también re-intrepretar los modelos de funcionamiento de la cóclea, incorporando los aspectos temporales anteriormente ignorados.

Sin embargo existe una limitación física que no permite medir con precisión arbitraria los eventos ocurridos en el plano tiempo-frecuencia. Cuanto mayor es la localización o

resolución en el tiempo, menor lo es en la frecuencia y viceversa. Ésto se conoce como *principio de incertidumbre* y da lugar a la utilización de dos tipos diferentes de espectrogramas para el estudio del habla. De acuerdo al ancho de banda de los filtros equivalentes se tienen los espectrogramas de banda angosta y los de banda ancha. En los espectrogramas de banda angosta la ventana temporal es relativamente ancha, 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 las denominadas *formantes*. En los espectrogramas de banda ancha la situación es exactamente la inversa y permiten extraer, en forma bastante sencilla, parámetros tales como el período de entonación.

La aparición de las computadoras y la tecnología digital permite volver a aplicar el análisis de Fourier a la señal del habla. La digitalización de las señales de sonido permite "introducirlas" en la computadora para realizar cálculos con ellas. Un problema inicial era que los cálculos para obtener la *transformada discreta de Fourier* (DFT) de una señal como ésta demandaban mucho tiempo. En 1965 Cooley y Tukey publican un trabajo en el cual proponen un algoritmo para realizar el cálculo rápido de la DFT mediante computadoras [29]. Ésto da lugar al resurgimiento de los estudios basados en espectrogramas para aprovechar la flexibilidad y potencialidades de esta nueva herramienta.

Las ideas de la transformada de Gabor se generalizan posteriormente en lo que se denomina *análisis por tramos* o ventanas, que permite "convertir" en local un análisis pensado para señales estacionarias [37].

1.3.5. Distribuciones tiempo-frecuencia

En 1948 se introduce la *distribución de Wigner-Ville* (WVD) como otra herramienta para analizar estructuras en el plano $t - f$ [127]. El enfoque utilizado para el cálculo de esta distribución es diferente al basado en representaciones atómicas ya presentadas. Ésto se debe a que en este caso la señal se compara con ella misma –modulada en la frecuencia y corrida en el tiempo– y no con una base de funciones. El resultado es la representación de la señal en términos de la *distribución tiempo-frecuencia* de su energía.

La DWV posee propiedades matemáticas interesantes, y constituye la base de las distribuciones $t-f$ cuadráticas. Estas distribuciones resultan en una teoría unificadora que aporta posibilidades adicionales para poder "resolver" los eventos significativos adecuadamente en tiempo y frecuencia para el análisis de señales complejas. Se ha demostrado que otras distribuciones cuadráticas se pueden obtener a partir de la distribución de Wigner-Ville mediante convoluciones con núcleos bidimensionales adecuados [127].

1.3.6. Análisis específicos para el habla

Dentro de las técnicas convencionales se deben mencionar aquellos análisis concebidos específicamente para el caso de la señal de voz. Éstos se han desarrollado a partir del estudio de las características perceptuales del oído o de un modelo de producción del habla y comenzaron a aplicarse a partir de la década de los 70's.

Por ejemplo, la técnica de LPC se basa en suponer un modelo lineal de tiempo discreto para la señal de voz [159]. Por otra parte los MFCC se sustentan en la supocisión de un modelo de producción del habla también lineal, junto con algunas propiedades adicionales, que permiten convertir una operación de convolución en una suma [37]. A ésto se agrega una escala psicoacústica que incluye aspectos perceptuales en la representación. Otro caso es el del PLP, introducido por Hermansky [74] para lograr una representación relativamente independiente del hablante, pero que conservara la información importante para la discriminación fonética. Además se han propuesto diferentes alternativas para lograr *representaciones auditivas tempranas* de la voz mediante modelos de oído, que contemplan los aspectos más sobresalientes del procesamiento hasta el nivel del nervio auditivo [36].

1.4. Técnicas no convencionales

A pesar de la sencillez y elegancia matemática de las bases de Fourier, las bases ortogonales en general y las técnicas basadas en modelos lineales, se ha visto que existen razones importantes para considerar ideas aún más generales.

Las restricciones impuestas a los elementos de las bases tradicionales son bastante artificiales y difíciles de explicar en términos de los sistemas sensoriales biológicos "equivalentes". La aparición de determinadas características, como por ejemplo cierta redundancia a nivel de los elementos de la "base", puede demostrarse útil en numerosas aplicaciones y está presente en los organismos vivos.

La idea de *marcos* y también la de *diccionarios* de funciones, que son diferentes generalizaciones de las bases ortogonales, aparecen principalmente de la mano de la *teoría de onditas*. Algunas características estadísticas de los coeficientes de las representaciones logradas tampoco son adecuadamente descriptas en términos de distribuciones gausianas o combinaciones de éstas. Por lo tanto se requiere la utilización de modelos que tengan en cuenta de manera explícita la estadística de orden superior. Estos modelos aparecen junto con el *análisis de componentes independientes* y las *representaciones ralas*.

Con estas ideas más generales, y con cierta inspiración biológica adicional, el desafío consiste en poder diseñar y/o construir sistemas de análisis y representación de señales que superen algunas de las restricciones actuales de los mismos.

1.4.1. Análisis basado en onditas

A pesar de los avances obtenidos mediante la STFT, las limitaciones del principio de incertidumbre seguían presentes. La opción del uso de la DWV poseía también algunos problemas prácticos importantes como el *artefacto* de los términos cruzados. Para intentar brindar una alternativa aparece a mediados de los años 80's la teoría de onditas [130]. La *transformada ondita continua* (CWT) fue "diseñada" especialmente para trabajar con señales transitorias o no estacionarias. La distribución de energía derivada de la CWT es el *escalograma*. Ésto da lugar a nuevas formas de descomposición que constituyen un "prisma" que analiza de diferente forma los eventos en el plano $t - f$.

Con la CWT se logra "seguir" con mayor precisión en frecuencia a los eventos más lentos y con mayor precisión en el tiempo a los eventos rápidos o cambios bruscos. Es en este sentido que resulta más parecido a lo realizado en el sistema auditivo (respecto al análisis de Gabor). Sin embargo la CWT no es la única forma de descomposición no uniforme posible.

De hecho es posible dividir el plano $t - f$ en *cuadrículas* de diferentes maneras (para el caso de representaciones discretas). En la Figura 1.5 se pueden ver las cuadrículas $t-f$ obtenidas a partir de diferentes diccionarios, junto con algunos de los átomos correspondientes. Para el caso de la *base de Dirac*, que constituye la base canónica para señales de tiempo discreto, se obtiene la mejor resolución temporal posible. El otro extremo corresponde a la base de Fourier, donde se logra la mejor resolución frecuencial. Para la STFT se muestran los casos de banda ancha y banda angosta ya descriptos. La transformada ondita en su versión *discreta* (DWT, diádica) particiona las frecuencias en octavas, con la mayor resolución frecuencial en las frecuencias bajas y mayor resolución temporal en las frecuencias altas. Ésto lo realiza en base al esquema denominado *análisis multiresolución*. La *transformada paquetes de onditas* (WPT) consiste en una generalización de este análisis multiresolución, lo que permite una mayor flexibilidad en la definición de una partición $t - f$. A pesar de esta flexibilidad es posible todavía relajar algunas restricciones implícitas para lograr una representación más general.

1.4.2. Representaciones ralas e independientes

Para extraer mediante el análisis aquellas características significativas de una señal, es necesario que la transformación pueda describir a cada una de las señales posibles en términos de un pequeño número de características del total disponible [145]. Ésto quiere decir que para que la representación de la señal resulte óptima es necesario lograr que sea suficientemente rala. Por ejemplo si se quiere describir una señal senoidal mediante la base de Fourier es posible hacerlo con un solo elemento. Ésto resulta particularmente eficiente en términos de *teoría de información* [71]. Sin embargo para señales más complejas, o que presenten alguna discontinuidad, la cantidad de coeficientes diferentes de cero puede aumentar enormemente (o volverse tan grande como la dimensión del espacio).

Se requiere entonces un diccionario que posea una cantidad importante de átomos diferentes, de manera tal de poder seleccionar sólo aquellos más relacionados con los rasgos distintivos de la señal a analizar, independientemente de la complejidad de ésta última. Este comportamiento de naturaleza "no lineal" deja de lado las restricciones de independencia lineal de las bases ortogonales. Podría decirse que resulta equivalente a utilizar una "base" adaptiva, o sea una base que se adapte a la señal analizada, eligiendo sólo los elementos que "mejor" la describen. Para armar este diccionario se pueden superponer diferentes diccionarios básicos o bien un conjunto de átomos individuales, que en su conjunto representen las diferentes características presentes en la señal bajo estudio. Otra característica deseable es que los coeficientes de la representación resulten estadísticamente independientes entre sí [71].

Con este enfoque es posible encontrar no sólo los coeficientes para lograr la representación, partiendo de un diccionario dado de antemano, sino inclusive el "diccionario

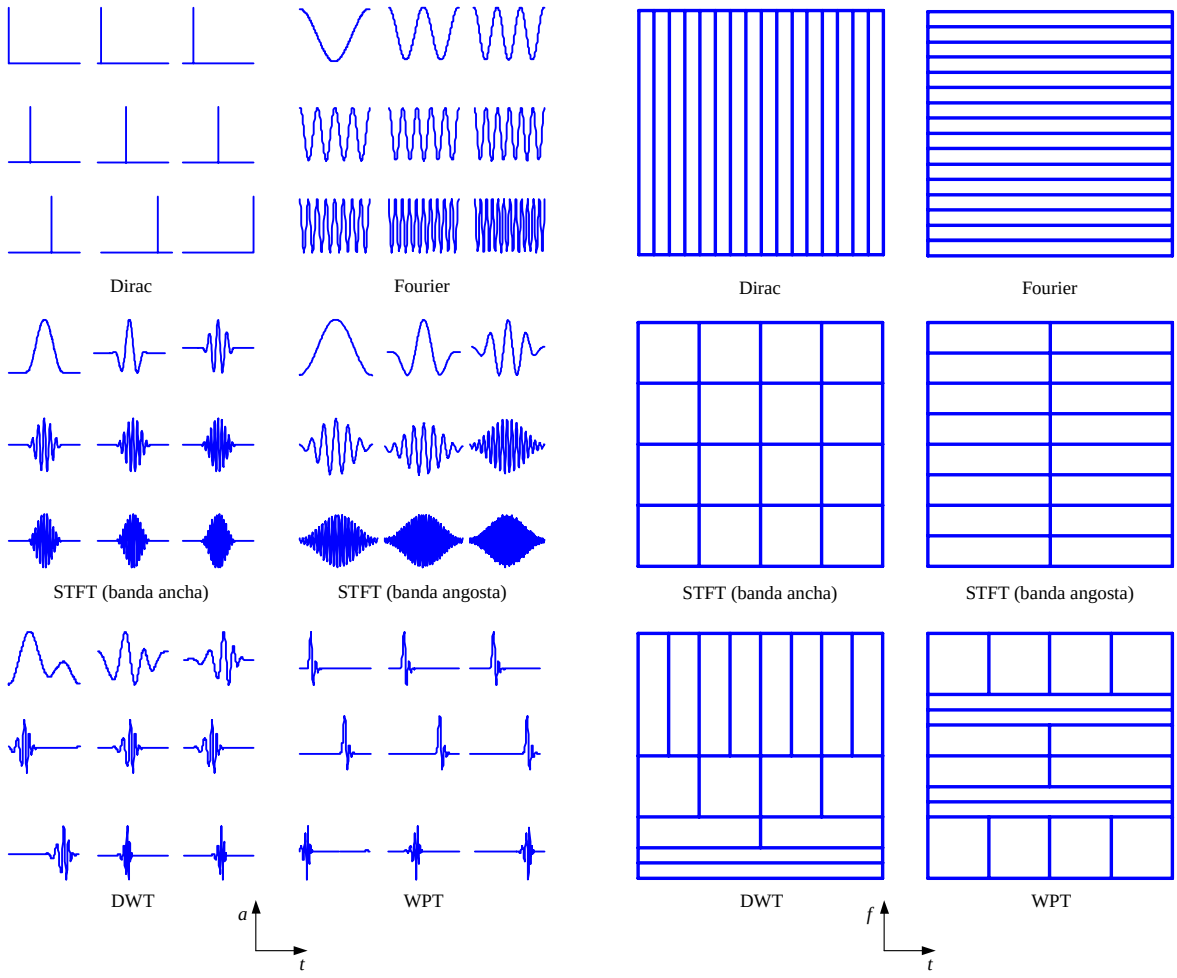

Figura 1.5: *Átomos de la base o diccionario (izquierda) y representación en el plano tiempo-frecuencia (derecha) para los distintos tipos de análisis. Para el caso de Fourier y STFT se muestra sólo la parte real. Tanto la DWT como la WPT están calculadas utilizando la ondita de Daubechies con 16 momentos nulos. Éstas son sólo algunas de las posibilidades para lograr una representación atómica. Nótese que en ninguno de los casos es posible evitar el principio de incertidumbre, pero si se puede poner más atención en los aspectos temporales o frecuenciales que más interesan para las diferentes zonas del plano, manteniendo constante el área de los átomos.*

óptimo" para cumplir con estas nuevas condiciones[5]. En la Figura 1.6 se pueden observar algunos átomos del diccionario óptimo obtenido a partir de datos del fonema */a/* mediante estas técnicas, junto con la disposición de algunos de ellos en el plano $t - f$.

Como se ha discutido en este Capítulo, el sistema auditivo posee características sobresalientes para el análisis de los sonidos del ambiente y en particular del habla humana. Se ha demostrado recientemente que ésto obedece a una adaptación para procesar los mismos de manera óptima en términos de la teoría de información. Se han desarrollado modelos de los sistemas sensoriales a nivel cortical que incorporan los aspectos de independencia estadística y dispersión[6] ya mencionados y que se han validado con experimentos realizados in vivo [145].

Las células de la corteza auditiva responden con mayor actividad ante determinados patrones $t-f$. A estos patrones se los conoce como *campos receptivos espectro-temporales* (STRF) de las células de la corteza auditiva [35]. Es posible ver al análisis cortical como ejecutado utilizando un diccionario óptimo adecuado. En la Figura 1.7 es posible ver la representación de algunos STRF, junto con los espectrogramas de los átomos obtenidos para un "diccionario óptimo" a partir de un conjunto de fonemas del español mediante las técnicas desarrolladas en esta tesis. La utilización de técnicas como éstas se encuadran en lo que podría denominarse *análisis "neuro-sensorial"*.

1.5. Comentarios de cierre del capítulo

Este trabajo está orientado al análisis y la representación de la señal de voz. Encontrar la representación óptima para esta señal es efectivamente sólo una parte del problema de diseñar sistemas de reconocimiento del habla con capacidades más cercanas a las humanas. Sin embargo, la representación de una señal se ha considerado como una parte fundamental de los problemas de clasificación o reconocimiento de patrones. Tanto es así que se ha asegurado que, de poderse encontrar una representación adecuada, el resto del problema quedaría prácticamente resuelto. La idea es que si se logra una representación que describa a la señal en términos de sus características más significativas, entonces se habrá "comprendido" adecuadamente la naturaleza de esta señal. Ésto significa que se habría logrado un buen modelo de ella, a partir del cual podrían tomarse decisiones confiables.

En este capítulo se presentaron en términos muy generales una serie de técnicas (convencionales y no convencionales) que conforman un amplio espectro de posibilidades para representar la señal de voz. Surge claramente la pregunta acerca de cuál resulta más adecuada en el contexto de los objetivos planteados en este trabajo. Es evidente también la dificultad en proporcionar una respuesta definitiva a esta pregunta. Se ha visto que los sistemas vivos realizan el procesamiento de las señales sensoriales en forma muy diferente a como lo hacen los sistemas artificiales actuales para intentar conseguir

[5]Se podría pensar entonces que simplemente se están reemplazando unas restricciones o condiciones por otras, sin embargo la diferencia radica en que estas nuevas condiciones parecen también haber sido incorporadas en los sistemas sensoriales de los organismos vivos [146].

[6]En este trabajo se utilizará el término dispersión (en inglés *sparsity*) para indicar cuan rala resulta la representación lograda.

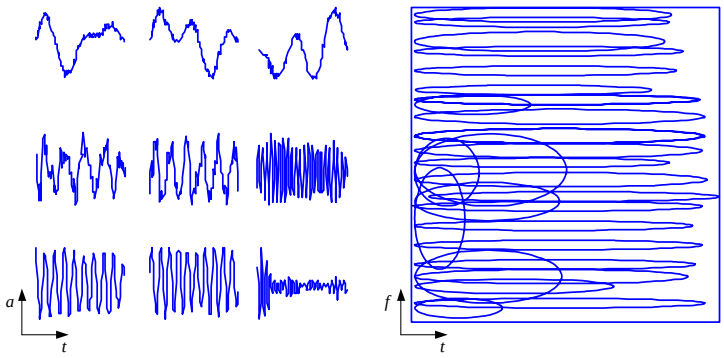

Figura 1.6: *Átomos (izquierda) y representación en el plano tiempo-frecuencia (derecha) para el diccionario óptimo obtenido a partir de datos del fonema /a/ del español mediante las técnicas de representación rala e independiente presentadas en esta tesis. Las frecuencias abarcan 4 KHz y el tiempo total es de 10 mseg. Es posible observar la superposición de algunos átomos en el plano $t - f$ debido a que no se requiere independencia lineal.*

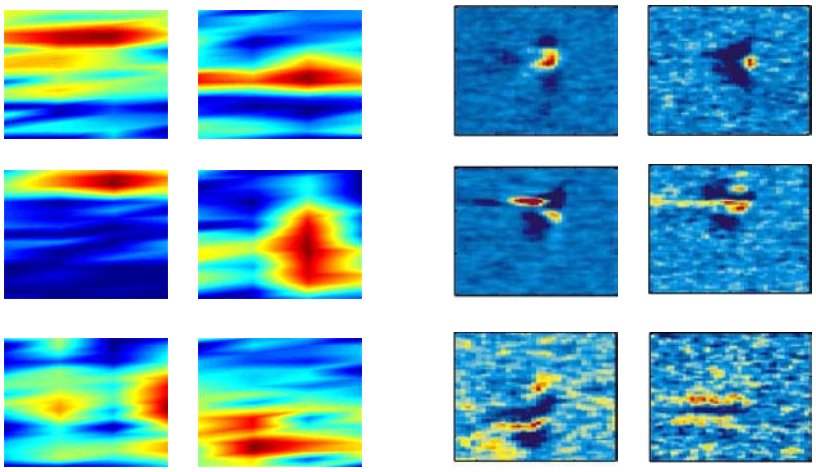

Figura 1.7: *Espectrogramas de los átomos obtenidos para un diccionario óptimo mediante las técnicas de representación rala e independiente de la Figura 1.6 (izquierda) y campos receptivos espectro-temporales de las células de la corteza auditiva [35] (STRF, derecha). Para los STRF se representa en las abscisas el tiempo (en mseg a partir del momento de estimulación), en las ordenadas la frecuencia (en octavas respecto a 120 Hz) y en colores la actividad de la célula nerviosa (en pulsos/segundo). De acuerdo con las propiedades del modelo utilizado para generar el diccionario es posible decir que los espectrogramas resultan "equivalentes" a los STRF auditivos.*

idénticos fines. Es posible entonces, como se ha hecho anteriormente a otros niveles, obtener algo de "inspiración" biológica que ayude en la búsqueda de esta respuesta. Utilizando la terminología del principio del capítulo es posible decir entonces que la meta consiste en encontrar el "prisma" que permita extraer las "componentes ocultas" en la señal del habla, de forma similar a como lo realiza el sistema auditivo humano.

Capítulo 2

Nociones preliminares y marco conceptual

"Porque nada hay oculto sino para ser descubierto, y no hay nada escondido sino para que venga a la luz."

(Marcos 4,22)

Contenido

2.1. Introducción

ESTE trabajo trata principalmente el problema de la representación "óptima" de la señal de voz para su posterior "interpretación" por parte de un sistema artificial. Para avanzar hacia la resolución de este problema es necesario comenzar por el principio, lo que implica establecer algunas nociones preliminares y definir el marco conceptual que permitirá la exposición de las ideas principales. Para abordar la solución de un problema tan complejo es necesario un enfoque interdisciplinario en el que confluyan ideas tomadas de diferentes campos, tales como: análisis de señales, teoría de la información, modelización de sistemas y análisis estadístico de datos. Además se requiere cierta comprensión de la fisiología de la comunicación humana, a lo que se dedicará un capítulo especial. La Matemática es el lenguaje "universal" que permitirá combinar estos enfoques y utilizarlos en forma de algoritmos para proveer las soluciones correspondientes. Por lo tanto

en el presente Capítulo se presentan una serie de nociones básicas de las distintas áreas, principalmente en lenguaje matemático, que serán necesarias para el desarrollo de esta tesis. Se alternarán enfoques continuos y discretos, por cuestiones de sencillez en la presentación de los diferentes conceptos y sus propiedades, explicitando cada contexto según sea necesario. El objetivo primordial es lograr que el material resulte bastante "auto-contenido", facilitando al lector el seguimiento del mismo. Sin embargo se darán por sentados algunos conocimientos básicos de álgebra matricial, estadística, reconocimiento de patrones y métodos de optimización para no engrosar demasiado el presente volumen.

El capítulo está organizado de la siguiente manera. En la primera parte, Sección 2.2, se incluyen los fundamentos del análisis de señales. A continuación, en la Sección 2.3 se exponen conceptos básicos relacionados con la teoría de información. La Sección 2.4 se dedica al enfoque de modelización de señales, donde se incluye la descripción de algunas técnicas que servirán de antecedente a los desarrollos de esta tesis. Finalmente se presentan las técnicas de análisis estadístico de datos en la Sección 2.5.

2.2. Análisis de señales

En el Capítulo 1 se introdujeron las ideas generales detrás del análisis de señales. En esta sección se presentan una serie de nociones matemáticas básicas que serán importantes en el resto del capítulo y de la tesis. En el Capítulo 4 se utilizarán varias de ellas para presentar las técnicas aplicables al análisis de la señal de voz más importantes. Aunque se requiere cierto cuidado, es posible establecer ciertas analogías entre los espacios euclídeos de la Geometría Clásica y algunos espacios funcionales. Estas analogías se utilizarán aquí principalmente con fines didácticos. El enfoque de esta sección es fundamentalmente de tiempo continuo. Por cuestiones de sencillez muchas de las nociones se definen en espacios vectoriales de dimensión finita. Es posible "extrapolar" algunas de estas nociones a espacios de dimensiones infinitas "bien comportados" de interés práctico, realizando las consideraciones correspondientes. Se prestará especial atención a los espacios de Hilbert que resultan de interés para las aplicaciones y desarrollos aquí considerados. Una parte importante de este trabajo está relacionada con un tipo particular de transformaciones lineales que permiten obtener diferentes representaciones de una señal en términos de sus "proyecciones" en un espacio de interés. El corazón de este tipo de transformaciones es el denominado *producto interno*. Para mayor detalle acerca de estos aspectos es posible consultar algunos de los textos clásicos de *análisis funcional* (por ejemplo Bachman y Narici (2000) [6]).

Definición 2.1 *Para el caso de dos señales continuas $x(t)$ y $y(t) \in L^2(\mathbb{C})$ el producto interno $\langle \cdot \rangle$ entre ellas queda definido como:*

$$\langle x(t), y(t) \rangle \triangleq \int_{-\infty}^{\infty} x(t) y^*(t) dt = c \, ,$$

donde $c \in \mathbb{C}$.

Desde el punto de vista conceptual esta operación entre dos señales nos devuelve una cantidad que nos da una idea del "aporte" o proyección de una señal en otra. En general esta cantidad suele normalizarse de manera de independizarse de la energía de la señal contra la cuál se está comparando.

De forma análoga el producto interno para el caso de dos señales de tiempo discreto $x[n]$ y $y[n] \in \mathbb{C}^N$ queda definido como:

$$\langle x[n], y[n] \rangle \triangleq \sum_{n=1}^{N} x[n] y^*[n] = c \, .$$

Esta idea de medir el grado de parecido entre una señal bajo estudio y un conjunto de señales con propiedades especiales de interés, constituye el fundamento de las transformaciones lineales mediante bases que se presentarán a continuación.

2.2.1. Transformaciones, bases y marcos

Una noción fundamental para el desarrollo de esta sección es la de *espacio vectorial* (*e.v.*). Dado un conjunto de señales o vectores $\mathbf{x} \in X$, junto con las operaciones de suma y producto por un escalar, este nuevo conjunto $\mathcal{X} = \{X, +, \cdot\}$ constituye un *e.v.*, si se cumplen las condiciones de la siguiente definición.

Definición 2.2 *Un conjunto $\mathcal{X}$ constituye un* espacio vectorial *sobre el* campo $\mathbb{F}$ *(por ejemplo $\mathbb{R}$ o $\mathbb{C}$) si, dadas:*

- *una operación de suma de vectores definida en $\mathcal{X}$, denotada como $\mathbf{x} + \mathbf{y}$ (donde $\mathbf{x}, \mathbf{y} \in \mathcal{X}$), y*
- *una operación de multiplicación escalar en $\mathcal{X}$, denotada como $a \cdot \mathbf{x}$ (donde $\mathbf{x} \in \mathcal{X}$ y $a \in \mathbb{F}$),*

se cumplen las siguientes 10 propiedades $\forall a, b \in \mathbb{F}$ y $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$:

1. $\mathbf{x} + \mathbf{y} \in \mathcal{X}$. *(Cerradura de la suma vectorial)*
2. $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$. *(Asociatividad de la suma vectorial)*
3. $\exists$ *un elemento neutro* $\mathbf{0} \in \mathcal{X}$, *tal que* $\forall \mathbf{x} \in \mathcal{X}, \mathbf{x} + \mathbf{0} = \mathbf{x}$. *(Elemento neutro)*
4. *para cada* $\mathbf{x} \in \mathcal{X}$, $\exists$ *un elemento* $\mathbf{z} \in \mathcal{X}$, *tal que* $\mathbf{x} + \mathbf{z} = \mathbf{0}$. *(Elemento cancelativo)*
5. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$. *(Conmutatividad de la suma vectorial)*
6. $a \cdot \mathbf{x} \in \mathcal{X}$. *(Cerradura de la multiplicación escalar)*
7. $a \cdot (b \cdot \mathbf{x}) = (a \cdot b) \cdot \mathbf{x}$. *(Asociatividad de la multiplicación escalar)*
8. $\exists$ *un elemento unitario* $1 \in \mathbb{F}$, *tal que* $\forall \mathbf{x} \in \mathcal{X}, 1 \cdot \mathbf{x} = \mathbf{x}$. *(Elemento unitario)*

9. $a \cdot (\mathbf{x} + \mathbf{y}) = a \cdot \mathbf{x} + a \cdot \mathbf{y}$. *(Distributividad respecto a la suma vectorial)*

10. $(a + b) \cdot (\mathbf{x}) = a \cdot \mathbf{x} + b \cdot \mathbf{x}$. *(Distributividad respecto a la suma escalar)*

El producto interno permite adoptar una perspectiva "geométrica" y utilizar terminología familar de los espacios de dimensión finita. De todos los espacios vectoriales topológicos de dimensiones infinitas, los *espacios de Hilbert* son los que "mejor se comportan" y los "más cercanos" a los espacios de dimensiones finitas.

Cada producto interno (según la Definición 2.1) en un *e.v.* da lugar a una norma de la siguiente forma:

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}. \tag{2.1}$$

Si se agrega entonces la operación de producto interno a un *e.v.* y es *completo* con respecto a la norma (2.1), entonces éste constituye un espacio de Hilbert[1] (que se denotará como $\mathcal{H}$). Completitud en este contexto significa que cualquier *secuencia de Cauchy*[2] de elementos del espacio converge también a un elemento dentro del espacio, en el sentido que la norma de las diferencias se aproxima a cero.

Cada espacio de Hilbert es también un *espacio de Banach* (*e.v.* normado y completo), pero no viceversa. Todos los *e.v.* de dimensiones finitas con producto interno (tales como el espacio euclídeo con el producto escalar ordinario) son espacios de Hilbert. Sin embargo, los casos de dimensiones infinitas son importantes en muchas aplicaciones.

Definición 2.3 *Dado un conjunto de señales* $\Phi = \{\phi_i\}_{i \in \Gamma}$, *se denomina* combinación lineal *de las mismas a una expresión* $x(t)$ *de la forma [135]:*

$$x(t) \triangleq \sum_{i \in \Gamma} a_i \phi_i(t) \ , \tag{2.2}$$

donde $a_i \in \mathbb{C}$ *y* Γ *es un conjunto* finito *de escalares definido de manera adecuada*[3].

Se dice que $x(t)$ es linealmente dependiente del conjunto Φ si y sólo si se puede escribir a $x(t)$ como una combinación lineal de las $\phi_i(t)$. En caso contrario se dice que $x(t)$ es linealmente independiente del conjunto Φ.

Al variar los coeficientes a_i, es decir, al generar todas las combinaciones lineales posibles de las $\phi_i(t)$, el resultado es un nuevo conjunto X de señales $x_j(t)$ que a su vez heredan muchas de las propiedades de las $\phi_i(t)$ que las generaron. A este nuevo conjunto se lo denomina *expansión lineal* del conjunto Φ.

Si ahora $\mathcal{X} = \{X, +, \cdot\}$ constituye un *e.v.* de acuerdo con la Definición 2.2 entonces se dice que el conjunto Φ genera el espacio $\mathcal{X}$ si, para toda señal $x_j(t) \in X$, existe el conjunto de escalares $A_j = \{a_i\}_j$ tales que $x_j(t)$ se pueda expresar como una combinación lineal de los elementos de Φ.

[1] Si no fuera completo se habla de un pre-Hilbert.

[2] Una secuencia de Cauchy es aquella en la cual sus términos se vuelven arbitrariamente cercanos a medida que la secuencia progresa.

[3] Entre los casos de interés para la definición de Γ el más sencillo es aquel para el cual $\Gamma = \{1, \cdots, N\}$ con $N \in \mathbb{N} < \infty$. Resulta también de interés en este trabajo la definición de un $\Gamma \subset \Gamma'$ para el caso donde los elementos para la combinación lineal se seleccionan de alguna manera particular a partir de un conjunto mayor de posibilidades. Es necesario recalcar que Γ debe ser siempre finito, aunque en este último caso Γ' podría ser infinito.

Definición 2.4 *Se dice que un conjunto de señales* $\Phi = \{\phi_i\}_{i\in\Gamma}$ *es* linealmente independiente *si la relación:*

$$\sum_{i\in\Gamma} a_i\phi_i(t) = 0 \ ,$$

sólo puede satisfacerse para el caso en que $a_i = 0 \ \ \forall i \in \Gamma$.

Dicho de otro modo, un conjunto es linealmente independiente si ninguna de sus señales puede expresarse como combinación lineal de las demás señales del mismo conjunto.

Definición 2.5 *Dado un espacio vectorial* $\mathcal{X}$ *y un subconjunto de señales* $\Phi = \{\phi_i\}_{i\in\Gamma} \in \mathcal{X}$*, se dice que* Φ *constituye una* base *(de Hamel) de* $\mathcal{X}$ *si:*

1. Φ *es linealmente independiente,*
2. Φ *genera a* $\mathcal{X}$.

Es decir, para que Φ sea una base del *e.v.* $\mathcal{X}$, cualquier vector perteneciente a $\mathcal{X}$ debe poder escribirse como una combinación lineal de los vectores de Φ y además, ninguno de los vectores de Φ debe poder escribirse como una combinación lineal de los demás.

Se puede probar que dos bases (de Hamel) de un mismo *e.v.* $\mathcal{X}$ poseen la misma cardinalidad $|\mathcal{X}|$. Se define la dimensión D de un *e.v.* $\mathcal{X}$ como $D \triangleq |\mathcal{X}|$ y se corresponde con el número de señales que tiene cualquier base de dicho espacio (una cantidad menor no es suficiente para generarlo). Se puede demostrar que cualquier subconjunto de $N > D$ señales de $\mathcal{X}$, será linealmente dependiente (en este caso son demasiadas señales).

Una propiedad importante de una base es la unicidad de la representación, es decir que cada elemento del espacio posee una representación única en términos de los elementos de esa base.

Definición 2.6 *Sea* $\mathcal{X}$ *un e.v. con producto interno y sea* $\Phi = \{\phi_i\}_{i\in\Gamma}$ *un subconjunto de* $\mathcal{X}$*, se dice entonces que el conjunto* Φ *es* ortogonal *si se verifica que para todos sus elementos:*

$$\langle\phi_i(t), \phi_j(t)\rangle = 0 \quad \forall i \neq j \quad \text{con} \quad i, j \in \Gamma \ , \tag{2.3}$$

si además:

$$\langle\phi_i(t), \phi_i(t)\rangle = 1 \quad \forall i \in \Gamma \ , \tag{2.4}$$

entonces se dice que el conjunto es ortonormal.

Obsérvese que *ortogonalidad* $\Rightarrow$ *independencia lineal* pero lo contrario no siempre se cumple (Ver Figura 2.1). Por lo tanto pueden existir bases (de Hamel) ortogonales y no-ortogonales. El caso ortogonal resulta de interés debido a que da lugar a una fórmula sencilla para el cálculo de los coeficientes de la transformación producida por la base.

Antes de continuar es necesario aclarar que el concepto de bases para espacios de dimensión finita o bases de Hamel es diferente del de bases para espacios de dimensión infinita. El caso general corresponde a las denominadas bases de Schauder en los espacios de Banach. Resulta de interés la definición de bases ortonormales para los espacios de Hilbert. Para definirlas se requiere previamente introducir el concepto de *densidad*.

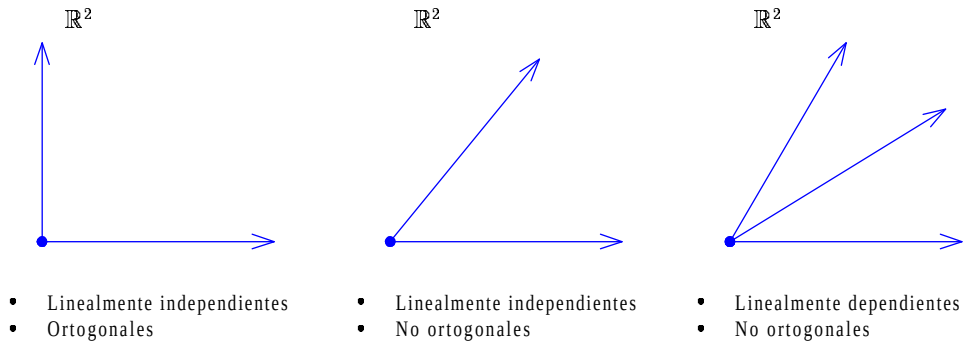

Figura 2.1: *Tres conjuntos de vectores en $\mathbb{R}^2$: base ortogonal (izquierda), base no ortogonal (centro), marco (derecha).*

Definición 2.7 *Un subconjunto Φ de un conjunto parcialmente ordenado H es* denso *en H si para cualquier $\mathbf{x} < \mathbf{y} \in H$, hay algún $\mathbf{z} \in \Phi$ tal que $\mathbf{x} < \mathbf{z} < \mathbf{y}$.*

Definición 2.8 *Sea $\mathcal{H}$ un espacio de Hilbert y sea $\Phi = \{\phi_i\}_{i\in\Gamma}$ un subconjunto de $\mathcal{H}$, se dice entonces que el conjunto Φ es una* base ortonormal *para $\mathcal{H}$ si:*

1. *el conjunto Φ es ortonormal,*
2. *la expansión lineal de Φ es densa en $\mathcal{H}$.*

En efecto si Φ constituye una base ortogonal (quitando la condición de normalidad de la definición anterior) para un espacio de Hilbert $\mathcal{H}$, entonces es posible probar que cualquier señal $x(t)$ se puede escribir como:

$$x(t) = \sum_{i\in\Gamma} \frac{\langle x(t), \phi_i(t)\rangle}{\|\phi_i(t)\|}\, \phi_i(t) = \sum_{i\in\Gamma} a_i\, \phi_i(t)\ . \tag{2.5}$$

Aún si Φ es *no numerable* se puede probar que solamente una cantidad numerable de términos en esta suma son distintos de cero y la expresión está por lo tanto bien definida. A esta suma se la llama también *expansión de Fourier de $x(t)$ en la base Φ*.

De esta forma cada coeficiente a_i en (2.5) es una medida del *parecido* entre la señal y el elemento correspondiente de la base. Los a_i se denominan *coeficientes de Fourier generalizados* y como se discutió anteriormente representan la *componente* de la señal $\phi_i(t)$ en $x(t)$. El término de normalización en (2.5) se vuelve innecesario para el caso ortonormal.

Otra propiedad interesante de una base ortogonal Φ de $\mathcal{H}$, es que existe una constante $A > 0$ para cualquier señal $x(t) \in \mathcal{H}$, tal que:

$$\sum_{i\in\Gamma} |\langle x(t), \phi_i(t)\rangle|^2 = A\, \|x(t)\|^2\ .$$

Para el caso ortonormal este resultado se conoce como *fórmula de Plancherel* y permite asegurar la conservación de la energía entre ambos espacios.

El requerimiento de ortogonalidad puede resultar innecesario en algunas aplicaciones, debido a que en general se impone por cuestiones de simplificación y no por características especiales del problema. Por ejemplo, no existe evidencia fisiológica de que las representaciones internas a nivel de los sistemas sensoriales constituyan bases ortogonales (Ver Capítulo 3). Más bien la evidencia apunta a suponer que existen representaciones altamente redundantes y mecanismos intrinsecamente no lineales.

Cuando se pierde el requerimiento de ortogonalidad, especialmente en un espacio de dimensión infinita, es necesario imponer alguna equivalencia de energía aunque sea parcial. Ésto es necesario para garantizar la estabilidad de la transformación. Puede inclusive perderse el requerimiento de independencia lineal y, bajo condiciones generales, ser todavía útil el conjunto para describir una señal en términos de sus proyecciones. Este último caso es el de los *marcos*, que se definirá a continuación.

Definición 2.9 *Dado un espacio de Hilbert $\mathcal{H}$, se dice que el conjunto de señales Φ constituye un* marco *de $\mathcal{H}$ si existen dos constantes $A > 0$ y $B > 0$ tales que para toda $x(t) \in \mathcal{H}$:*

$$A \left\| x(t) \right\|^2 \leqslant \sum_{i \in \Gamma} \left| \langle x(t), \phi_i(t) \rangle \right|^2 \leqslant B \left\| x(t) \right\|^2 \ ,$$

donde A y B se denominan cotas del marco. Cuando $A = B$ se dice que el marco es ajustado.

Podría decirse que los marcos constituyen un concepto más general que las bases. Ambos permiten la descomposición de la señal en términos de productos internos, pero en este caso no es necesaria la independencia lineal del conjunto. Para la reconstrucción de la señal se recurre nuevamente a la colección de las proyecciones sobre los elementos del marco. Sin embargo para esta reconstrucción se utiliza otro conjunto de señales, denominado marco dual $\mathring{\Phi}$:

$$x(t) = \sum_{i \in \Gamma} \langle x(t), \phi_i(t) \rangle \ \ \mathring{\phi}_i(t) \ . \tag{2.6}$$

El marco dual queda definido por:

$$\left\langle \phi_i(t), \mathring{\phi}_j(t) \right\rangle = \delta[i-j] \ \ \forall i, j \in \Gamma \ . \tag{2.7}$$

Se suele asociar a los marcos con la idea de redundancia, ésto significa que poseen más elementos de los necesarios para representar una señal del espacio. Esta redundancia puede ser útil en varias aplicaciones, especialmente provee cierta robustez frente a los efectos del ruido. Se debe aclarar que la reconstrucción propuesta por (2.6) a partir de los $\mathring{\phi}_i(t)$ no es la única posible. De esta forma, de acuerdo al grado de redundancia, se podrían emplear sólo los coeficientes menos afectados por el ruido para la reconstrucción. Los elementos del marco dual están relacionados en el caso de tiempo discreto y dimensiones finitas con las denominadas matrices pseudoinversas (que se discutirá en el Capítulo 6).

Existen casos de conjuntos de funciones aún más generales que todavía son útiles para la descripción y análisis de señales. En estos casos suelen ser de interés algunas

características del conjunto, como por ejemplo su localización tiempo-frecuencia. Ésto da lugar a la siguiente definición.

Definición 2.10 *Considere una familia general de señales* $\Phi = \{\phi_\gamma(t)\}_{\gamma\in\Gamma}$ *con características particulares que dependen de un parámetro* γ *(que puede ser un parámetro multi-índice) con* $\phi_\gamma(t) \in L^2(\mathbb{R})$ *y* $\|\phi_\gamma\| = 1$*. A cada una de estas señales* $\phi_\gamma(t)$ *se las denomina* átomos *y al conjunto* Φ *se lo denomina* diccionario*.*

2.3. Teoría de información

La teoría de información provee un marco útil para expresar varias de las ideas presentadas en esta tesis o para valorar los resultados obtenidos, por lo que es importante examinar aquí los conceptos más importantes de esta teoría. Las definiciones se presentarán principalmente para el caso unidimensional [71]. Es posible encontrar más detalles en los libros clásicos del área (por ejemplo Cover y Thomas (1991) [30]).

2.3.1. Información y entropía

La *información* $I(p)$ asociada con un evento de probabilidad p se define como [71]:

$$I(p) \triangleq \log(1/p) \,, \tag{2.8}$$

Como medida de información esta sencilla fórmula posee dos propiedades muy interesantes:

1. Cuanto más improbable es un evento es cuando más información aporta, puesto que un evento cierto ($p = 1$) no aporta ninguna información adicional.

2. La información que se obtiene de observar dos eventos independientes (cuya probabilidad conjunta es el producto de sus probabilidades individuales p_1 y p_2) es la suma de las informaciones individuales: $I(p_1p_2) = I(p_1) + I(p_2)$.

Para un sistema completo es útil tener un modelo para el mecanismo que "genera" la información. El más simple de estos modelos es el correspondiente a la *fuente de información discreta sin memoria*. Ésto significa que la fuente genera una secuencia de símbolos en forma aleatoria (seleccionándolos a partir de un alfabeto o conjunto de símbolos posibles). Además cada símbolo dentro de esta secuencia no tiene dependencia con el anterior. En adelante se supondrá que todas las fuentes son de este tipo.

Definición 2.11 *Sea* $\overline{X}$ *una* fuente de información discreta sin memoria*, que emite una secuencia de símbolos de un alfabeto finito* $\mathcal{A}_{\overline{X}} = \{x^1, x^2, \ldots, x^n\}$*, donde cada símbolo ocurre en forma independiente con una probabilidad dada por la función* $p_{\overline{X}}(x)$*, para*

$x \in \mathcal{A}_{\overline{X}}$. Entonces se define a la entropía $\mathcal{H}$ *de la fuente* $\overline{X}$ *como la información media obtenida por cada símbolo:*

$$\begin{aligned} \mathcal{H}(\overline{X}) &\triangleq \underset{x \in \mathcal{A}_{\overline{X}}}{\mathcal{E}} \left[\log \tfrac{1}{p(x)} \right] \\ &= \sum_{x \in \mathcal{A}_{\overline{X}}} p(x) \log \tfrac{1}{p(x)} \\ &= - \sum_{x \in \mathcal{A}_{\overline{X}}} p(x) \log p(x) \ , \end{aligned} \tag{2.9}$$

donde se ha utilizado $p(x)$ en forma abreviada para $p_{\overline{X}}(x)$, y $\mathcal{E}[\cdot]$ corresponde al operador valor esperado[4]*.*

Si el logaritmo en (2.9) es en base 2, entonces la entropía se mide en bits.

De acuerdo con la Definición 2.11, $\mathcal{H}(\overline{X}) \geq 0$. En un extremo, si todos los n posibles valores para $\overline{X}$ son igualmente probables, entonces la entropía toma su máximo valor igual a $\log n$. En el otro extremo, si existe un solo valor posible entonces se trata del caso determinístico y $\mathcal{H}(\overline{X}) = 0$. De esta manera es posible asociar la entropía con la "agudeza" del pico principal de la distribución de probabilidad de los valores de $\overline{X}$, cuando más "picuda" es $p(x)$, más pequeña es la entropía. Esta propiedad será utilizada más adelante. En términos de codificación, la entropía da una idea acerca de la longitud promedio del código más corto posible para representar los valores de $\overline{X}$, ésto es la *eficiencia* de la codificación[5]. En este sentido $\mathcal{H}$ es también una medida de la compresibilidad de una fuente de información.

La entropía puede ser generalizada también para el caso de una fuente que tome valores continuos, en cuyo caso se la suele denominar como *entropía diferencial*:

$$h(\overline{X}) \triangleq - \int_S p(x) \log p(x) dx \ ,$$

donde S es el conjunto sobre el cual $p(x) > 0$.

Un problema con la entropía diferencial es que no se comporta tan bien como su contraparte discreta, en particular puede tomar valores negativos y es sensible a cambios de escala.

Para evitar algunos de estos problemas es que suele definirse la *negentropía* como:

$$j(\overline{X}) \triangleq -h(\overline{X}) + h(\overline{X}_G) \ ,$$

donde $\overline{X}_G$ es una variable aleatoria ($v.a.$) normalmente distribuida con igual varianza σ^2 que $\overline{X}$. Es fácil demostrar que la negentropía es invariante a la escala de $\overline{X}$ y

[4] No deben confundirse los x^i en esta definición, que constituyen los posibles valores o símbolos de $\overline{X}$, con los x_i que constituían las señales del conjunto X en la sección anterior.

[5] En *teoría de la comunicación* un *código* es una regla que permite convertir una pieza de información en otra forma de representación, no necesariamente del mismo tipo. *Codificación* es el proceso mediante el cual se realiza esta conversión de la información contenida en el mensaje a fin de establecer la comunicación entre el transmisor y el receptor (Ver Figura 3.1). Aquí código se utilizará como sinónimo de *representación*.

además, debido a que la distribución normal maximiza la entropía para una varianza fija, está garantizada su no negatividad [71].

Otra forma equivalente para poder realizar comparaciones mediante la entropía diferencial, es la de igualar primero las varianzas.

2.3.2. Entropía conjunta y condicional

Si ahora se consideran dos fuentes $\overline{X}$ y $\overline{Y}$ es posible definir algunas cantidades relacionadas con la dependencia entre ellas.

Definición 2.12 *Sean $\overline{X}$ y $\overline{Y}$ dos fuentes con alfabetos $\mathcal{A}_{\overline{X}}$ y $\mathcal{A}_{\overline{Y}}$ respectivamente, entonces se define la* entropía conjunta $\mathcal{H}(\overline{X},\overline{Y})$ *como:*

$$\begin{aligned}\mathcal{H}(\overline{X},\overline{Y}) &\triangleq \underset{\substack{x\in\mathcal{A}_{\overline{X}}\\ y\in\mathcal{A}_{\overline{Y}}}}{\mathcal{E}}[\log p(x,y)] \\ &= -\sum_{x\in\mathcal{A}_{\overline{X}}}\sum_{y\in\mathcal{A}_{\overline{Y}}} p(x,y)\log p(x,y)\ .\end{aligned} \tag{2.10}$$

Además es posible definir también la entropía condicional *como:*

$$\begin{aligned}\mathcal{H}(\overline{Y}|\overline{X}) &\triangleq \underset{\substack{x\in\mathcal{A}_{\overline{X}}\\ y\in\mathcal{A}_{\overline{Y}}}}{\mathcal{E}}[\log p(y|x)] \\ &= -\sum_{x\in\mathcal{A}_{\overline{X}}} p(x)\sum_{y\in\mathcal{A}_{\overline{Y}}} p(y|x)\log p(y|x) \\ &= -\sum_{x\in\mathcal{A}_{\overline{X}}}\sum_{y\in\mathcal{A}_{\overline{Y}}} p(x,y)\log p(y|x)\ .\end{aligned} \tag{2.11}$$

2.3.3. Entropía y complejidad

La entropía clásica de Shannon $\mathcal{H}$, o varias nociones generalizadas a partir de ella, pueden utilizarse para caracterizar la complejidad de sistemas dinámicos no lineales o señales derivadas de éstos [28, 181].

La entropía de Harvda-Charvat-Daróvczy-Tsallis o *q–entropía* [72, 32, 207], que depende de un parámetro real $q \neq 1$, puede escribirse como:

$$\mathcal{H}_q(\overline{X}) \triangleq (q-1)^{-1}\sum_{x\in\mathcal{A}_{\overline{X}}} p(x) - p(x)^q. \tag{2.12}$$

La q–entropía se ha aplicado también a la detección de cambios suaves (en inglés *slight*) en los parámetros de un sistema mediante el análisis de las señales involucradas [205, 204].

2.3.4. Entropía relativa e información mutua

A veces es importante tener una idea de la "distancia" entre dos distribuciones de probabilidad p y r. La entropia relativa puede interpretarse en este sentido.

Definición 2.13 *Se define a la* entropía relativa *o* divergencia de Kullback-Leibler *entre dos distribuciones de probabilidad* $p(x)$ *y* $r(x)$ *como:*

$$\begin{aligned} \mathcal{D}_{KL}(p\|r) &\triangleq \underset{x\in\mathcal{A}_{\overline{X}}}{\mathcal{E}}\left[\log\tfrac{p(x)}{r(x)}\right]_{p(x)} \\ &= \sum_{x\in\mathcal{A}_{\overline{X}}} p(x)\log\tfrac{p(x)}{r(x)}\ , \end{aligned} \tag{2.13}$$

donde $\mathcal{E}[\cdot]_{p(x)}$ *corresponde al valor esperado sobre la distribución* $p(x)$*.*

En el sentido estricto $\mathcal{D}_{KL}(p\|r)$ no constituye una distancia métrica porque no es simétrica ni tampoco cumple con la desigualdad triangular. Sin embargo tiene la propiedad de que $\mathcal{D}_{KL}(p\|r) \geq 0$, cumpliéndose la igualdad si y sólo si $p = r$ [71]. La misma definición se utiliza para distribuciones multi-dimensionales.

En el caso de la q–entropía para $q \in \mathbb{R} - \{1\}$, la correspondiente q–entropía relativa está dada por [203]:

$$\mathcal{D}_q(p\|r) \triangleq \frac{1}{1-q}\sum_{x\in\mathcal{A}_{\overline{X}}} p(x)\left[1-\left(\frac{p(x)}{r(x)}\right)^{q-1}\right]. \tag{2.14}$$

Otra medida de gran utilidad es la denominada *información mutua* (MI).

Definición 2.14 *La* información mutua $\mathcal{I}(\overline{X};\overline{Y})$ *entre dos fuentes* $\overline{X}$ *y* $\overline{Y}$ *se define como la entropía relativa entre su distribución conjunta y el producto de sus distribuciones marginales:*

$$\begin{aligned} \mathcal{I}(\overline{X};\overline{Y}) &\triangleq D_{KL}(p(x,y)\|p(x)p(y)) \\ &= \sum_{\substack{x\in\mathcal{A}_{\overline{X}}\\ y\in\mathcal{A}_{\overline{Y}}}} p(x,y)\log\tfrac{p(x,y)}{p(x)p(y)}\ . \end{aligned} \tag{2.15}$$

De las propiedades estadísticas básicas se desprende que información mutua igual a cero es condición necesaria y suficiente para la independencia estadística. Si X y Y forman parte de un mismo código, la MI es una medida de la redundancia entre ellas. Un esquema conceptual que ilustra las diferentes relaciones entre las cantidades obtenidas a partir del concepto de entropía se presenta la Figura 2.2.

2.4. Modelización de señales

El problema de representación de una señal puede plantearse en términos de encontrar un modelo adecuado de la misma. A partir de este modelo, que se denotará con (la letra) $\mathcal{M}$, es posible generar la señal mediante un conjunto de entradas que estarán relacionadas con diversas características de la misma (modelo *directo*). Además, es también posible

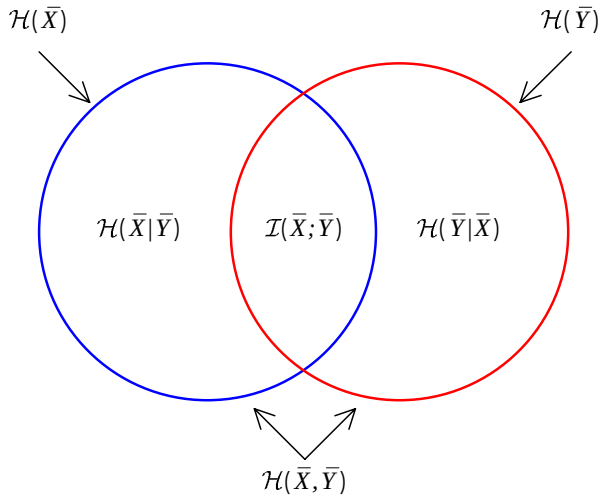

Figura 2.2: *Representación de las relaciones entre diferentes cantidades entrópicas para dos v.a.* $\overline{X}$ *y* $\overline{Y}$ *(adaptado de [71]).*

"invertir" el modelo para obtener, ahora mediante $\mathcal{M}^{-1}$, las características significativas partiendo de una señal determinada[6]. Para realizar la estimación de los parámetros de un modelo se requiere una gran cantidad de *datos* del entorno. A la salida de un modelo se la suele llamar *representación* o *código* . Por ejemplo, el modelo que utilizan los sistemas sensoriales biológicos para capturar una adecuada descripción de su *entorno* es en realidad un modelo *inverso*[7] (Ver Figura 2.3). .

Ésto constituye fundamentalmente un cambio de enfoque, frente al clásico de análisis de señales, y permite establecer una serie de hipótesis que debe cumplir el modelo mediante el cual se obtiene una representación. En esta sección se revisarán brevemente los fundamentos de este enfoque, la terminología básica y las hipótesis más importantes que conducen a diferentes métodos de representación.

2.4.1. Pautas para evaluar un modelo

Dado que un buen modelo implica una buena representación de la señal, es importante tener algunos criterios que permitan evaluar este aspecto. Si se dispone de un conjunto de modelos es necesario también decidir acerca de cuál es el que mejor representa una señal. Para ello es posible identificar los siguientes factores que pueden ayudar en la tarea [71]:

Completitud. Si un modelo es completo (o *sin pérdida*), todos los aspectos de la entrada están representados en su salida. Ésto permite una reconstrucción completa

[6]La distinción entre modelo directo e inverso constituye un aspecto relativo a la forma en la que se plantea el modelo en relación con las características causa-efecto del fenómeno bajo estudio. Muchos problemas prácticos terminan en modelos inversos que resultan más difíciles de tratar, según se discutirá más adelante.

[7]Esta interpretación puede ser importante en el contexto de la señal de voz y su representación sensorial interna según se discutirá más adelante.

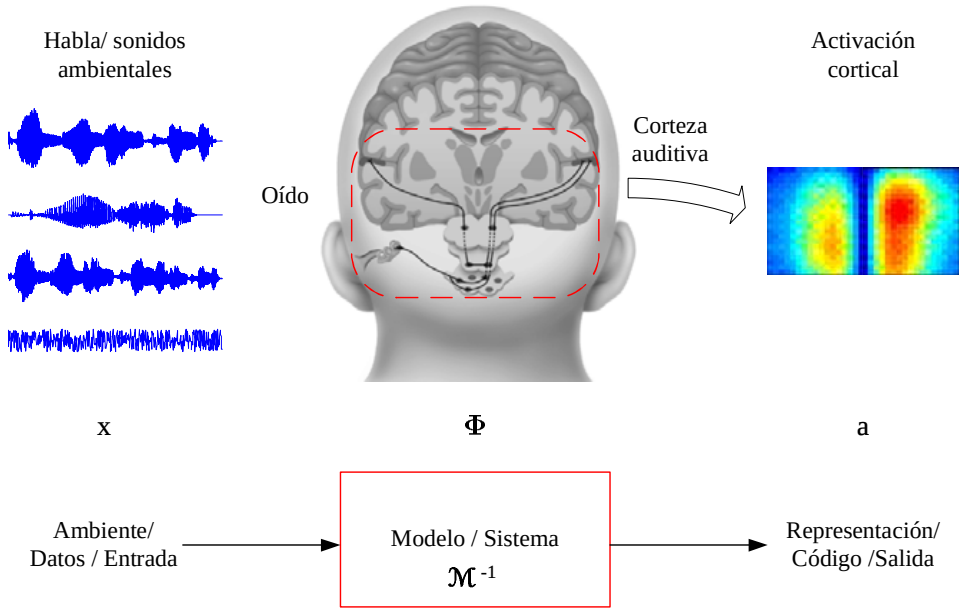

Figura 2.3: *Esquema de las partes que componen un modelo y su correspondiente terminología (abajo), ejemplificado para el sistema sensorial auditivo humano (que resulta en realidad un modelo inverso, arriba). En este contexto biológico el proceso de estimación de los parámetros del modelo se denomina* aprendizaje *y el proceso que permite generar la salida del modelo se conoce como* inferencia.

de la señal a partir de la representación del modelo. El caso *con pérdida* aparece en presencia de ruido, imprecisión o porque el modelo no posee suficiente potencia expresiva para representar totalmente a la entrada. Ésto no necesariamente constituye un problema, debido a que pueden existir aspectos de la entrada que no resulten de interés para el sistema de procesamiento subsecuente.

Complejidad. En el caso de obtener varios modelos completos, o "igualmente" completos entre sí, se debería disponer de otro elemento que permita seleccionar al mejor de entre ellos. El principio de la denominada *navaja de Occam* sugiere elegir el más simple[8]. Por supuesto que se requiere definir aquí que se entiende por *complejidad* y "dónde" debe medirse. La complejidad podría medirse en términos intrínsecos al modelo (por ejemplo su estructura) o en términos de la representación lograda por el mismo (por ejemplo la cantidad de elementos del código). No es posible disminuir ambas "complejidades" al mismo tiempo (si se mantiene la completitud), por lo que en general se debe tomar una decisión en función de otros factores. Existe también un compromiso entre completitud y complejidad, o sea que si se quiere mantener una representación simple en general se debe sacrificar algo de la habilidad del modelo para representar todos los aspectos de la entrada.

[8]La navaja de Occam (o navaja de Ockham, o principio de economía) se atribuye al fraile franciscano inglés del siglo XIV Guillermo de Ockham. En su forma más simple indica que las explicaciones nunca deben multiplicar las causas sin necesidad. Cuando dos explicaciones se ofrecen para un fenómeno, la explicación completa más simple es preferible.

Generalización. Los buenos modelos "concuerdan" con la realidad. Es difícil definir qué se entiende por esta concordancia, pero una posible medida la constituye la capacidad del modelo para generalizar o predecir el comportamiento de la realidad en base a nuevos datos del ambiente.

Ajuste al propósito. Un modelo constituye sólo un medio para llegar a un fin, una base para describir un problema o decidir un curso de acción. De esta forma su valor debe ser determinado en función de su utilidad para alcanzar su meta final.

En las secciones siguientes precisaremos mejor algunos de estos aspectos y sus implicancias.

Propósito de este trabajo

Para poder valorar la última pauta presentada, es decir el ajuste al propósito, se requiere precisar más el propósito de este trabajo. Se puede decir que éste consiste en encontrar una representación óptima de la señal de voz en términos de:

1. Realce de las pistas acústicas y los eventos más relevantes de la señal.
2. Independencia con respecto a cambios no "sustanciales" en la señal.
3. Robustez al ruido y a las perturbaciones.
4. Clasificación dinámica de los patrones acústicos generados en función de unidades fonético-lingüísticas:
   a) Discriminación entre patrones correspondientes a diferentes unidades.
   b) Secuenciación temporal de estas unidades.
   c) Generalización a señales no "observadas" previamente (en forma similar a lo discutido previamente).

Podría agregarse que es importante también lograr algún correlato con las representaciones internas de los sistemas biológicos equivalentes, que incorporan explícitamente estos aspectos. Por otro lado, resulta difícil atacar todos los aspectos simultáneamente, por lo que es necesario dividir el problema en varias etapas. Los primeros tres aspectos pueden ser abordados mediante una etapa que contenga un modelo adecuado de la señal de voz. Este modelo debe tener en cuenta las características perceptuales importantes de la misma y debe desechar cualquier otra información. Para abordar el diseño de esta etapa de forma independiente de la siguiente es necesario utilizar métodos de tipo *no supervisado*. Ésto significa que sean capaces de "descubrir" estas características en forma automática y sin referencias a la identidad de las unidades fonético-lingüísticas correspondientes. En este trabajo se pretende avanzar en la construcción de esta etapa por lo que no se prestará la misma atención al resto de los aspectos mencionados, que requieren un enfoque más supervisado y que deberá formar parte de una etapa posterior. Sin embargo se verá que el tipo de enfoque utilizado permite también atacar en forma "temprana" alguna de las otras dimensiones.

Se presentarán a continuación dos criterios que pueden aplicarse al "diseño" de los modelos y que permiten aproximarse en la dirección de algunas de las características o pautas mencionadas para generar un buen modelo para nuestro problema:

1. Maximizar la independencia estadística entre los elementos de la representación.
2. Lograr que esta representación sea lo más rala posible.

Redundancia e independencia estadística

Se ha establecido que la *reducción* de la redundancia es un paso importante hacia el "completo entendimiento" de los datos [8]. En particular ésto es especialmente aplicable si el objetivo consiste en realizar una *detección de características* de bajo nivel. Estas "características" representan las dependencias existentes en un conjunto de datos. En el caso de las imágenes, por ejemplo, las características visuales tales como las líneas dan cuenta de que la intensidad de un punto de la imagen tiende a ser altamente dependiente de la de sus vecinos. Así mismo, en el caso de la señal de voz estas dependencias aparecen tanto a nivel temporal como espectral. Al mismo tiempo se espera que la dependencia entre diferentes características sea baja, de otro modo éstas deberían juntarse en una o más características "mayores". A veces tampoco es posible eliminar fácilmente todas las dependencias, debido a que en general éstas se extienden a diferentes niveles o escalas[9] [142]. Se puede decir entonces que un *código basado en características* deberá tener poca redundancia, o sea que deberá existir cierta independencia estadística entre los coeficientes de la representación. Además esta independencia estadística permite modelar procesos complejos de manera más sencilla, debido a que los mismo pueden descomponerse en procesos independientes. Por ello un código o representación de este tipo es óptimo en términos de completitud y complejidad [71]. Otra ventaja de la independencia estadística es que si se produce algún tipo de daño en algún elemento del código este no se propaga al resto.

Ha podido demostrarse que la reducción de la redundancia constituye un mecanismo subyacente en muchas etapas de procesamiento dentro del sistema nervioso, más específicamente en los sistemas sensoriales, incluyendo la vía auditiva [191, 19]. Sin embargo actualmente se prefiere utilizar en este contexto el término *explotación* en lugar de reducción [8]. En este sentido se entiende que cada etapa "aprovecha" la redundancia presente en los datos, con lo que se revaloriza su importancia, aunque no necesariamente existe siempre una reducción[10]. Algunos de estos aspectos se discutirán en el Capítulo siguiente.

Códigos locales, completamente distribuidos y ralos

Muchas veces es importante lograr una representación en términos de un número pequeño de elementos descriptores tomados de un conjunto grande. Los modelos que

[9] Para el análisis de esta clase de datos con dependencia a varias escalas es necesario un enfoque del tipo multiescala o multiresolución.

[10] De hecho la redundancia a veces aumenta y ésto puede servir como mecanismo para lograr cierta inmunidad al ruido.

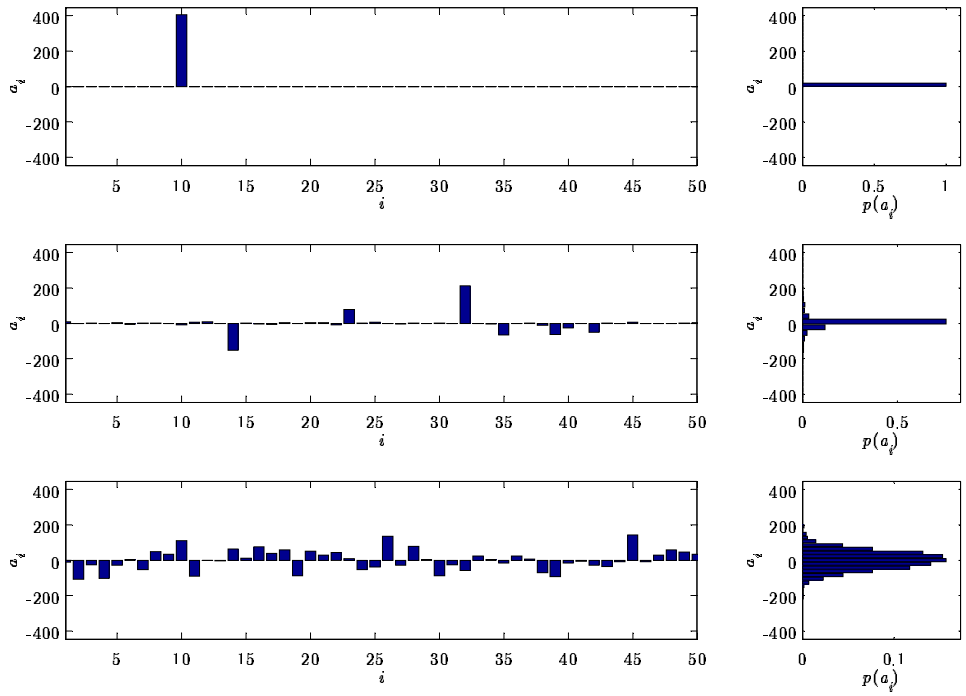

Figura 2.4: *Códigos locales (arriba), ralos (centro) y completamente distribuidos (abajo). Éstos dos últimos se generaron con igual varianza.*

producen este tipo de representaciones pueden considerarse óptimos en el balance de varios de los aspectos discutidos en la sección anterior. A estas representaciones se las denomina *ralas*. En realidad la codificación rala es sólo una de las posibilidades dentro de un espectro que va desde los códigos *locales* a los *completamente distribuidos* [71]. En la Figura 2.4 se muestra un ejemplo de la apariencia típica que presentan los códigos de cada tipo. En los denominados códigos locales existe un solo elemento activo por cada patrón. La ventaja de este tipo de códigos es la facilidad para realizar una interpretación o clasificación de los patrones, sin embargo poseen varias desventajas tales como: la cantidad de dimensiones, la posibilidad de generalización, y la dependencia entre las dimensiones. En el otro extremo se encuentran los códigos completamente distribuidos (o "densos"), donde todos los elementos son utilizados para representar cada patrón. Este tipo de códigos poseen en general menos dimensiones, pero a costa de sacrificar la facilidad de clasificación y sin una garantía de independencia entre dimensiones. Entre ambas situaciones aparecen los códigos ralos como una posibilidad de dar una solución óptima a estos problemas: no están afectados por una explosión combinatoria en su tamaño, aunque sin embargo son capaces de representar componentes separadas de los datos de manera directa [109]. En la Tabla 2.1 se presenta una comparación cualitativa de varios de los aspectos discutidos para los diferentes códigos. Desde el punto de vista de aprendizaje automático suele asociarse a los códigos locales con los métodos de aprendizaje de tipo competitivo (o basados en modelos de causa única), mientras que los distribuidos se asocian con los métodos cooperativos (o basados en modelos de múltiples causas). Otra vez los códigos ralos resultan de una combinación de las ventajas de am-

Tabla 2.1: *Comparación cualitativa entre los diferentes esquemas de codificación (adaptado de [50],(*) ver el texto para una discusión acerca de este aspecto).*

Tipo de código	Capacidad de representación	Capacidad de memoria	Facilidad de interpretación	Capacidad de generalización	Robustez al ruido
local	muy baja	limitada	muy fácil	ninguna	ninguna
ralo	alta	alta	fácil	buena	alta
denso	muy alta	baja	difícil	buena	muy alta(*)

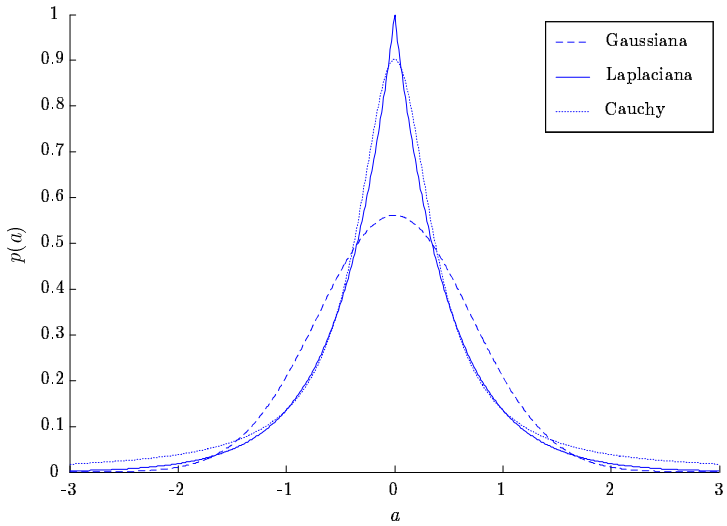

Figura 2.5: *Comparación entre una densidad de probabilidad gaussiana, una laplaciana y una de Cauchy. Para poder realizar la comparación las dos primeras poseen igual varianza, mientras que en el caso de Cauchy el parámetro de escala se ha fijado para aproximarse a la laplaciana debido a que su varianza resulta infinita. Obsérvese como los picos más agudos de la laplaciana y la de Cauchy, que exceden al de la gaussiana, se compensan con las colas más largas de ambas distribuciones.*

bos enfoques. Una ventaja de los códigos densos es su tolerancia al ruido, debido a su gran redundancia. Sin embargo mucha menos redundancia es suficiente para producir un comportamiento robusto y en muchos casos prácticos los códigos ralos pueden resultar inclusive más robustos que los densos [50]. Desde el punto de vista probabilístico los códigos distribuidos suelen asociarse con densidades de probabilidad gaussianas (de sus dimensiones o coeficientes), o incluso uniformes, debido a que casi siempre están activados. Los códigos locales, en el otro extremo, se asimilan a densidades tipo delta de Dirac. Por otro lado, los códigos ralos se asocian a densidades con picos importantes en cero, como las laplacianas o las de Cauchy (Ver Figura 2.5).

Un aspecto a considerar, según se desprende de esta discusión, es la cantidad de dimensiones de la representación. Si tenemos pocas dimensiones, las características tienden a "mezclarse" resultando difícil o hasta imposible generar códigos ralos e independientes sin sacrificar la completitud [22, 50]. Si utilizamos tantas dimensiones de salida como las presentes en los datos de entrada, estamos en el caso denominado *completo*. Dado que en general, en los sistemas reales, las causas que originan los datos suelen ser más que

las dimensiones de los mismos, se prefiere trabajar con códigos llamados *sobrecompletos* [146]. Ésto permite proyectar los datos en un espacio de características en el cual resultan ser, bajo determinadas condiciones, más fáciles de analizar y separar, o inclusive de eliminar el ruido [208].

De acuerdo a todo lo discutido, un código ralo –y posiblemente sobrecompleto– sería óptimo en términos de generalización y ajuste al propósito para este trabajo.

El problema del ruido

En la Sección 1.3.1 se ha revisado el concepto de ruido y los problemas que acarrea en el análisis de una señal de interés. El ruido es inherente a cualquier proceso de medición de alguna variable física. Por lo tanto este aspecto no debería despreciarse si se quiere modelar adecuadamente una señal. Un buen camino para intentar disminuir los efectos del ruido consiste en su inclusión explícita dentro del modelo considerado. Sin embargo existe un problema intrínseco a la definición de ruido, que está relacionado con su carácter subjetivo. La discriminación de cuál es la señal útil y cuál es el ruido depende estrictamente del criterio del observador, lo que puede ser difícil de incluir en un modelo. Se puede decir entonces que el ruido es una señal y por lo tanto transporta información, pero es información de algo en lo que no se está interesado. Por ejemplo, la variación entre los diferentes hablantes podría considerarse como ruido para un sistema de ASR, pero sería la información útil para un sistema de identificación del hablante. Varias de las medidas que se revisarán en la sección siguiente pueden resultar "ciegas" para considerar estos aspectos dependientes de la aplicación. Para evitar este inconveniente se requiere muchas veces incluir conocimiento específico de manera supervisada, es decir, una caracterización exhaustiva tanto de la señal como del ruido. Otra posibilidad consiste en utilizar un enfoque no supervisado pero que reúna algunas características de robustez intrínsecas como las discutidas en la sección anterior.

Otros aspectos a considerar

Se ha discutido acerca del ruido y de la dificultad de su inclusión en el modelo de una señal o sistema. Sin embargo ésta no es la única dificultad que puede aparecer en el proceso de modelización. Existen dificultades intrínsecas que pueden plantearse desde dos enfoques diferentes aunque complementarios: tomando como referencia el conocimiento interno del sistema o una determinada perspectiva de la respuesta a su entorno.

Para el primer enfoque, se puede comenzar suponiendo que se tiene acceso a las entradas y salidas del sistema bajo estudio. En un caso ideal es posible armar lo que se denomina un modelo tipo *caja blanca*, lo que significa que también se tiene acceso completo al "interior" del sistema. Entonces es posible conocer perfectamente sus principios de funcionamiento, su estructura y los valores de sus variables internas. Esta estrategia es posible solamente en el caso de modelos de sistemas relativamente sencillos, debido a nuestra capacidad limitada para "abarcar" todas las características de un sistema complejo. En el extremo opuesto aparecen los modelos tipo *caja negra*. Aquí todavía se tiene acceso a entradas y salidas pero se desconoce totalmente lo que ocurre en su interior. El caso intermedio son los modelos tipo *caja gris* donde el conocimiento interno es parcial

o impreciso. En los casos reales se trabaja en alguna de éstas dos últimas situaciones, y es necesario recurrir a técnicas de identificación de sistemas para estimar los parámetros internos desconocidos.

En lo anterior se supuso el acceso a las entradas y/o salidas de un sistema, aunque ésto no siempre es posible. Desde el enfoque alternativo importa además su relación con las *causas* y *efectos* del fenómeno considerado. Para modelar un fenómeno se trata, muchas veces, de inferir las causas que lo producen a partir de la observación de sus efectos. Ésto significa que la entrada del modelo corresponde a los efectos y su salida a las causas del fenómeno. Ésto se conoce como un *problema inverso*, que es diferente del *problema directo* en el cual los efectos (salida) se determinan a partir de las causas (entrada).

Los problemas directos son relativamente bastante más fáciles de encarar que los inversos. Para el caso determinístico existe una transformación única entre las causas y sus efectos, lo que constituye un aspecto clave para lograr un problema *bien condicionado* [71].

Un problema bien condicionado es aquel satisface los siguientes postulados (denominados de Hadamard [70]):

1. Existe una solución.

2. La solución es única.

3. La solución depende continuamente de los datos.

Para los problemas de este tipo las dificultades sólo aparecen cuando es necesario realizar simplificaciones excesivas o cuando no se conocen perfectamente las causas o entradas. Como contraste, los problemas inversos son generalmente *mal condicionados*, es decir que violan cualquiera de los postulados de Hadamard[11]. Estos problemas suelen ser muy sensibles al ruido o imprecisiones en los valores de sus parámetros debido a la falta de dependencia continua de los datos, lo que desemboca en soluciones inestables. Además pueden existir muchas explicaciones posibles para una observación particular. En estas circunstancias se debe considerar cual de las explicaciones resulta más *verosímil* (más cercana a la verdad). Todo ésto trae aparejadas complejidades adicionales que no aparecen en el problema directo equivalente. Como ejemplo de un problema directo se puede mencionar el de *síntesis de voz*: dado un mensaje de texto que se quiere comunicar y un conjunto de parámetros relacionados con las características del aparato fonador, es posible "calcular" la señal de voz resultante. El problema inverso correspondiente sería el de ASR, donde lo que se tiene es la señal de voz y lo que se desea averiguar es el mensaje contenido en esta señal. El estado de maduración relativo de ambos campos da una idea de las diferencias de complejidad de ambos problemas: mientras que es posible generar voz de muy buena calidad, todavía quedan sin resolver problemas importantes del ASR (como la falta de robustez mencionada en el capítulo anterior).

[11]Muchos de los modelos presentados en este trabajo son de este tipo. Desde el punto de vista de optimización se puede decir que son problemas *NP-completos* debido a que el tiempo requerido para encontrar una solución óptima crece de manera exponencial con el tamaño de los mismos.

2.4.2. Medidas de "calidad"

En la Sección 2.4.1 se han considerado algunas pautas generales para poder evaluar una representación, desde el punto de vista de la modelización: completitud, complejidad, generalización y ajuste al propósito. Además se han descripto someramente las ventajas de lograr una representación rala e independiente y su relación con las pautas mencionadas. Sin embargo no se ha discutido como realizar una valoración cuantitativa de la "calidad" de un modelo o representación en términos de estas características. En esta sección se presentarán diferentes alternativas para abordar este problema, algunas de las cuales se utilizarán para analizar los resultados del presente trabajo.

En lo que sigue el enfoque será principalmente discreto. Se considerará además que la señal $\mathbf{x} \in \mathbb{R}^N$ se genera mediante un modelo $\mathcal{M}$ con parámetros $\mathbf{\Phi} \in \mathbb{R}^{N \times M}$ en base a su representación interna $\mathbf{a} \in \mathbb{R}^M$, y afectado por el ruido ε en forma aditiva[12]:

$$\mathbf{x} = \mathcal{M}(\mathbf{\Phi}, \mathbf{a}) + \boldsymbol{\varepsilon} \, , \tag{2.16}$$

por lo que esta expresión se denomina ecuación del *modelo generativo*.

Existen varias formas para valorar una representación determinada, es decir la "bondad" de la codificación de los datos $\mathbf{x}$ en términos de los coeficientes $\mathbf{a}$, mediante un modelo $\mathcal{M}$ adecuado con parámetros $\mathbf{\Phi}$. Es posible dividirlas en dos grupos principales: aquellas originadas en la estadística o en la información presente en los coeficientes y las relacionadas con la dispersión de los coeficientes con respecto a alguna norma o distancia.

En el primer grupo es posible encontrar ejemplos como la curtosis o la entropía, que pueden ser útiles en el contexto de representaciones ralas o independientes. También puede calcularse el mínimo número de bits necesario para codificar la información contenida en los coeficientes, con un enfoque más orientado hacia la eficiencia de la codificación o la compresión de los datos.

En el segundo grupo están aquellas que utilizan alguna norma o medida de la dispersión promedio de los coeficientes $\mathbf{a}$ para un conjunto de datos. A pesar de denominarlas "normas" muchas de ellas no verifican estrictamente todas las propiedades necesarias, como por ejemplo la de homogeneidad o la desigualdad triangular [69].

Otras medidas pueden resultar útiles para cuantificar el grado de "ajuste" del modelo a los datos, como la denominada *evidencia*, o el error cuadrático medio (MSE). Además de las medidas de calidad consideradas en esta sección, existen otras específicas de cada aplicación. Ésto tiene que ver más con el aspecto de ajuste del modelo al propósito, discutido anteriormente. Por ejemplo, si la aplicación consiste en la clasificación de fonemas en base a la representación de la voz lograda por el modelo, entonces una buena medida de la calidad de la representación puede ser el porcentaje de fonemas bien clasificados. Este tipo de medidas también se discutirán y utilizarán en el capítulo de aplicaciones.

Medidas estadísticas y de información

Como ya se ha explicado, una característica importante para lograr una buena representación mediante un modelo es lograr la independencia estadística entre los coeficientes

[12] La utilización de una notación similar a la de las secciones anteriores resulta premeditada con el objeto de resaltar las conexiones entre los diferentes enfoques.

de la misma. A continuación se define que se entiende por independencia estadística en este contexto.

Definición 2.15 *Sea* $\mathbf{a} \in \mathbb{R}^M$ *el vector de los coeficientes de una representación determinada, cuya densidad de probabilidad es* $p(\mathbf{a})$*, entonces se dice que estos coeficientes son estadísticamente independientes si:*

$$p(\mathbf{a}) = \prod_{i=1}^{M} p(a_i) \ , \tag{2.17}$$

donde $p(a_i)$ *corresponde a la densidad marginal unidimensional respecto a la dimensión* i*-esima*[13]*.*

Se sabe que es posible capturar parte de la "estructura" de los datos a partir de las regularidades estadísticas de los mismos. Si se diseña un buen modelo estas regularidades son explotadas por éste y se traducen en su representación de la realidad. De aquí la importancia de medir diferentes aspectos de la estadística de la representación lograda.

La completitud de un modelo puede medirse en términos de información mutua $\mathcal{I}(\mathbf{x};\mathbf{a})$ entre los datos $\mathbf{x}$ y la representación $\mathbf{a}$ del modelo. Esta cantidad se denomina *transferencia de información* (de la entrada a la salida) del modelo. Si se supone que el sistema está libre de ruido, cualquier información a la salida puede provenir sólo de la entrada, entonces $\mathcal{H}(\mathbf{a}|\mathbf{x}) = 0$. De esta manera $\mathcal{I}(\mathbf{x};\mathbf{a}) = \mathcal{H}(\mathbf{a})$, y es posible maximizar la transferencia de información simplemente maximizando la entropía de la salida. Este es el fundamento de la técnica conocida como *infomax* [11]. Sin embargo estimar esta entropía puede resultar dificultoso en la práctica.

La complejidad de la representación producida por un modelo también puede ser medida en términos de su entropía $\mathcal{H}(\mathbf{a})$: suponiendo que cuanto más baja sea ésta, menor será la complejidad. Alternativamente puede utilizarse también $\mathcal{H}_q(\mathbf{a})$ para este propósito. La complejidad del modelo en sí mismo puede ser medida por la longitud (número de bits) de una descripción del mismo[14]. La combinación de ambas medidas en una sola expresión es la base del principio de *descripción de mínima longitud* (MDL) [162]. El problema subyacente de como "modelar el modelo" puede ser difícil, y requiere de algunas consideraciones que no se abordarán en el presente trabajo. Por ello se hará énfasis nuevamente en cuantificar la entropía de la representación.

Aquí aparece nuevamente el compromiso entre completitud y complejidad ya mencionado en la Sección 2.4.1, que se manifiesta ahora a través de la entropía de la representación. Una solución que no compromete a la completitud del modelo consiste en considerar la entropía de cada elemento de la salida en forma *separada*, y tratar entonces de minimizar su suma[15]. El mínimo, para una entropía total fija, resulta cuando cada

[13]Para las medidas de esta sección –relacionadas con la estadística y teoría de información– se sobreentiende que se trata de *v.a.*, aunque no se las denote expresamente así por razones de simplicidad. Por ejemplo, se ha reemplazado $\overline{\mathbf{A}}$ por $\mathbf{a}$ y se ha utilizado $p(\mathbf{a})$ en forma abreviada para $p_{\overline{\mathbf{A}}}(\mathbf{a})$.

[14]Ésto está sujeto a poder encontrar una manera adecuada de describirlo.

[15]Ésto resulta posible por la propiedad de sub-aditividad de la entropía de Shannon, por lo que no se aplica necesariamente a otros casos.

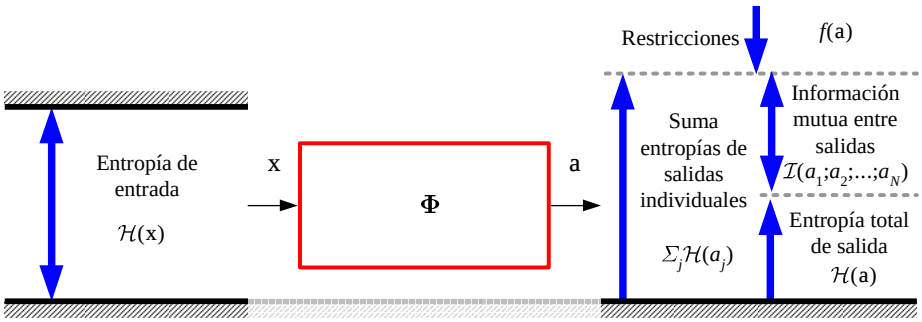

Figura 2.6: *Representación de la relación entre diversas medidas de entropía en un modelo o sistema (discreto). La entropía de la entrada es fijada por el ambiente, y provee un límite superior para la entropía total de la salida (en el caso sin ruido). La suma de las entropías de cada salida individual puede estar por encima o por debajo del nivel de entropía de entrada, pero siempre debe ser al menos tan grande como la entropía total de salida. Cualquier diferencia corresponde a la información mutua entre las salidas. Las restricciones imponen cierta presión para disminuir esta cantidad de manera de lograr, en el límite donde ambas líneas punteadas se juntan, un código factorial (adaptado de [71]).*

uno de sus elementos son estadísticamente independientes. Esta situación se discutió en la sección anterior y las representaciones con esta característica se denominan *códigos de baja entropía* [71]. De esta forma es posible utilizar la suma de la entropía $\mathcal{H}$ sobre los coeficientes individuales para medir la eficiencia de la codificación de $\mathbf{x}$:

$$\mathcal{H}_s(\mathbf{a}) = \sum_j \mathcal{H}(a_j) = -\sum_j \sum_i p_i(a_j) \log p_i(a_j) \ .$$

Al minimizar la suma de las $\mathcal{H}(a_j)$ se "destruye" también la información mutua $\mathcal{I}$ entre los coeficientes a_j, lo que termina, con las restricciones adecuadas, en un código con coeficientes estadísticamente independientes (también denominado *código factorial*) [71] (Ver Figura 2.6). Como se discutió anteriormente es deseable también obtener códigos ralos y ambos criterios pueden aplicarse simultáneamente con buenos resultados [179]. Para valorar este aspecto introduciremos algunas medidas especiales.

Es preciso tener cuidado con $\mathcal{H}$ como estimador de la eficiencia de la codificación ya que no incluye ninguna "medida de desajuste". Es decir que los datos podrían no representarse muy bien con el modelo y sin embargo tener una entropía baja. Si el modelo en consideración genera totalmente los datos (es decir que es completo) entonces $\mathcal{H}$ es una medida razonable del costo de la codificación [116]. Ésto se puede controlar estimando el error de aproximación del modelo. Para independizarse un poco de la cantidad de dimensiones en este trabajo se utiliza el promedio de las entropías para cada dimensión de $\mathbf{a}$.

Otro aspecto importante a tener en cuenta está dado por el costo de la codificación en bits. Puede utilizarse $\mathcal{H}$ para calcularlo por medio de la menor cantidad de bits requerida para codificar los patrones [188]:

$$\#bits \geq \mathcal{H}_{bits}(a_j) = -\sum_i p_i(a_j) \log_2 p_i(a_j) \ .$$

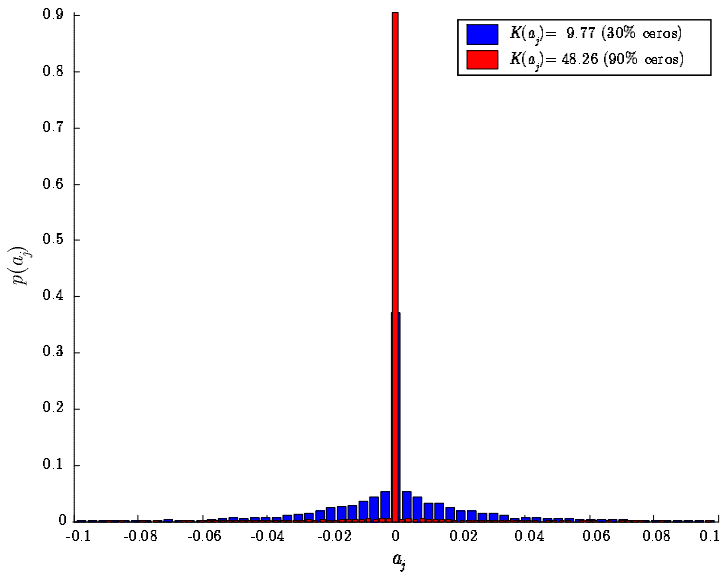

Figura 2.7: *Histograma y curtosis de dos distribuciones de valores ralas. Los datos fueron generados mediante una distribución mixta formada por una delta de dirac en cero más una exponencial simétrica, ésto permite controlar fácilmente el porcentaje de valores exactamente iguales a cero.*

Esta forma de calculo sufre de similares desventajas que $\mathcal{H}$. Para evitarlas existen otras maneras de estimar este costo, por ejemplo si se conoce la relación entre la entrada y los parámetros del modelo es posible estimar [116]:

$$\#bits \geq -\log_2 P(\mathbf{x}|\mathbf{\Phi}) - N \log_2 (\sigma_\varepsilon) \ ,$$

donde σ_ε es la desviación estándar del ruido aditivo. Sin embargo la exactitud de la estimación depende de la de $-\log_2 P(\mathbf{x}|\mathbf{\Phi})$ que puede ser complejo de estimar, por lo que en este trabajo se utilizará la expresión anterior.

Una de las formas de cuantificar la dispersión desde el punto de vista estadístico es la de utilizar el momento de 4º orden o curtosis. La curtosis $\mathcal{K}$ resulta una buena medida de dispersión para distribuciones simétricas unimodales [71]. Si se toma ahora a_j como $v.a.$:

$$\mathcal{K}(a_j) \triangleq \frac{\mathcal{E}\left[a_j^4\right]}{\mathcal{E}\left[a_j^2\right]^2} - 3 \ .$$

En la Figura 2.7 se puede apreciar la variación de esta cantidad en relación con el número de valores iguales a cero de un coeficiente a_j con diferentes distribuciones de probabilidad. La curtosis generalmente aumenta cuando la entropía disminuye por lo que a veces se la puede relacionar con la independencia (aunque es necesario tener cuidado con esta aproximación). Suele usarse como medida de "no gaussianidad". Si el valor es positivo se habla de distribuciones supergaussinas, y subgaussianas si es negativo. El problema principal es que es muy propensa a los *valores fuera de rango* (en inglés *outliers*). En este trabajo se utiliza el promedio de la curtosis en todas las dimensiones del vector $\mathbf{a}$.

Medidas de dispersión

Una medida obvia de la dispersión o "raleza" de una representación es simplemente el número de términos distintos de cero dentro del vector. Ésto es lo que hace la *norma cero* o ℓ_0, es decir cuenta el número de coeficientes no nulos [40]:

$$\|\mathbf{a}\|_0 \triangleq \#\{j : a_j \neq 0\} \ .$$

Sin embargo esta norma es muy poco robusta frente a pequeñas perturbaciones de los elementos considerados como "cero". En la práctica el "cero" se define en función de algún umbral pequeño θ, debido a que si no la norma es muy "severa". La versión aquí utilizada está normalizada con la dimensión del vector, y θ se escoge en función del valor máximo de los coeficientes de todos los patrones. Ésto provee cierta invarianza a la escala de los mismos.

Si se considera la norma ℓ_q:

$$\|\mathbf{a}\|_q \triangleq \left(\sum_j |a_j|^q\right)^{1/q} \ ,$$

con $0 < q \leq 1$, es posible apreciar que cuanto más pequeño se seleccione q más énfasis se pone en una representación más rala. En la Figura 2.8 se pueden apreciar los valores que toma la norma ℓ_q para diferentes valores de q en un espacio bidimensional de los coeficientes $\mathbf{a}$.

De hecho,

$$\lim_{q\to 0} \|\mathbf{a}\|_q^q = \|\mathbf{a}\|_0 \ .$$

Por lo anterior ℓ_1 se utiliza frecuentemente como aproximación práctica para la norma ℓ_0, especialmente en problemas de optimización [40, 21]. Minimizar con respecto a esta norma es análogo a requerir que los coeficientes tengan una distribución de probabilidad *a priori* laplaciana:

$$p(a_j) = \frac{1}{2\sigma}\, e^{\frac{-|a_j|}{\sigma}} \ ,$$

lo que conecta este tipo de medidas con las estadísticas presentadas en la sección anterior.

La *norma de mínimo volumen* [69] es también extremadamente "severa". La idea es similar a ℓ_0 ya que consiste también en contar las componentes del vector distintas de cero, lo que analíticamente se puede expresar de la siguiente forma: Sea $\theta > 0$ un número pequeño, del orden de la precisión de la máquina. Entonces:

$$\frac{a_j^2}{a_j^2+\theta} \approx \begin{cases} 0 & \text{si } a_j \approx 0 \\ 1 & \text{si } a_j \neq 0,\ a_j \gg \theta \\ > 0, < 1 & \text{de otra forma .} \end{cases}$$

De aquí, la definición de la norma:

$$minvol\,(\mathbf{a}) \triangleq \sum_{j=1}^{M} \frac{a_j^2}{a_j^2+\theta} \ .$$

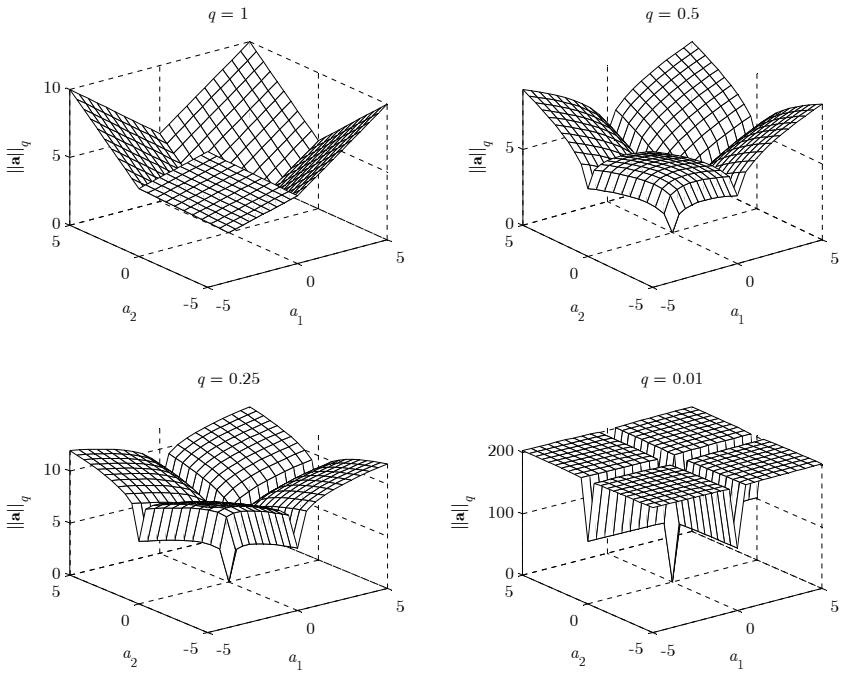

Figura 2.8: *Valores que toma la norma ℓ_q para diferentes valores de q en un espacio bidimensional. Es posible apreciar como a medida que q se hace más chico, la norma resulta más severa asignando valores más pequeños sólo a los puntos más cercanos al origen.*

Se puede decir que *minvol* constituye una aproximación a ℓ_0 cuando se utiliza un umbral pequeño para decidir que coeficientes se consideran diferentes de cero.

Como otro criterio posible aparece la *norma de Cauchy*, que se ha utilizado con el objeto de resolver problemas de procesamiento de señales. En este caso se supone que la distribución a priori de los coeficientes de la representación corresponde a una distribución de Cauchy, con parámetro de escala σ_c (que permite ajustar la severidad):

$$p(a_j) = \frac{1}{2\pi\sigma_c^2} \frac{1}{1 + \frac{a_j^2}{2\sigma_c^2}} \; .$$

De forma similar al caso de función de densidad laplaciana, su larga "cola" da origen a una distribución rala de los parámetros, cuya norma de medición es:

$$S\left(\mathbf{a}\right) \triangleq \sum_{j=1}^{M} \log_2 \left(1 + \frac{a_j^2}{2\sigma_c^2}\right) . \tag{2.18}$$

En [69] se extiende el empleo de esta norma para encontrar soluciones ralas de sistemas lineales indeterminados de aplicación general similares a los planteados en este trabajo.

También puede mencionarse la *norma D*, definida como el cociente entre el máximo valor absoluto de las componentes y el módulo del vector:

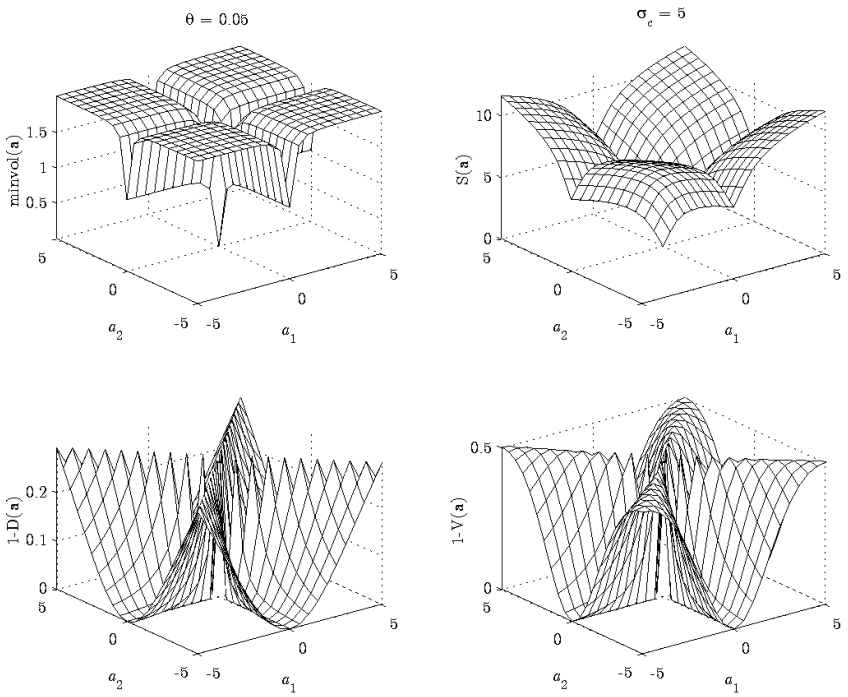

Figura 2.9: *Valores que toman las diferentes normas en un espacio bidimensional: minvol (arriba-izquierda), Cauchy (arriba-derecha), D (abajo-izquierda) y varimax (abajo derecha). Comparar con la Figura 2.8.*

$$D\left(\mathbf{a}\right) \triangleq \frac{\max\limits_{j=1,\ldots,M} |a_j|}{|\mathbf{a}|} \ ,$$

Esta norma toma valores próximos a 1 cuando hay una componente que predomina sobre las demás. Maximizar $D(\mathbf{a})$ (o minimizar $1 - D(\mathbf{a})$) es un posible camino de búsqueda de soluciones con pico en una sola componente, aunque las demás no tomen valores totalmente próximos a cero.

Un criterio similar lo provee la *norma varimax*:

$$V\left(\mathbf{a}\right) \triangleq \frac{\sum\limits_{j=1}^{M} a_j^4}{\left(\sum\limits_{j=1}^{M} a_j^2\right)^2} \ ,$$

Obsérvese la similitud con la expresión de la curtosis $\mathcal{K}(a_j)$.

En la Figura 2.9 se puede apreciar una comparación entre los valores que toman las distintas normas consideradas en esta sección en un espacio bidimensional.

Una manera práctica e intuitiva de comparar la dispersión de distintas representaciones consiste en graficar la media de los coeficientes ordenados por su magnitud (normalizada con el valor máximo). De esta forma, cuanto más rala sea la representación

más rápida será la caída a medida que se avanza en el sentido del orden relativo de los coeficientes.

Medidas de ajuste

Casi todas las medidas mencionadas se aplican sólo sobre los coeficientes de la representación, por lo tanto no dan cuenta del grado de aproximación o ajuste de los datos mediante el modelo considerado. Por ello es necesario definir algunas medidas para este propósito.

El valor esperado del *error cuadrático medio* (MSE) de la reconstrucción de los datos mediante el modelo puede dar una idea del ajuste que se requiere:

$$MSE(\varepsilon) \triangleq \mathcal{E}[\|\mathbf{x} - \mathcal{M}(\mathbf{\Phi}, \mathbf{a})\|_2] \ ,$$

Su valor está acotado inferiormente por la varianza del ruido σ_ε del modelo generativo.

Otra forma consiste en estimar $P(\mathbf{x}|\mathbf{\Phi})$ o sea la evidencia o probabilidad de los datos dado el modelo, es decir cuan bien describe este modelo a los datos. Más adelante se mostrará como estimar esta cantidad para un caso determinado.

2.5. Análisis estadístico de datos

Existe una familia de técnicas que permiten realizar un análisis estadístico de los datos con el fin de extraer de ellos algunas "componentes interesantes". Para resolver este problema se ha propuesto una variedad de métodos de transformación acompañados con ciertos criterios de optimización que determinan su funcionamiento y utilización. Entre los métodos se pueden mencionar el *análisis de componentes principales* (PCA), el *análisis de factores* (FA), la *búsqueda de proyecciones* (PP), y más recientemente el *análisis de componentes independientes* (ICA). Entre los criterios de optimización empleados se pueden citar los que buscan una reducción de la dimensión de los datos, obtener una mayor simplicidad con la transformación empleada, ajustarlos a algunas propiedades estadísticas significativas, o bien criterios orientados a alguna aplicación específica.

Varias de estas técnicas se pueden plantear desde dos enfoques diferentes pero con idénticos o similares resultados. Ésto es suponiendo un modelo estadístico que genera los datos, o bien un modelo determinístico más relacionado con el enfoque de análisis de señales discutido en la Sección 2.2. Este último enfoque permite verlas como técnicas de análisis de señales donde la base o el diccionario no es fijo sino que se adapta con la o las señales a analizar[16]. En este sentido lo que se denominan componentes en un contexto puede asimilarse con los átomos o elementos de una base o diccionario en el otro contexto. Por supuesto que es posible cambiar el enfoque en cualquier momento, y ésto se aprovechará especialmente cuando facilite la interpretación de algún resultado, o bien para presentar algoritmos específicos. En esta sección se utilizará mayoritariamente

[16]Este es el caso contrario al de la mayoría de las técnicas de análisis de señales descriptas en el Capítulo 4 donde el diccionario se fija de antemano.

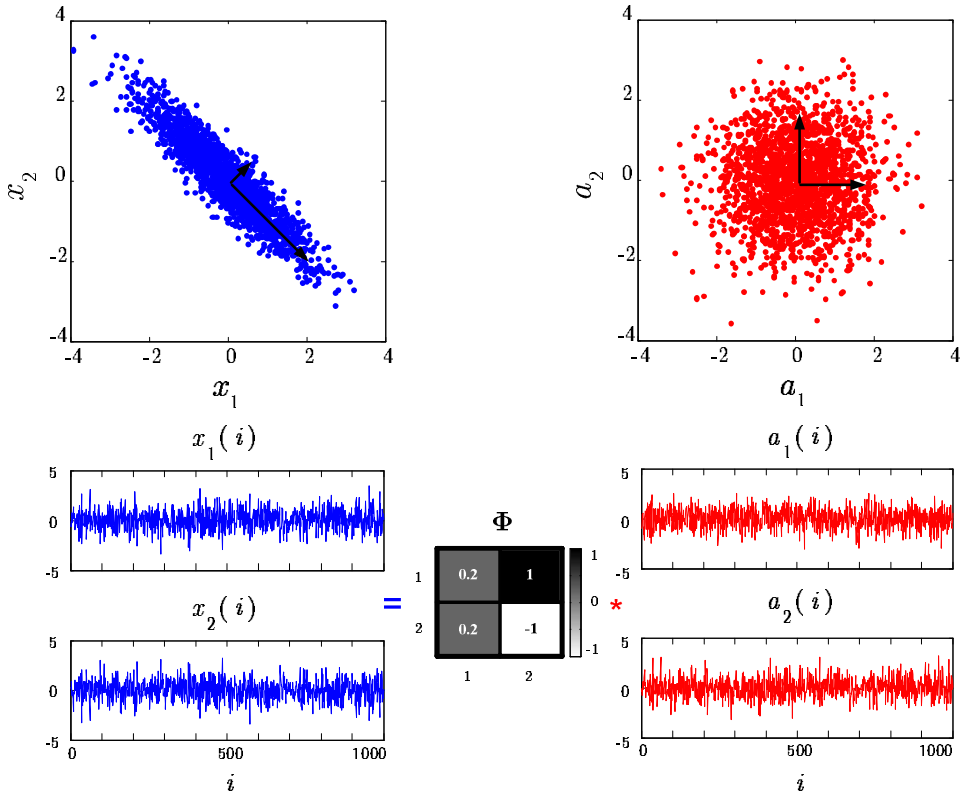

Figura 2.10: *Ilustración del modelo generativo para PCA: Datos gaussianos correlacionados* **x** *(arriba izquierda), fuentes o causas gaussianas no correlacionadas* **a** *(arriba derecha) y vectores que apuntan en las direcciones de máxima varianza en ambos casos. Muestras de los datos, la matriz de mezcla, y las fuentes (abajo). Obsérvese que las columnas de la matriz de mezcla se corresponden con las direcciones de máxima varianza de los datos y son ortogonales.*

el primer enfoque para presentar las técnicas de PCA e ICA que serán útiles para el desarrollo de este trabajo. Para una revisión con mayor detalle puede consultarse la bibliografía específica disponible [85, 163].

2.5.1. Análisis de componentes principales

La técnica de PCA [151] constituye uno de los métodos estadísticos clásicos más populares para análisis de datos, extracción de características y compresión. Es posible formularla en términos de un modelo estadístico que supone que los datos **x** han sido generados a partir de una mezcla lineal **Φ** de un conjunto de causas o fuentes **a**, las componentes principales propiamente dichas, con distribuciones gaussianas (y media cero) y sin la influencia de ruido externo. Se supone también que las columnas de la matriz de mezcla son ortogonales entre si, y que corresponden precisamente a las direcciones de las componentes principales (Ver Figura 2.10). Para ello se reescribe la ecuación (2.16) del modelo generativo como sigue:

$$\mathbf{x} = \mathbf{\Phi a} \ , \tag{2.19}$$

donde el vector de datos $\mathbf{x} \in \mathbb{R}^N$, el vector de fuentes $\mathbf{a} \in \mathbb{R}^M$ y la matriz de mezcla $\mathbf{\Phi} \in \mathbb{R}^{N \times M}$, con $M \leq N$.

El objetivo del método consiste precisamente en encontrar las componentes $\mathbf{a}$ que mejor expliquen los datos $\mathbf{x}$ en el sentido de las direcciones de máxima varianza. Ésto puede definirse de una forma intuitiva utilizando la siguiente formulación recursiva [86]. Sea $\mathbf{w_1}$ la dirección de la primer componente principal:

$$\mathbf{w}_1 = \underset{\|\mathbf{w}\|_2=1}{\arg\max} \, \mathcal{E}\left[\left(\mathbf{w}^T\mathbf{x}\right)^2\right] \ , \tag{2.20}$$

donde $\mathbf{w}_1$ es un vector columna de la misma dimensión que $\mathbf{x}$. De esta forma la primer componente principal corresponde a la proyección en la dirección en la cual su varianza se maximiza. Habiendo determinado las primeras $k-1$ componentes principales, la k-ésima componente principal se determina como la componente principal del residuo:

$$\mathbf{w}_k = \underset{\|\mathbf{w}\|_2=1}{\arg\max} \, \mathcal{E}\left[\left(\mathbf{w}^T\left\{\mathbf{x} - \sum_{m=1}^{k-1} \mathbf{w}_m\mathbf{w}_m^T\mathbf{x}\right\}\right)^2\right] \ ,$$

Finalmente las componentes principales están dadas ahora por $a_m = \mathbf{w}_m^T\mathbf{x}$.

En la práctica el cálculo de las componentes principales se realiza mediante el siguiente procedimiento que resulta más eficaz. Inicialmente, y para evitar ambigüedades de magnitud intrínsecas al planteo del problema, se restringe a que las fuentes posean varianza unitaria. La matriz de mezcla $\mathbf{\Phi}$ puede ser escrita entonces en términos de su *descomposición en valores singulares* (SVD) [80], como:

$$\mathbf{\Phi} = \mathbf{Q}\mathbf{\Lambda}\mathbf{P}^T \ ,$$

donde $\mathbf{Q} \in \mathbb{R}^{N \times M}$ y $\mathbf{P} \in \mathbb{R}^{M \times M}$ son matrices con columnas ortogonales y $\mathbf{\Lambda} \in \mathbb{R}^{M \times M}$ es una matriz diagonal.

Debido a las propiedades de $\mathbf{P}$ es posible reescribir una expresión alternativa más simple para $\mathbf{\Phi}$ y sin pérdida de generalidad como:

$$\mathbf{\Phi} = \mathbf{Q}\mathbf{\Lambda} \ .$$

Una de las características más importantes de PCA es que se puede estimar $\mathbf{\Phi}$ sólo contando con la estadística de segundo orden. Para ello es necesario estimar la matriz de covarianza de los datos:

$$\mathbf{\Sigma_x} = \ \mathcal{E}\left[\mathbf{x}\mathbf{x}^T\right] \ .$$

Dado que las fuentes gaussianas mezcladas en forma lineal producen a su vez datos con distribución gaussiana, es posible escribir también que:

$$\mathbf{\Sigma_x} = \mathbf{\Phi}\mathbf{\Sigma_a}\mathbf{\Phi}^T = \mathbf{Q}\mathbf{\Lambda}^2\mathbf{Q}^T \ ,$$

donde $\mathbf{\Sigma_a}$ es la matriz de covarianza de las fuentes.

Ésto puede expresarse en la forma de una ecuación de *autovalores*:

$$\mathbf{\Sigma_x}\mathbf{Q} = \mathbf{Q}\mathbf{\Lambda}^2 \ ,$$

donde las columnas de $\mathbf{Q}$ son los autovectores y corresponden a las componentes principales de los datos, y los autovalores miden la varianza de cada una de las M fuentes. Generalmente éstos se ordenan en orden decreciente lo que permite resolver la ambigüedad de orden de las columnas para la solución final.

Si se quiere recuperar las fuentes $\mathbf{a}$ a partir de los datos $\mathbf{x}$, basta con multiplicarlos[17] por $\mathbf{W} = \mathbf{\Lambda}^{-1}\mathbf{Q}^T$. De esta forma los mismos se decorrelacionan lo cual equivale, en este caso gaussiano, a encontrar las componentes estadísticamente independientes. Además la multiplicación por $\mathbf{\Lambda}^{-1}$ normaliza también las varianzas. Este proceso suele denominarse *blanqueo* o *esferización* de los datos y es frecuente su utilización previa a la aplicación de otras técnicas más elaboradas como la de ICA que veremos en la sección siguiente porque "remueve" la estadística de segundo orden.

2.5.2. Análisis de componentes independientes

Como se ha visto, el éxito de PCA en encontrar las componentes significativas de los datos depende de que todas las fuentes posean una distribución gaussiana[18]. La técnica de ICA surge como una alternativa para lograr una representación completamente independiente de los datos en condiciones más generales, mediante una transformación adecuada que solucione una ecuación análoga a la (2.19). Se puede decir que se trata no de un solo algoritmo, sino más bien de una familia de algoritmos para solucionar problemas que se pueden escribir de manera "similar". Debido a que el objetivo de ICA consiste precisamente en encontrar una representación de nuestra señal en términos de un conjunto de coeficientes estadísticamente independientes el criterio de optimización empleado consiste en tratar de maximizar esta independencia en forma directa o indirecta.

Considerando las variables $\mathbf{x}$ y $\mathbf{a}$ como *v.a.* con media cero, es posible definir los siguientes modelos ICA:

Modelo General o Clásico: Consiste en encontrar una transformación lineal $\mathbf{a} = \mathbf{W}\mathbf{x}$ tal que las fuentes a_i sean tan independientes como sea posible en el sentido de maximización de alguna función objetivo $G = g\left(a_1, \cdots, a_M\right)$ que mide dicha independencia. Ésta es la definición más general en el sentido de que no hace suposiciones con respecto a los datos; resulta un poco vaga debido a que se debe definir también una medida de independencia para las componentes a_i.

Modelo ICA libre de ruido: Consiste en estimar el modelo generativo $\mathbf{x} = \mathbf{\Phi}\mathbf{a}$, donde las variables latentes a_i en el vector $\widehat{a} = \left(a_1, \cdots, a_M\right)^{\mathrm{T}}$ se suponen independientes. La matriz de mezcla $\mathbf{\Phi}$ es constante y $\in \mathbb{R}^{N\times M}$. Para encontrar las fuentes se utiliza finalmente la inversa de la matriz de mezclas $\mathbf{W} = \mathbf{\Phi}^{-1}$. Este modelo fue introducido por Jutten y Herault en 1991 [96], y se puede ver como la primera formulación explícita de ICA.

[17] Es aconsejable trabajar con datos que posean media cero o bien sustraerles la media como paso previo.

[18] Aunque ésto no resulta una limitación importante en algunas aplicaciones también se requiere que se cumplan las restricciones de ortogonalidad sobre las columnas de la matriz de mezcla.

Modelo ICA con ruido: Consiste en estimar el siguiente modelo generativo: $\mathbf{x} = \mathbf{\Phi a} + \boldsymbol{\varepsilon}$, donde $\mathbf{a}$ y $\mathbf{\Phi}$ son los mismos que en el caso anterior y $\boldsymbol{\varepsilon}$ es un vector de ruido aleatorio N-dimensional. Esta definición reduce el problema a la estimación de un modelo de variables latentes que no resulta sencillo de resolver. Una posible vía para encarar su solución consiste en incluir restricciones adicionales para que la representación lograda en términos de las fuentes a_i sea rala además de independiente [118, 116].

Modelo ICA sobrecompleto: Consiste en solucionar un modelo ICA pero cuando $M > N$. En los modelos anteriores el número de mezclas lineales observadas N debía ser al menos tan grande como el número de componentes independientes M, es decir, $N \geq M$. En el caso de $N < M$ o sobrecompleto es aún posible identificar la matriz de mezcla, pero las realizaciones de las fuentes independientes no son directamente identificables debido a que la matriz $\mathbf{\Phi}$ no se puede invertir [14]. En este caso se puede encontrar una matriz $\mathbf{W}$ pero que depende en realidad de $\mathbf{x}$ y de $\mathbf{\Phi}$, lo que lleva a una relación bastante más compleja. Hay una serie de trabajos que tratan este último caso denominado ICA con bases sobrecompletas (o *mayor cantidad de fuentes que de mezclas* en terminología clásica de ICA), y también aquí aparecen relaciones importantes con las representaciones ralas [14, 146, 118, 116, 88].

Estos últimos dos modelos resultan de especial interés para la solución de algunos de los problemas considerados en esta tesis, en particular en la versión que contempla simultáneamente el caso sobrecompleto y con ruido. Por ello se tratarán con mayor detalle en los capítulos posteriores, especialmente en su relación con las representaciones ralas.

En el trabajo de Comon [27] se imponen restricciones adicionales al supuesto de independencia que determinan las denominadas *condiciones de identificabilidad* del modelo ICA:

1. Todas las componentes independientes a_i deben ser no gaussianas, salvo una como máximo.

2. La matriz de mezclas $\mathbf{\Phi}$ debe ser de rango completo (en las columnas).

Como ya se mencionó, para el caso de variables aleatorias gaussianas, independencia estadística implica no correlación, por lo que cualquier representación que las decorrelacione dará hasta cierto grado componentes independientes. Es por eso que sólo se puede admitir hasta un miembro gaussiano dentro del conjunto a_i.

Las relaciones entre PCA e ICA son significativas. Ambos métodos formulan una función objetivo general, que define una cualidad "interesante" para la representación lineal de la señal, para luego maximizarla. Ambas metodologías plantean hipótesis "contradictorias" de gaussianidad y no gaussianidad respectivamente. Además PCA utiliza estadística de segundo orden y enfatiza la reducción de la dimensión, mientras que ICA requiere la utilización de estadística de orden mayor y la reducción de la dimensión no

es su objetivo fundamental, por lo que ésta puede mantenerse o incluso aumentarse. En el caso de *PCA no lineal* existe todavía una relación mayor [140, 141].

Como PCA supone distribuciones gaussianas, tiene problemas para encontrar las "componentes principales" si los datos fueron generados mediante distribuciones con curtosis positiva grande. Tampoco está definida la solución para el caso sobrecompleto. En la Figura 2.11 se puede ver una comparación entre la solución de un problema en dos dimensiones por PCA e ICA, para los casos completo y sobrecompleto.

Funciones objetivo para ICA

La estimación del modelo ICA se realiza formulando una función objetivo, y utilizando un algoritmo de optimización para minimizarla o maximizarla. A esta función se la denomina también como *función de contraste*, de pérdida o de costo, según los diferentes autores y métodos. Siguiendo a [86] aquí se utilizará la primera denominación con idéntico significado y se describirán brevemente los casos más usuales. Para ello se realizará una distinción entre los casos de evaluación de *múltiples unidades* en forma simultánea, y los casos de evaluación de una *única unidad* por vez en forma secuencial. Para una revisión más exhaustiva remitirse a [85, 163].

Funciones de contraste multi-unidad Estas funciones estiman todas las componentes independientes simultáneamente.

1) *Verosimilitud*: En el modelo libre de ruido es posible formular la verosimilitud de los datos y después estimar el modelo mediante un método de *máxima verosimilitud* [86]. Sea $\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_M)^T = \mathbf{\Phi}^{-1}$ como en la sección anterior, entonces la verosimilitud logarítmica se determina como:

$$\mathcal{L}(\mathbf{W}) = \sum_{i=1}^{I} \sum_{m=1}^{M} \log p_m \left(\mathbf{w}_m^T \mathbf{x}(i) \right) + I \ln |\det \mathbf{W}| \, ,$$

donde p_m es la función de distribución de probabilidad de las componentes independientes que se supone conocida, y $\mathbf{x}(i), i = 1, \cdots, I$ corresponde a las diferentes realizaciones de las variables observadas. Por lo anterior este método posee la desventaja de que debe tenerse conocimiento de las densidades de probabilidad de las componentes independientes y, aunque éstas pueden estimarse, el proceso resulta generalmente bastante complicado.

2) *Entropía de una red neuronal*: Otra posibilidad consiste en utilizar una función de contraste no lineal similar a las utilizadas en las *redes neuronales artificiales*[19] [12]. De esta forma se trata de maximizar la entropía de esta red neuronal (este principio denominado *infomax* ya se mencionó en la Sección 2.4.2). Sea $\mathbf{x}$ la entrada a una red neuronal cuyas salidas son de la forma $g_m \left(\mathbf{w}_m^T \mathbf{x} \right)$, donde g_m son funciones escalares no

[19] Una red neuronal artificial es un sistema de procesamiento de información o señales compuesto por un gran número de elementos simples de procesamiento, llamados neuronas artificiales o simplemente nodos. Dichos nodos están interconectados por uniones directas llamadas conexiones y cooperan para realizar procesamiento en paralelo con el objetivo de resolver una tarea computacional determinada.

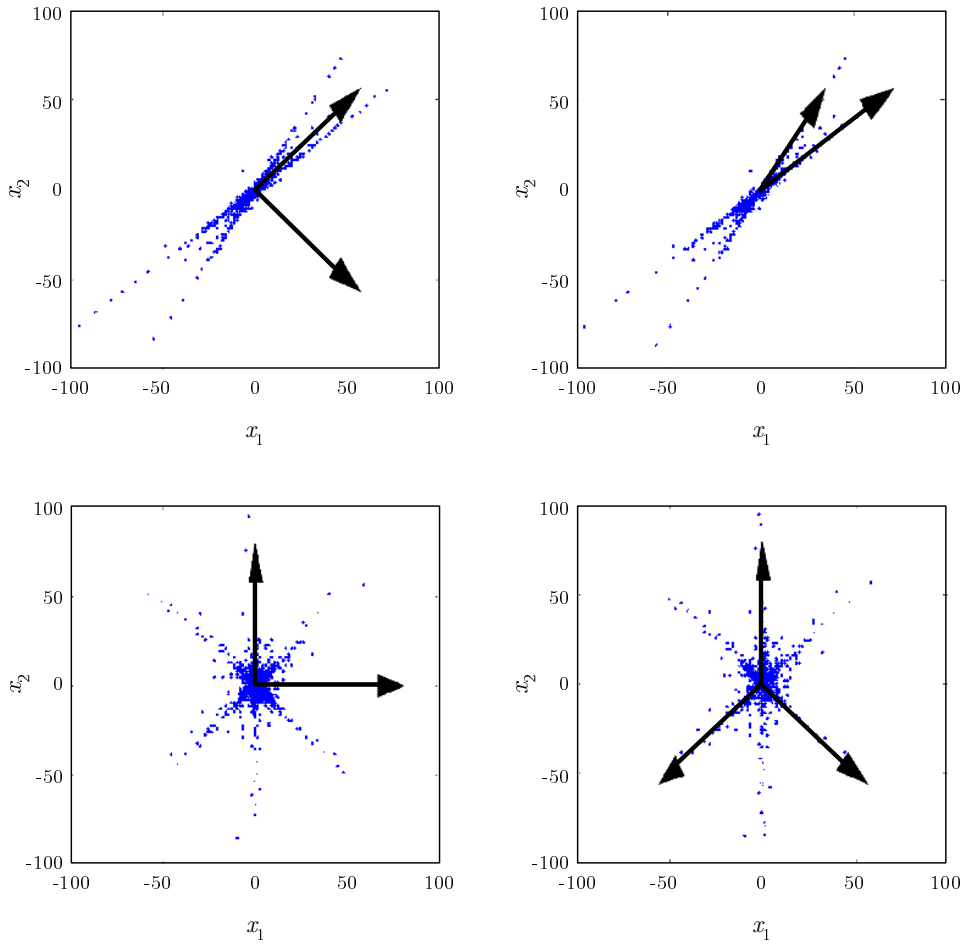

Figura 2.11: *Ilustración de vectores base correspondientes a las columnas de* $\mathbf{\Phi}$ *en un espacio de datos bidimensionales con dos (arriba) o tres (abajo) fuentes ralas: PCA encuentra vectores base ortogonales (arriba izquierda), la representación ICA encuentra vectores base independientes (arriba derecha), ICA no puede modelar adecuadamente la distribución de los datos con tres fuentes pero (abajo izquierda), la representación ICA sobrecompleta encuentra 3 vectores base que se ajustan a la distribución de los datos (abajo derecha; adaptado de [111]).*

lineales y los $\mathbf{w}_m$ son vectores de peso de las neuronas. Se desea maximizar la entropía de las salidas:

$$\mathcal{L}'(\mathbf{W}) = \mathcal{H}\left(g_1\big(\mathbf{w}_1^T\mathbf{x}\big), \cdots, g_M\big(\mathbf{w}_M^T\mathbf{x}\big) \right) .$$

Es posible demostrar que el principio de máxima entropía aplicado sobre esta red es equivalente a la estimación de máxima verosimilitud anterior, si las funciones g_m se eligen como las funciones de distribución acumulada [86]. Una desventaja es que resulta muy sensible a los datos fuera de rango para algunas formas de las funciones de distribución de probabilidad de las componentes independientes.

3) *Información mutua*: Según se discutió en la Sección 2.4.2 también es posible utilizar para este propósito el concepto de MI, que toma en cuenta la estructura completa de la dependencia de las componentes. Como se explicó su valor es cero sólo en el caso de que las componentes sean estadísticamente independientes. Por lo tanto el encontrar la transformación que minimice la MI entre las componentes a_m es una manera natural de estimar el modelo ICA [27]. Como se puede inferir de las relaciones presentadas en la Sección 2.3.4 este planteo equivale a determinar la divergencia Kullback-Leibler de la función de distribución de probabilidad conjunta de $\mathbf{a}$, y el producto de las funciones de distribución de probabilidad marginales sobre los a_m. Los problemas de utilizar MI se basan en la dificultad de su estimación que requiere el empleo de *cumulantes de alto orden*[20].

4) *Correlaciones cruzadas no lineales*: Se ha utilizado también para obtener las componentes independientes el principio de cancelación de correlaciones cruzadas no lineales [86]:

$$\mathcal{E}[g_1(a_i)g_2(a_j)] ,$$

donde g_1 y g_2 son funciones no lineales impares adecuadamente escogidas.

5) *PCA no lineal*: Una aproximación a la obtención de las componentes independientes la dan los criterios de PCA no lineales con la idea de introducir la no linealidad en la función objetivo utilizada para la versión recursiva de PCA (2.20) [86]:

$$\mathbf{w}_1 = \underset{\|\mathbf{w}\|_2=1}{\arg\max}\ \mathcal{E}\Big[g\big(\mathbf{w}^T\mathbf{x}\big)^2\Big] .$$

Si la $g(\cdot)$ se elige como una función no lineal de las distribuciones de probabilidad de las componentes independientes y los datos han sido previamente blanqueados, entonces esta versión no lineal estima las componentes independientes.

Funciones de contraste de unidad única Son funciones cuya optimización permite la estimación de una componente independiente a la vez, permitiendo encontrar las restantes por medio de iteraciones. Resulta conveniente estimar las componentes independientes de esta manera cuando no se requiere conocer su número de antemano.

[20]En este contexto cuando se habla de estadística de alto orden se refiere a orden mayor a dos, por ejemplo el cumulante de tercer orden para un *proceso aleatorio discreto estacionario* con media cero $x[n]$ está dado por $C_{3x} = \mathcal{E}[x[n]x[n+k]x[n+l]]$, con $0 \leq l \leq k \leq \infty$.

1) *Negentropía*: Es posible utilizar la negentropía como función de contraste debido a su propiedad ya mencionada de anularse para el caso gaussiano. El problema con su determinación radica otra vez en estimar la función de distribución de probabilidad de las fuentes por lo que se pueden utilizar métodos de aproximación como los siguientes:

Aproximación utilizando momentos estadísticos de alto orden:

La forma clásica de aproximar la negentropía es utilizar momentos de alto orden como [90]:

$$j(\mathbf{y}) \approx \frac{1}{12}\ \mathcal{E}\left[\mathbf{y}^3\right]^2 + \frac{1}{48}\mathcal{K}(\mathbf{y})^2\ ,$$

donde la variable aleatoria $\mathbf{y} = \mathbf{w}^T\mathbf{x}$ se supone con media cero y varianza uno. La validez de esta estimación es limitada ya que se ha demostrado que las aproximaciones basadas en cumulantes para la negentropía son inexactas y en muchos casos sensibles a datos fuera de rango [84].

Aproximación utilizando el principio de máxima entropía:

Debido a las limitaciones anteriores se desarrollaron nuevas aproximaciones como las basadas en el principio de máxima entropía [90]:

$$j(\mathbf{y}) \approx \sum_{i=1}^{P} k_i \left\{\ \mathcal{E}[G_i(\mathbf{y})] - \ \mathcal{E}[G_i(\boldsymbol{\nu})]\right\}\ ,$$

donde P es el orden de la aproximación, k_i son constantes mayores que cero y $\boldsymbol{\nu}$ es una variable aleatoria gaussiana de media cero y varianza uno; G_i es prácticamente cualquier función no cuadrática.

2.6. Comentarios de cierre del capítulo

En este capítulo se han revisado varias técnicas que servirán de base para el desarrollo de este trabajo. Aunque muchas de ellas están íntimamente relacionadas con el *reconocimiento de patrones* y las *redes neuronales*, este aspecto no se ha explorado totalmente por cuestiones de espacio. Posibles fuentes para revisar este enfoque, que puede resultar útil en las aplicaciones, son [1, 143, 71]. Las técnicas de *búsqueda por gradiente*, *optimización* e incluso la teoría de *estimación*, juegan también un papel importante en la implementación práctica de varios de los métodos descriptos. Para una revisión de estos temas es posible consultar los capítulos especiales en [85].

Capítulo 3

Bases fisiológicas de la comunicación

"Pero ellos no entendían nada de esto, eran cosas ininteligibles para ellos, no entendían lo que les decía."

(Lucas 18,34)

Contenido

3.1. Introducción

LA comunicación verbal, tanto escrita como oral, diferencia claramente al hombre del resto de las criaturas. El habla constituye además nuestra forma de comunicación más importante. Las sorprendentes características de este sistema natural no han podido aún ser emuladas por medios artificiales. A los efectos de encontrar una representación de la señal óptima para el diseño de nuevos dispositivos tecnológicos se debe comprender la naturaleza del habla y su forma de producción. Así mismo, es necesario interpretar los aspectos fundamentales del procesamiento llevado a cabo por el sistema auditivo que permiten extraer las características significativas de la señal de voz. Es posible entonces discernir cuales son los parámetros relevantes que deberían preservarse en esta representación y bajo que principios correspondería codificar los mismos. Por todo esto

se requiere el estudio de los fundamentos anatómicos y fisiológicos involucrados en el proceso de comunicación oral humana.

En este punto surge claramente la cuestión acerca de cuanto debe acercarse un mecanismo diseñado por el hombre a este proceso natural, para intentar resolver el problema planteado en este trabajo. Se puede decir que el criterio aquí será acercarse lo necesario como para capturar en los dispositivos artificiales aquellos aspectos esenciales que permitan asegurar algunas capacidades de utilidad práctica. Entre estas capacidades deseables es posible mencionar el lograr la independencia de su desempeño bajo diferentes condiciones como ser cambios en el volumen y la velocidad de pronunciación, en la identidad del hablante (por cambios de rasgos particulares o cuestiones regionales), o en las interferencias del ambiente acústico circundante.

Se denomina comunicación al proceso de transmisión y recepción de información. En el hombre el habla es utilizada para transmitir información de un hablante a un oyente. En la Figura 3.1 se aprecia una diagrama simplificado del proceso de comunicación oral humano. A modo de comparación se muestra también un diagrama en bloques de un sistema genérico de comunicación donde se destacan los elementos constitutivos básicos. La fuente (o emisor) es la encargada de seleccionar los signos que formarán el mensaje. El transmisor es el responsable de codificar adecuadamente el mensaje en una señal. Esta señal es transportada a través del canal, recibida y decodificada en el receptor. Finalmente el significado del mensaje es interpretado en el destino. Nótese que para la adecuada transmisión de un mensaje ambas partes deben compartir un código y un conjunto de signos comunes. Toda interferencia en este proceso es considerada ruido. El código puede incluir información redundante que permita disminuir los efectos del ruido. Esta redundancia puede tomar la forma de signos o reglas repetidas o "superfluas" a distintos niveles del proceso. Esto facilita la interpretación adecuada a pesar de perdidas de información durante el proceso. Con estos elementos en mente se puede resumir el proceso de comunicación humano planteado en la Figura 3.1 de la siguiente forma. Este comienza con una idea o pensamiento que el hablante desea transmitir al oyente [37]. El hablante traduce este pensamiento a través de una serie de procesos neurológicos y movimientos musculares para producir una onda de presión sonora. Esta señal es recibida por el sistema auditivo del oyente, procesada, y convertida nuevamente en una señal neurológica. A partir de ello el oyente forma una idea del mensaje recibido.

Esta explicación resumida esconde algunos aspectos importantes. Para realizar su tarea el hablante convierte la idea a transmitir en una estructura lingüística. Esto se realiza mediante la selección de las palabras y el orden de las mismas que mejor representen la idea, basada en reglas asociadas con el lenguaje en particular. Se agregan también algunas características adicionales como por ejemplo la entonación. En estas primeras etapas se incluye una redundancia importante en el sentido explicado anteriormente. A continuación, el cerebro produce una serie de comandos motores que mueven diversos músculos del sistema vocal para producir la onda de presión sonora deseada. Esta onda acústica es recibida por el sistema auditivo del hablante y convertida nuevamente en una secuencia de pulsos neurológicos. Esto produce la realimentación necesaria para controlar su propia producción de voz. El proceso de percepción en el oyente comienza cuando recibe la onda de presión sonora en el oído externo y la convierte en impulsos

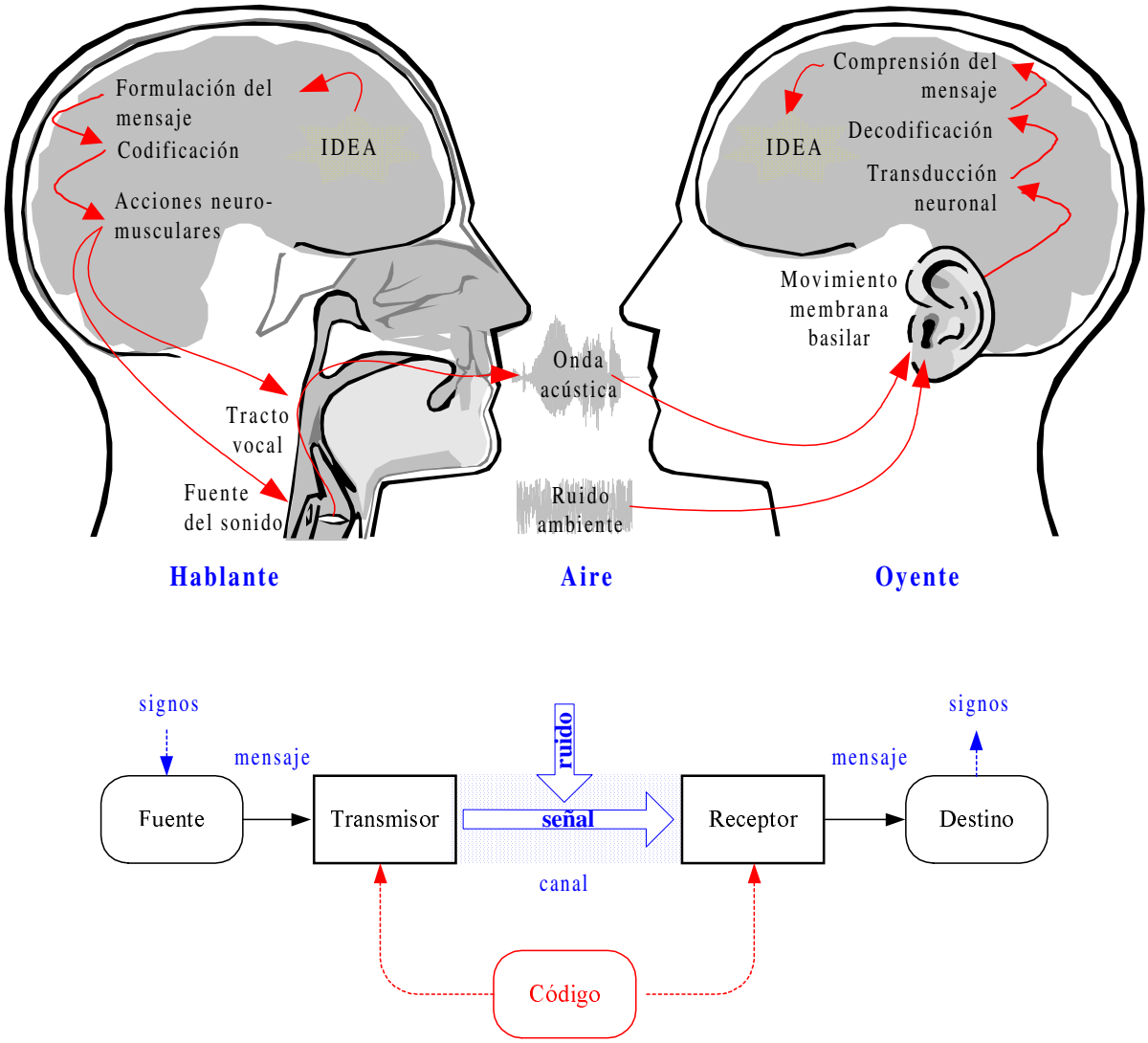

Figura 3.1: *Diagrama simplificado del proceso de comunicación oral de un mensaje en el hombre (arriba). Se resaltan solo las etapas y órganos intervinientes más importantes del proceso, en un único sentido. Diagrama en bloques de un sistema de comunicación genérico (abajo). La aplicación de la terminología y los conceptos derivados de la teoría matemática de la comunicación al proceso de comunicación oral humana permite comprender mejor este complejo proceso.*

neurológicos al pasar por el oído medio e interno. Finalmente interpreta estos pulsos en la corteza auditiva del cerebro para determinar cuál fue el mensaje (lo que implica también la comprensión del significado del mensaje).

Todo este complejo proceso tiene sus bases en los órganos del aparato fonador, el sistema auditivo, y el procesamiento realizado a nivel cerebral en ambos sentidos, requiriendo también para su comprensión una perspectiva lingüística. El aparato fonador y el sistema auditivo no pueden tratarse tampoco de manera aislada. Según Greenberg [63] el aparato vocal humano está probablemente optimizado para producir la comunicación de señales, con propiedades que aprovechan la habilidad del sistema auditivo de codificar la información de una manera robusta, o tolerante a fallas. El espectro del habla está sesgado hacia las bajas frecuencias, que son particularmente resistentes a alteraciones debidas al ruido de fondo. El nivel de presión sonoro de la mayor parte del habla es suficientemente alto como para asegurar que esa información espectral de baja frecuencia se extienda por una amplia serie de canales de frecuencia auditiva. La periodicidad glótica asegura que el sistema pueda seguir o rescatar el habla en condiciones de ruido, acústicamente adversas, y la modulación de la longitud de las sílabas ayuda al cerebro a juntar entidades espectrales dispares en unidades más significativas. Dentro de este marco, la importancia del sistema auditivo para el discurso, está en que precondiciona la representación nerviosa para maximizar la fiabilidad y la tasa de transmisión de información. El cerebro por consiguiente necesita sólo seguir el rastro de estas características en la señal, "confiando" en que son sólo estos rasgos los que codifican la información importante.

Durante el desarrollo de este capítulo se explicarán con mayor detalle todos estos mecanismos para poder dilucidar aquellos aspectos que se deberían preservar en la representación de la señal de voz. El enfoque pretende ser integrador, incluyendo esquemas y diagramas que faciliten la comprensión de las funciones y su relación con las estructuras anatómicas involucradas. Para un estudio más detallado el lector se deberá remitir a la extensa bibliografía específica disponible para cada área (por ejemplo [24, 158, 129, 99]).

Este capítulo se organizará siguiendo un orden similar al de la exposición anterior acerca del proceso de comunicación oral humana. Los aspectos funcionales del proceso son relativamente independientes del idioma considerado, aunque este análisis se limitara al idioma español (principalmente en su versión *argentina rioplatense* [129]). En primer lugar se describirán el mecanismo de producción del habla y los órganos involucrados. Esto incluye la descripción de los principales tipos de sonidos o fonemas que es posible generar mediante el aparato fonador. Luego se presentarán aspectos relacionados con la señal de voz propiamente dicha mostrando algunos ejemplos típicos. Posteriormente se esbozarán los principios y elementos que intervienen en la percepción de los sonidos del habla y la audición. Se enfatizarán aquí los fundamentos de la codificación de la señal de voz a nivel neurosensorial por considerarse de importancia para los objetivos planteados.

3.2. Mecanismo de producción del habla

Para comenzar se esbozarán brevemente los mecanismos involucrados en la producción del habla. Como se mencionó en la sección anterior el proceso de comunicación comienza en el hablante con la traducción de una idea a patrones de variación de la presión sonora en la señal de voz. Para ello el primer paso se realiza principalmente en la corteza cerebral involucrando varias áreas de manera simultánea o alternada. Este proceso es bastante complejo ya que el cerebro debe enviar las ordenes adecuadas al aparato fonador para codificar la información acústica a transmitir por medio de una serie de reglas lingüísticas a diferentes niveles[1]. Cada uno de estos niveles impone ciertas restricciones y "estructura" que forman parte del "código" compartido entre el hablante y el oyente [43, 158, 120] :

Fonológico: se encarga de la representación o modelado de las características físicas de los sonidos utilizados para la producción del habla (fonemas). *No todos los sonidos posibles de generar constituyen fonemas.*

Fonético: se ocupa de la descripción de las variaciones en la pronunciación de los fonemas que aparecen dentro de una palabra o cuando las palabras son dichas juntas en una frase (coarticulación, fusión de sílabas, etc.). *La realización particular de un fonema depende principalmente de su contexto.*

Morfológico: realiza una descripción del modo en que los morfemas (unidades de significación) son combinados para formar palabras. (formación de plurales, conjugación de verbos, etc.). *No todas las combinaciones de morfemas son admitidas.*

Léxico: se ocupa de definir las palabras válidas y el sentido que estas poseen. *No todos las combinaciones de fonemas constituyen palabras permitidas.*

Sintáctico: consiste en las reglas de formación de frases, dando lugar a una limitación del numero de frases. *No todas las combinaciones de palabras son frases autorizadas.*

Prosódico: consiste en una descripción de la fluctuación en la acentuación y entonación durante el transcurso de una frase. *No se admite cualquier patrón de fluctuación.*

Semántico: se ocupa del significado de las palabras y las frases que puede ser visto también como una restricción sobre el alcance del mensaje. *No todas las frases gramaticalmente válidas tienen significado.*

Pragmático: se ocupa de las reglas de conversación. *La respuesta de un interlocutor no debe ser solamente una frase con significado sino también una respuesta razonable acerca de lo que se esta diciendo.*

En la mayoría de las personas las funciones más importantes asociadas con el lenguaje se localizan en el hemisferio izquierdo. A pesar de este predominio del lado izquierdo,

[1]Existe información que se codifica simultáneamente en varios niveles para proveer la necesaria redundancia para aumentar la robustez de la comunicación.

el contenido emocional del lenguaje está gobernado principalmente por el hemisferio derecho. En la Figura 3.2 se puede apreciar un diagrama de las diferentes partes funcionales de la corteza relacionadas con la producción y la comprensión del habla. Dos áreas conocidas como el área de Wernicke y el área de Broca son las más importantes y están involucradas en el almacenamiento de información relacionada con el habla [155]. Ambas áreas se comunican mediante una vía bidireccional denominada *fascículo arqueado*. El área de Wernicke guarda información necesaria para colocar las palabras de un vocabulario previamente aprendido en forma de una conversación con sentido. El área de Broca almacena información necesaria para la producción del habla. Esta última es precisamente la responsable de la programación de la corteza motora para mover la lengua, los labios y los músculos del aparato fonador para articular las diferentes palabras. A continuación la corteza ejecuta este programa que permite coordinar adecuadamente los distintos órganos y partes del aparato fonador para producir la señal sonora requerida. La percepción de su propia voz, en conjunto con la del ruido ambiente, le permite al hablante un continuo monitoreo y control de su fonación. Los cambios producidos en la misma debido a la presencia de ruido se denominan *efecto Lombard* [121, 108, 94] y tienen por objeto minimizar los efectos del ruido. Todo esto conlleva también la necesaria activación de la corteza auditiva en el proceso de producción del habla.

Se debe aclarar que en una conversación normal, además de la comunicación por medio del habla, se utilizan otros medios de transmisión de información no verbales. Un ejemplo de ello son los gestos. Sin embargo estos medios alternativos no se incluirán en este desarrollo. Para la percepción de la señal de voz en condiciones adversas otra información visual como la del movimiento de los labios puede mejorar la inteligibilidad. Este aspecto es procesado en zonas de integración sensorial de la corteza y tampoco será analizado en este trabajo.

3.2.1. Aparato fonador

La forma en la que los cambios en la configuración del aparato fonador modifican las características de la señal acústica serán examinados a continuación. En la Figura 3.3 se observa un esquema simplificado del aparato fonador en conjunto con una sección sagital del mismo (que no incluye a los pulmones). La zona comprendida entre la laringe (glotis) y los labios constituye el *tracto vocal* propiamente dicho. Este está formado por las cavidades supraglóticas, faríngeas, oral y nasal. El aparato fonador se puede considerar como un sistema que transforma energía muscular en energía acústica. La teoría acústica de producción del habla describe este proceso como la respuesta de un sistema de filtros a una o más fuentes de sonidos. En la representación simbólica, y suponiendo linealidad, si $H(f)$ es la función de transferencia del filtro que representa el tracto vocal en un instante dado y $X(f)$ la fuente de excitación, el producto $Y(f) = H(f).X(f)$ representa el sonido resultante. La fuente $X(f)$ indica la perturbación acústica de la corriente de aire proveniente de los pulmones. A veces suele agregarse a este modelo la función transferencia $L(f)$ del fenómeno de radiación a la salida de los labios. Es decir que los sonidos del habla son el resultado de la excitación acústica del tracto vocal, el cual varía constantemente sus características. En este proceso los órganos fonatorios

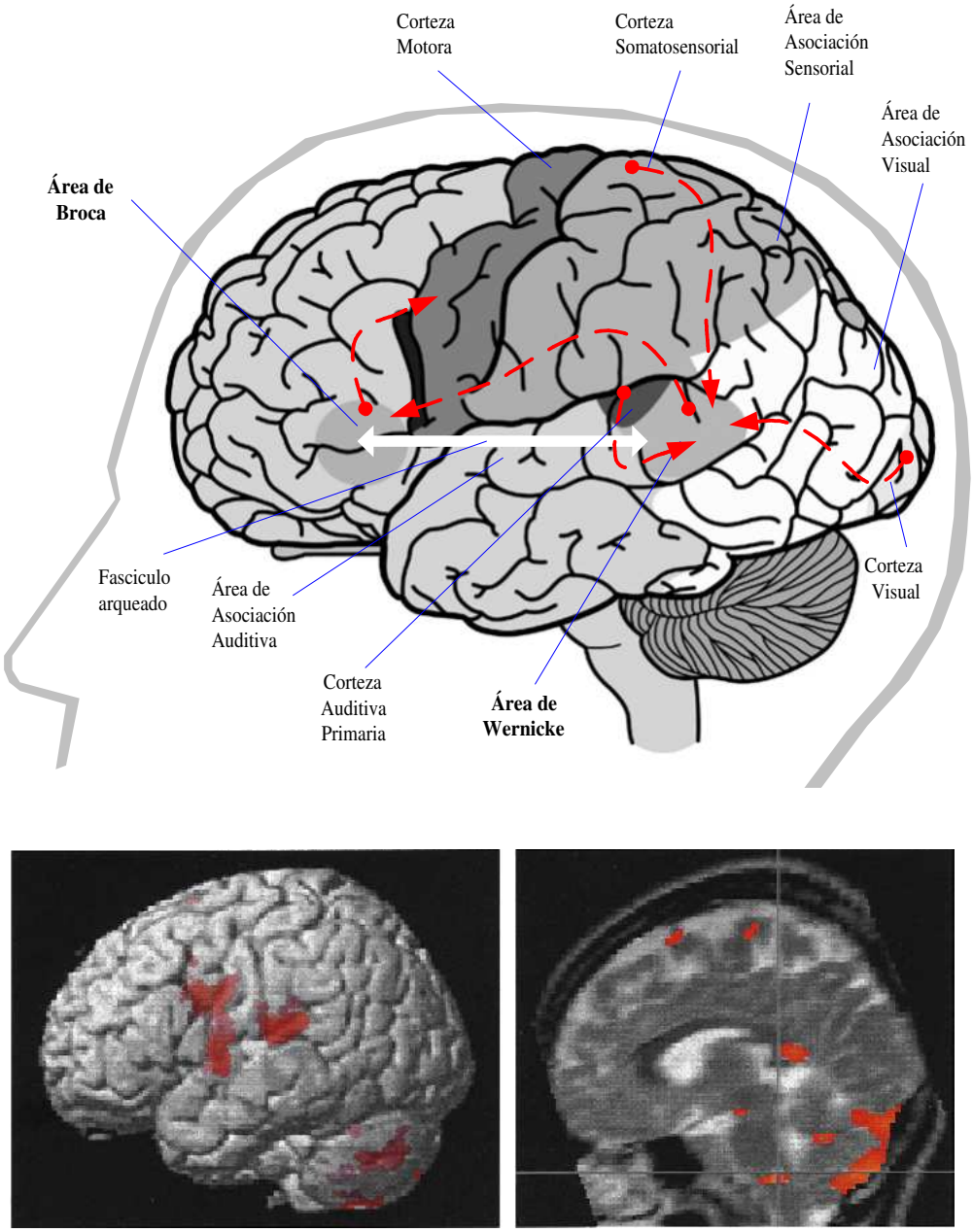

Figura 3.2: *Diagrama de las principales áreas cerebrales implicadas en la producción y comprensión del habla (arriba). Las cortezas sensorial, auditiva, visual y motora primarias muestran la relación del las áreas del lenguaje de Broca y de Wernicke con las áreas menos especializadas que, no obstante están incluidas en el proceso. Áreas activas en la corteza cerebral (abajo izquierda) y actividad de la via motora en el tronco cerebral (abajo derecha) durante la producción del habla (tomado de [79]).*

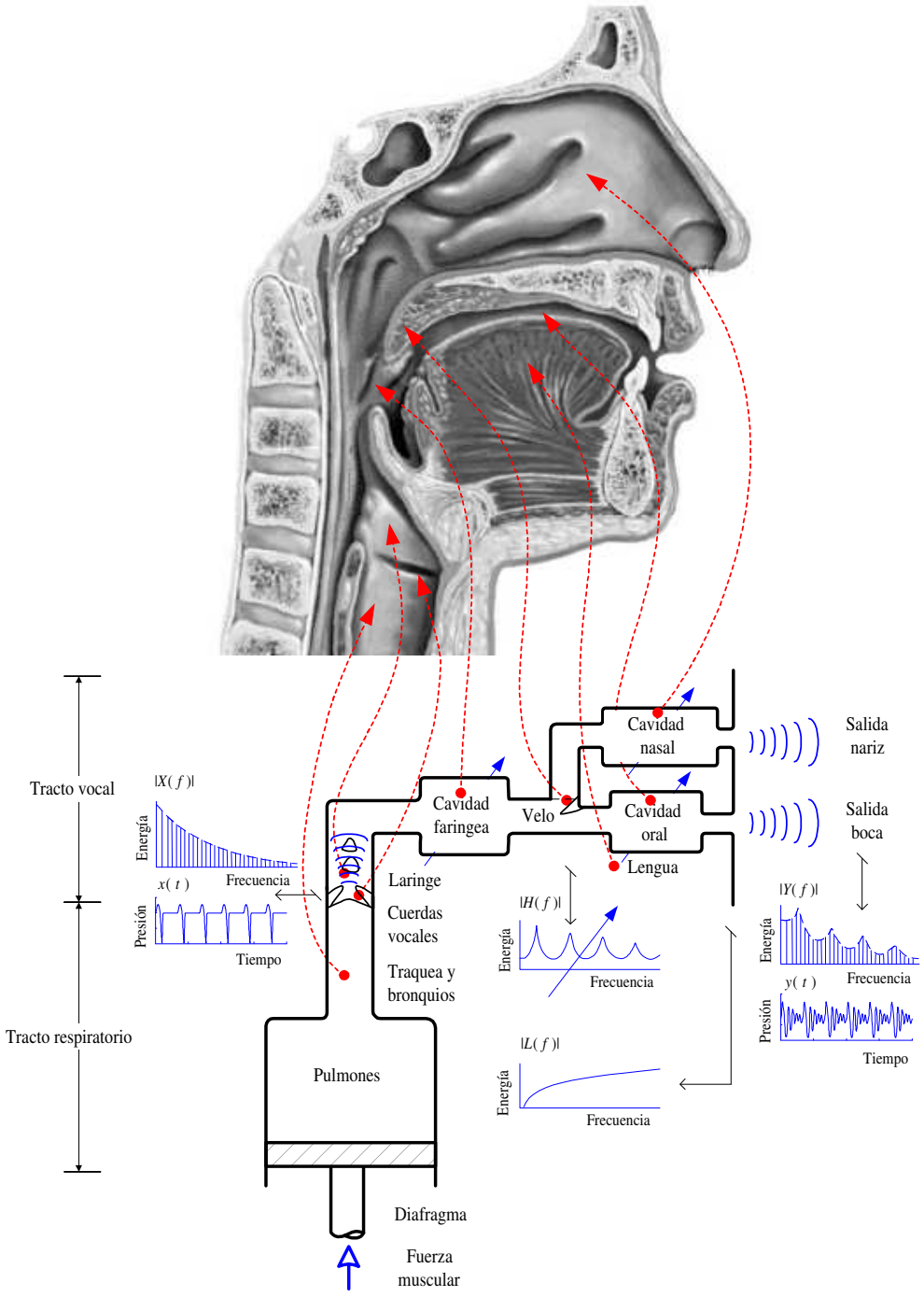

Figura 3.3: *Corte sagital anatómico del aparato fonador (arriba) y diagrama esquemático del mismo que ilustra su funcionamiento (abajo). En el diagrama se ejemplifican las señales temporales, sus correspondientes espectros y sus funciones de transferencia espectrales, para el caso de producción de un fonema sonoro. La suposición subyacente es que se trata de un sistema lineal.*

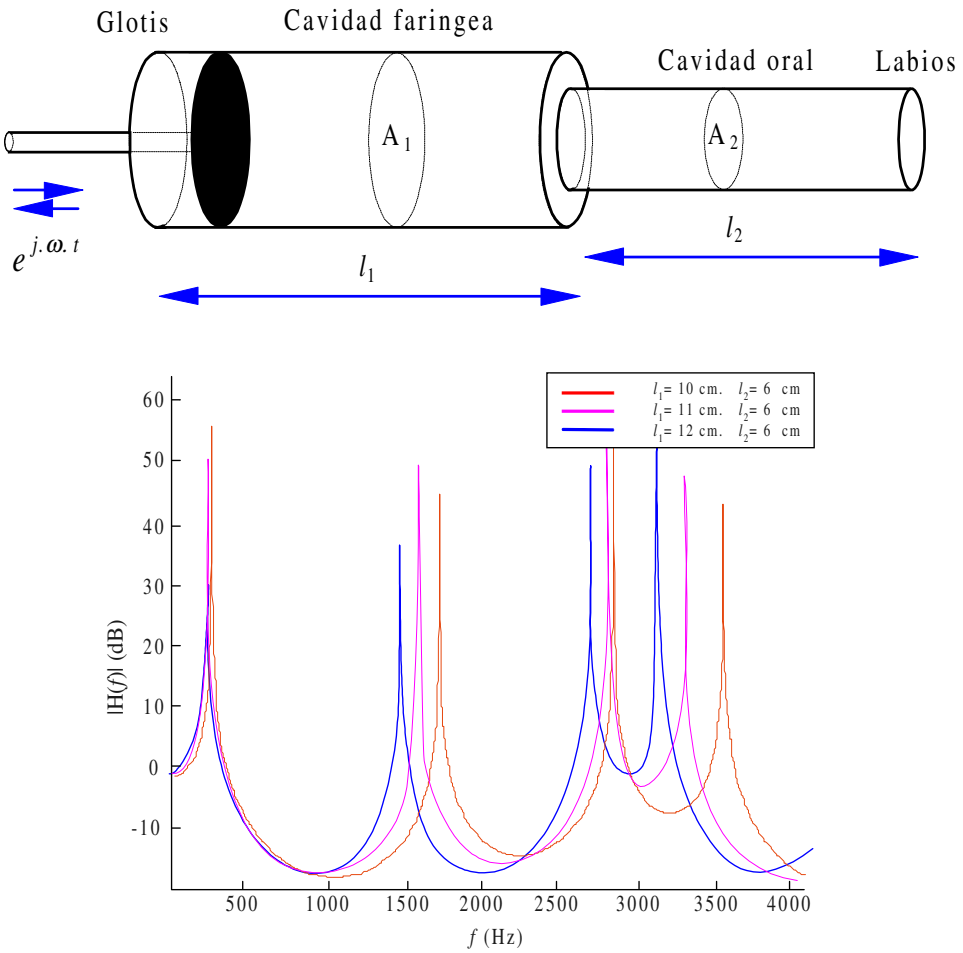

Figura 3.4: *Modelo de dos tubos sin pérdida para el tracto vocal (arriba) y respuesta en frecuencia del mismo para diferentes longitudes de la cavidad faríngea (abajo). Aquí es posible observar como varían las frecuencias de resonancia en función de las diferentes configuraciones, lo que constituye un filtro acústico variante en el tiempo.*

desarrollan distintos tipos de actividades, tales como movimientos de pistón que inician una corriente de aire, movimientos o posiciones de válvula que regulan el flujo de aire, y al hacerlo generan sonidos o en algunos casos simplemente modulan las ondas generadas por otros movimientos.

Para comprender la forma en la que el tracto vocal varía sus características muchas veces se utiliza un modelo sencillo de dos tubos uniformes sin pérdida que varían su ancho o su longitud. En la Figura 3.4 se puede apreciar un modelo de este tipo, junto con distintas respuestas en frecuencias para diferentes configuraciones. Esto permite explicar no solo las diferencias entre los sonidos producidos por un mismo hablante, sino también las existentes entre los sonidos de diferentes hablantes, debido a sus diferencias anatómicas.

El sistema respiratorio constituye la principal fuente de energía para producir sonidos en el aparato fonador humano. La energía es proporcionada en forma de flujo o corriente

de aire y presiones que, a partir de las distintas perturbaciones, generan los diferentes sonidos. De esta forma se pueden identificar tres mecanismos generales en la excitación del tracto vocal:

1. Las cuerdas vocales modulan un flujo de aire que proviene de los pulmones dando como resultado la generación de pulsos cuasiperiódicos.

2. Al pasar el flujo de aire proveniente de los pulmones por una constricción en el tracto vocal se presenta la generación de ruido de banda ancha.

3. El flujo de aire produce una presión en un punto de oclusión total en el tracto vocal; la rápida liberación de esta presión, por la apertura de la constricción, causa una excitación de tipo plosivo, intrínsecamente transitoria.

El aparato respiratorio actúa también en la regulación de parámetros tan importantes como la energía (intensidad), la frecuencia fundamental de la fuente cuasiperiódica, el énfasis y la división del habla en varias unidades (sílabas, palabras, frases).

La laringe juega un papel fundamental en el proceso de producción del habla. En la Figura 3.5 se aprecia un corte longitudinal de la misma junto con un diagrama funcional. La función fonatoria de la laringe se realiza mediante un mecanismo en el que intervienen las cuerdas vocales, los cartílagos en los que se insertan y los músculos laríngeos intrínsecos, y que depende también de las características del flujo de aire proveniente de los pulmones. La forma de onda de los pulsos generados puede representarse en forma simplificada como una onda triangular. En el hombre, la frecuencia de esta onda de vibración de las cuerdas vocales varía entre 100 y 170 Hz, en las mujeres entre 180 y 280 Hz y en los niños puede superar los 300 Hz. Los valores de esta vibración glótica (o frecuencia glótica) se modifican en forma voluntaria y son los responsables de la frecuencia fundamental (denominada F_0) producida al hablar (ver Figura 3.8 más adelante).

El tracto vocal puede mantener una configuración relativamente abierta y actuar sólo como modulador del tono glótico o estrechar o cerrar el paso de la corriente de aire en una zona específica. El tracto actúa como filtro acústico, principalmente en los sonidos con componente glótica, pudiendo modificar sus parámetros en forma continua. Si se observan los espectros de los sonidos vocálicos, éstos proporcionan información sobre todos los aspectos relevantes de la configuración del tracto en ese instante. Es decir, todas las resonancias del tracto, resultantes de su configuración, pueden observarse directamente en el espectro del sonido vocálico. En la Figura 3.6 pueden observarse los sonogramas[2] de las cinco vocales del español junto con sus respectivas envolventes espectrales donde se pueden apreciar claramente estas resonancias a través de los picos espectrales.

3.2.2. Sonidos y fonemas

Como se ha mencionado las unidades lingüísticas básicas del habla son los fonemas. En realidad los fonemas son modelos de los sonidos que pueden diferir luego en su expre-

[2]En este trabajo se denominará "sonograma" a las gráficas de variación de la presión sonora en función del tiempo.

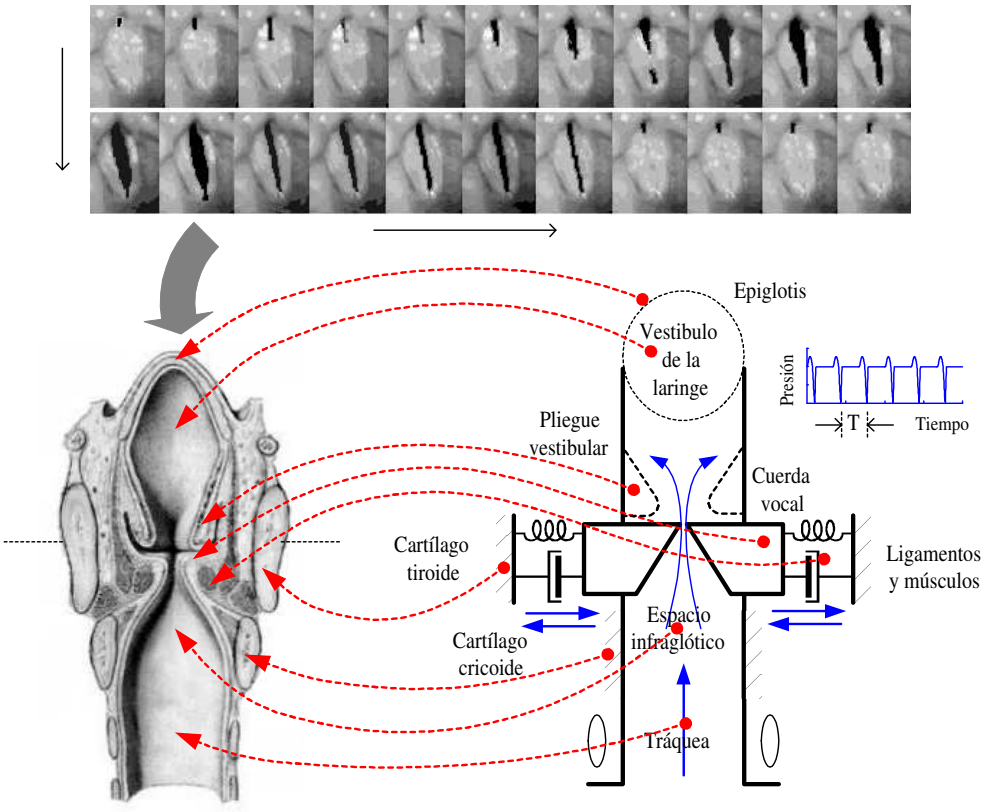

Figura 3.5: *Corte longitudinal de la laringe (abajo izquierda) junto con el diagrama funcional correspondiente (abajo derecha). Se muestra también el aspecto de la glotis en diferentes instantes de la vibración de las cuerdas vocales (arriba). Durante esta secuencia de apertura y cierre de las cuerdas vocales se producen variaciones bruscas en la presión sonora a la salida de las mismas, lo que puede representarse a partir de una señal periódica de período T.*

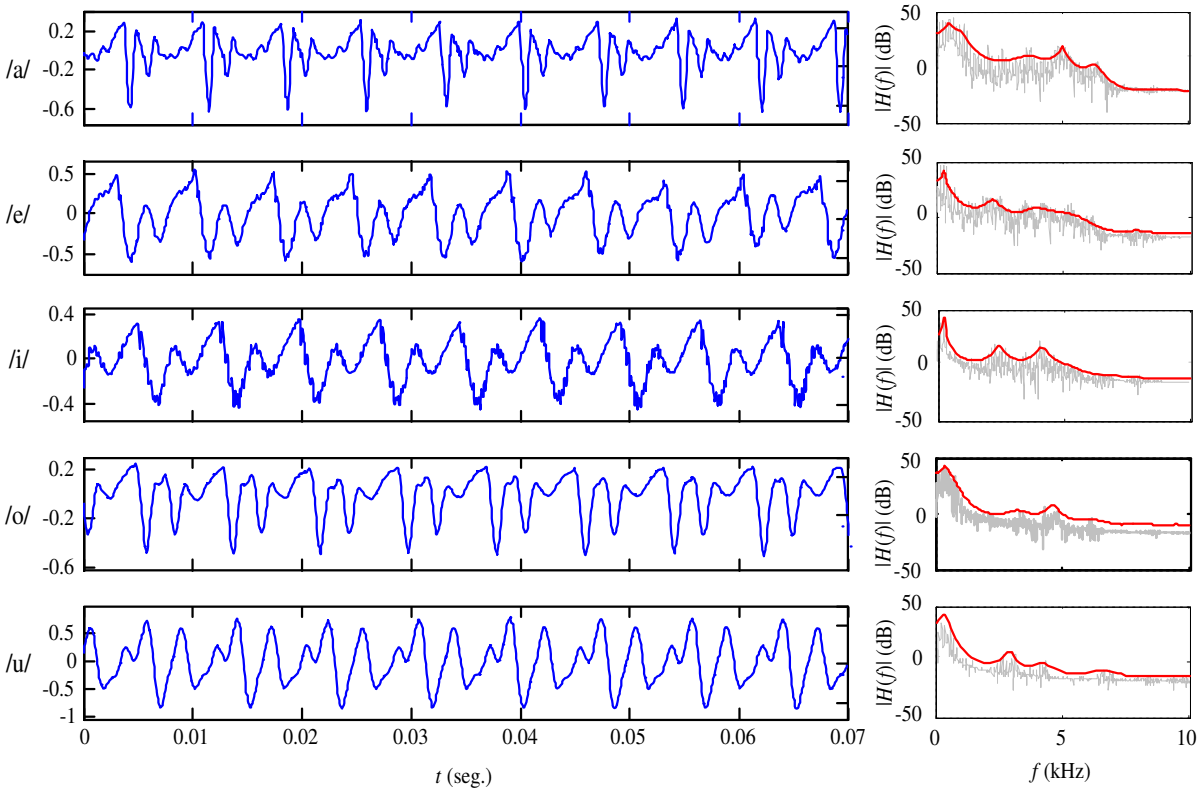

Figura 3.6: *Ejemplos de sonogramas (izquierda) y espectros (derecha) de las vocales del español pronunciadas en forma sostenida y aislada por un hablante masculino nativo. A pesar de la similitud de algunas de sus formas de onda temporales es posible discriminarlas a partir de las resonancias o picos espectrales.*

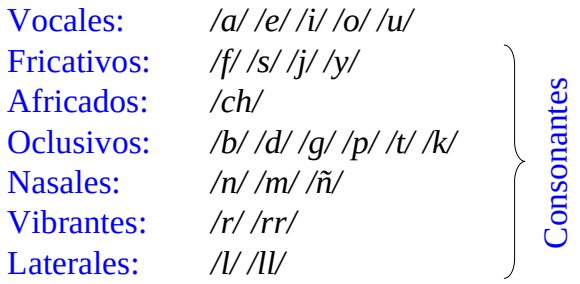

Figura 3.7: *Cuadro simplificado de clasificación de los fonemas del español* rioplatense. *De acuerdo con las características acústicas y los gestos articulatorios que dan lugar a cada tipo de sonido la principal división se da entre las vocales y las consonantes.*

sión acústica[3]. Se los puede definir como el conjunto mínimo de unidades que permite decir cualquier palabra en un idioma determinado. Dos fonemas son distintos si el cambio de uno por otro cambia la palabra (por ejemplo *boda* vs. *moda*). En la Figura 3.7 puede apreciarse un cuadro que muestra los fonemas de uso corriente en nuestro idioma[4].

Se consideraran ahora las configuraciones del tracto que corresponden a cada fonema ya que –como se dijo antes– toda configuración presenta características propias de resonancia que, junto con la fuente de excitación actuante, dan al sonido su peculiar cualidad fonética. Por ello los fonemas se agrupan en vocálicos y consonánticos. Esta división se sustenta tanto en las características acústicas como en los gestos articulatorios que dan lugar a cada tipo de sonido. La duración temporal de los fonemas no es uniforme. Para dar una idea general se puede decir que las vocales son más largas (en el orden de los 100 mseg promedio) que las consonantes (en el orden de los 20 mseg promedio).

Vocales

En la articulación de vocales y sonidos tipo vocálicos, el tracto presenta una configuración relativamente abierta y la fuente de excitación es siempre glótica. Las propiedades de estos sonidos persisten por un tiempo apreciable o cambian muy lentamente mientras se mantenga la configuración del tracto.

Los pulsos glóticos estimulan el tracto vocal que actúa como sistema resonador. Este puede modificar su configuración y con ello sus frecuencias de resonancia como una especie de filtro acústico adaptativo. Esta posibilidad de variación es la que permite al hablante producir muchos sonidos diferentes. La forma del tracto en la producción de las

[3]Se denominan alófonos a las diferentes realizaciones de un mismo fonema. También se utiliza el término *fono* como sinónimo de alófono.

[4]Existen alfabetos fonéticos para aplicaciones tecnológicas con adaptaciones particulares para el español rioplatense [68], tales como:

- SAMPA: `http://www.phon.ucl.ac.uk/home/sampa/spanish.htm`
- Worldbet: `http://www.ling.gu.se/~jimh/courses/ipa.ps`

Sin embargo, por razones de sencillez y salvo que se indique lo contrario, para hacer referencia a los fonemas se utilizará la grafía más cercana (a su pronunciación) encerrada entre /•/.

vocales esta controlada principalmente por la posición de la lengua, de la mandíbula y de los labios. Los sonidos vocálicos se pueden clasificar por sus distintas características acústicas[5] [120]:

Zonas de estrechamiento: Por estudios sistemáticos de radiografías de articulaciones vocálicas se han localizado tres zonas principales de producción de la constricción. Esto depende de la posición de la lengua, los labios, y la boca. De esta manera los sonidos vocálicos se agrupan en *anteriores* (*/i/*, */e/*), *medios* (*/a/*), y *posteriores* (*/o/*, */u/*) según la posición de la constricción.

Abertura de la boca: Esta abertura cuya configuración y grado están determinadas por la acción de los labios y del maxilar inferior, da lugar a importantes diferenciaciones acústicas y fonéticas. Así se tienen en forma relativa a las vocales *abiertas* (*/a/*), *medias* (*/e/*, */o/*) y *cerradas* (*/i/*, */u/*).

Grado de estrechamiento: De esta manera se describen los sonidos vocálicos según el grado de estrechamiento en la región de menor área o constricción máxima, en *estrechos* (*/i/*, */u/*, */o/*) y *amplios* (*/e/*, */a/*).

Longitud del tracto: La longitud del tracto se modifica redondeando los labios, subiendo y bajando la posición de la laringe. Así se tienen las vocales *labializadas* (*/o/*, */u/*) y *deliabializadas* (*/a/*).

Consonantes

Los sonidos consonánticos se producen con una configuración relativamente cerrada del tracto vocal. El cierre o estrechamiento del canal se realiza en zonas especificas del tracto vocal por acción de partes especificas de las estructuras articulatorias. Entre los factores que determinan la cualidad del sonido resultante, se deben distinguir aquellos que hacen al modo de articulación (cierre o estrechamiento) de los que señalan la zona o lugar de articulación (lugar donde se produce cierre o estrechamiento). La participación de la fuente glótica, la naturaleza del cierre o estrechamiento y la transmisión a través de la cavidad oral y/o nasal, constituyen los principales factores del modo de articulación.

Las consonantes, por otro lado, pueden ser agrupadas en los siguientes tipos articulatorios:

Fricativas: se caracterizan por ser ruidos aleatorios generados por la turbulencia que produce el flujo de aire al pasar por un estrechamiento del tracto. Pueden ser sonoros como */y/* si hay componente glótica o sordos como */f/*, */s/* o */j/* (también */z/* en otras versiones del español) si no la hay.

Africadas: si los fonemas comienzan como oclusivos y la liberación del aire es fricativa se denominan africados. Por ejemplo la */ch/*.

[5]En el español las dos primeras características son las más importantes para diferenciar entre las vocales.

Oclusivas: se producen por el cierre momentáneo total o parcial del tracto vocal seguido de una liberación más o menos abrupta del aire retenido. Por ejemplo las totales */p/*, */t/*, */k/* o las parciales */b/*, */d/*, */g/*. Estas últimas son sonoras.

Nasales: son producidas a partir de excitación glótica combinada con la constricción del tracto vocal en algún punto del mismo. Por ejemplo */m/*, */n/* o */ñ/*.

Vibrantes: éstas son producidos al pasar el aire por la punta de la lengua y producir su vibración. Tienen componente glótica. Por ejemplo */r/* y */rr/*.

Laterales: estas se producen cuando se hace pasar la señal sonora glótica por los costados de la lengua. Por ejemplo */l/* y */ll/*.

Semivocales: están formadas por la unión de dos de los anteriores hasta el punto de convertirse en otro sonido (por ejemplo dos vocales). Algunos consideran en este grupo a las vibrantes (*/rr/*) y las laterales (*/ll/*).

3.2.3. Segmentos, suprasegmentos y sílabas

De lo dicho anteriormente, se podría inferir que el habla es, de alguna manera, un fenómeno secuencial "discreto", es decir una sucesión de fonemas. De hecho, como se verá más adelante, es posible asignar *etiquetas* a los diferentes trozos de señal asociados con estos fonemas. Sin embargo si se observa la señal de la voz, la representación acústica de una frase, se verán muy pocas pausas o intervalos entre los sonidos. De esta forma el habla constituye un continuo acústico, producido por un movimiento ininterrumpido de los órganos del aparato fonador. A pesar de la naturaleza continua de la voz los oyentes pueden segmentarla en sonidos.

Aquellas características de la voz de una escala temporal superior al fonema se denominan suprasegmentales. Estas características están determinadas principalmente por la entonación, la cual determina la prosodia. Las variables que intervienen en la entonación son las variaciones de frecuencia fundamental o F_0, la duración y variaciones de energía y sonoridad.

La prosodia en las uniones puede ser caracterizada por silencios, duración en las vocales, o por formas como puede ser la presencia de sonoridad o aspiración. Por ejemplo en la frase "perdonar, no matar" existe una pausa después de "perdonar" pero si la coma cambia de lugar "perdonar no, matar" el silencio se produce después de "no" cambiando totalmente el significado del mensaje.

La sílaba constituye una unidad lingüística de escala temporal mayor que la del fonema. Si bien para una lengua la cantidad de sílabas es muy superior a la de fonemas, en general la variabilidad acústica de estas unidades es también mucho menor. Por ello algunos investigadores prefieren su utilización como unidad de modelado del habla.

3.3. Señal de voz

Hasta ahora se han descripto los distintos tipos de fonemas y la forma en la que se originan en el aparato fonador. Sin embargo se han hecho pocas referencias a los aspectos relacionados con la señal de voz propiamente dicha, que constituye el substrato del que se obtendrá una representación adecuada. Los aspectos discutidos en la presente sección están más relacionados con la fonética acústica que con la fonología.

Se comenzará por analizar las vocales, por constituir el caso más sencillo. En la Figura 3.6 pueden observarse el sonograma de las vocales del español pronunciadas en forma sostenida y aislada junto con sus respectivos espectros. En este caso se aprecia un cierto parecido entre */o/* y */u/* o entre */e/* y */i/*, lo cual es de suponer porque se puede decir que son vocales 'cercanas' según se verá a continuación. Como ya se mencionó en los espectros de los sonidos vocálicos pueden observarse todas las resonancias del tracto. Estas resonancias aparecen como picos en el espectro y se denominan formantes. Las formantes se numeran a partir del 1. Las formantes, principalmente F_1 y F_2, constituyen un medio para caracterizar a las vocales. De hecho, la presencia de formantes, y en particular de F_0 evidencia si se trata de un trozo sonoro o sordo (con o sin componente glótica). A pesar de la notación F_0 no constituye estrictamente una formante sino, como ya se indicó, la frecuencia fundamental que está directamente relacionada con la entonación de una frase o emisión[6]. En la Figura 3.8 aparece el espectro de una */i/* y su correspondiente envolvente espectral (estimada mediante un modelo autoregresivo) donde se aprecian claramente los picos y se muestran las distintas formantes. En la Figura 3.9 se puede apreciar un gráfico de la distribución de las vocales del español –o mapa de formantes– para hablantes masculinos en función de F_1 y F_2. Se puede observar que mediante estas características es posible separar o modelar fácilmente a las diferentes vocales. En el gráfico se muestra también la relación del valor de las formantes con los atributos articulatorios discutidos en la Sección 3.2.2 y el denominado *triángulo de las vocales*. Las formantes de esta figura han sido obtenidas de vocales aisladas pronunciadas en forma sostenida. En el caso del discurso continuo las formantes siguen siendo un rasgo distintivo importante para las vocales. Sin embargo en este caso es preciso seguir también la evolución de los patrones formánticos debido a que las clases no se encuentran tan bien separadas [77]. Este fenómeno está relacionado con el hecho, explicado anteriormente, que la voz constituye en realidad un fenómeno continuo. A lo largo de una frase las variaciones en la morfología del tracto vocal y las características de la excitación dan como resultado un cambio permanente del espectro de la señal resultante. En el caso más general estos patrones espectrales permiten caracterizar a los distintos fonemas mediante determinadas *pistas acústicas* que son requeridas para poder diferenciarlos. En la Figura 3.10 se pueden apreciar estos patrones para la frase del español "¿Cómo se llama el mar que baña Valencia?", segmentada y etiquetada. Se pueden destacar algunas pistas acústicas presentes en el espectrograma de esta figura. Se observa la corta duración y la *explosión* de la oclusiva */k/*. La estructura formántica de las vocales está evidenciada por las regiones más oscuras de conjuntos equiespaciados de lineas paralelas en dirección

[6]En el modelo lienal de producción de la voz fuente-filtro discutido en la Sección 3.2.1 F_0 es una caracterítica de la fuente mientras que F_1 y F_2 corresponden a características del filtro.

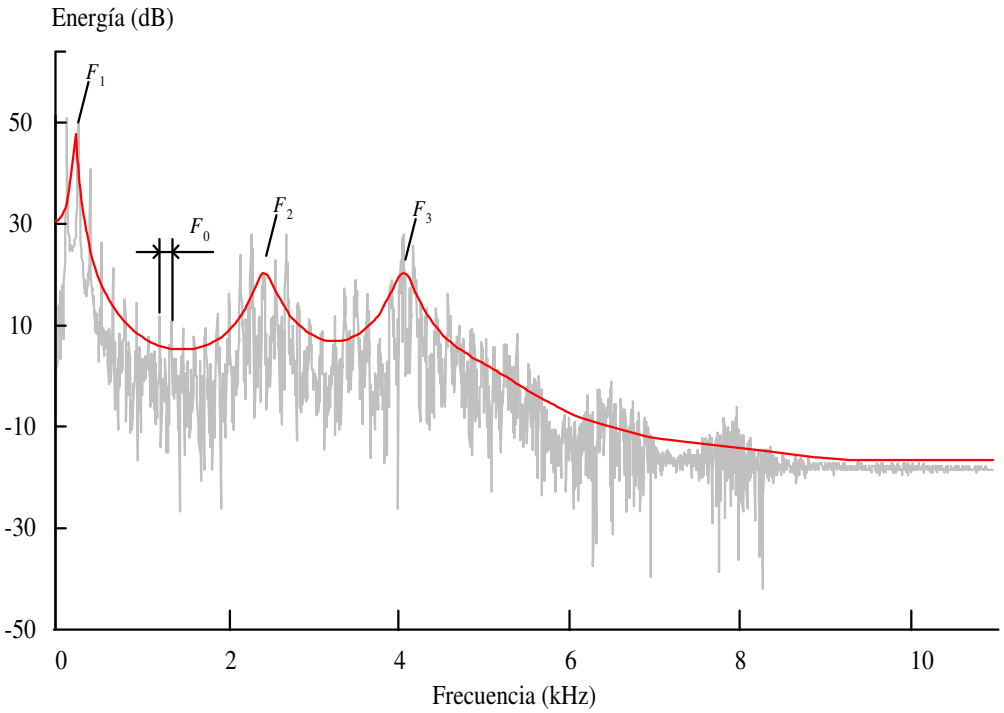

Figura 3.8: *Espectro de una vocal /i/ pronunciada en forma sostenida y su envolvente, donde se resaltan las frecuencias formantes (F_1, F_2 F_3 y la frecuencia fundamental F_0). F_0 corresponde a la frecuencia glótica y es uno de las componentes de la entonación del habla, mientras que el resto constituyen las formantes que permiten discriminar entre las vocales. Su variación temporal permite también diferenciar entre los diferentes fonemas sonoros.*

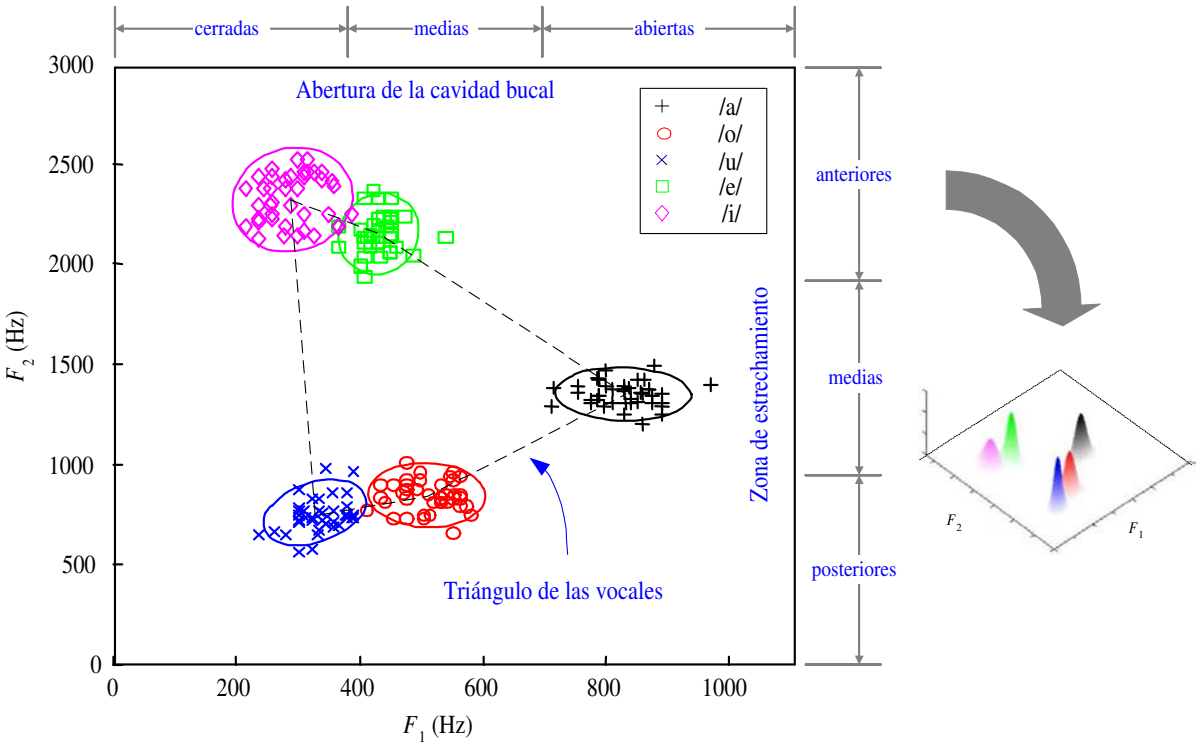

Figura 3.9: *Mapa de las formantes obtenido a partir de datos experimentales para las vocales del español pronunciadas en forma sostenida por un conjunto de hablantes masculinos. Para dibujar las elipses se ha supuesto una distribución gaussiana bidimensional para cada clase [5]. Sobre el mapa se ha superpuesto el clásico triángulo de las vocales del español, mostrando además sobre ambos ejes la relación de F_1 con la abertura de la boca y de F_2 con las zonas de estrechamiento del tracto vocal.*

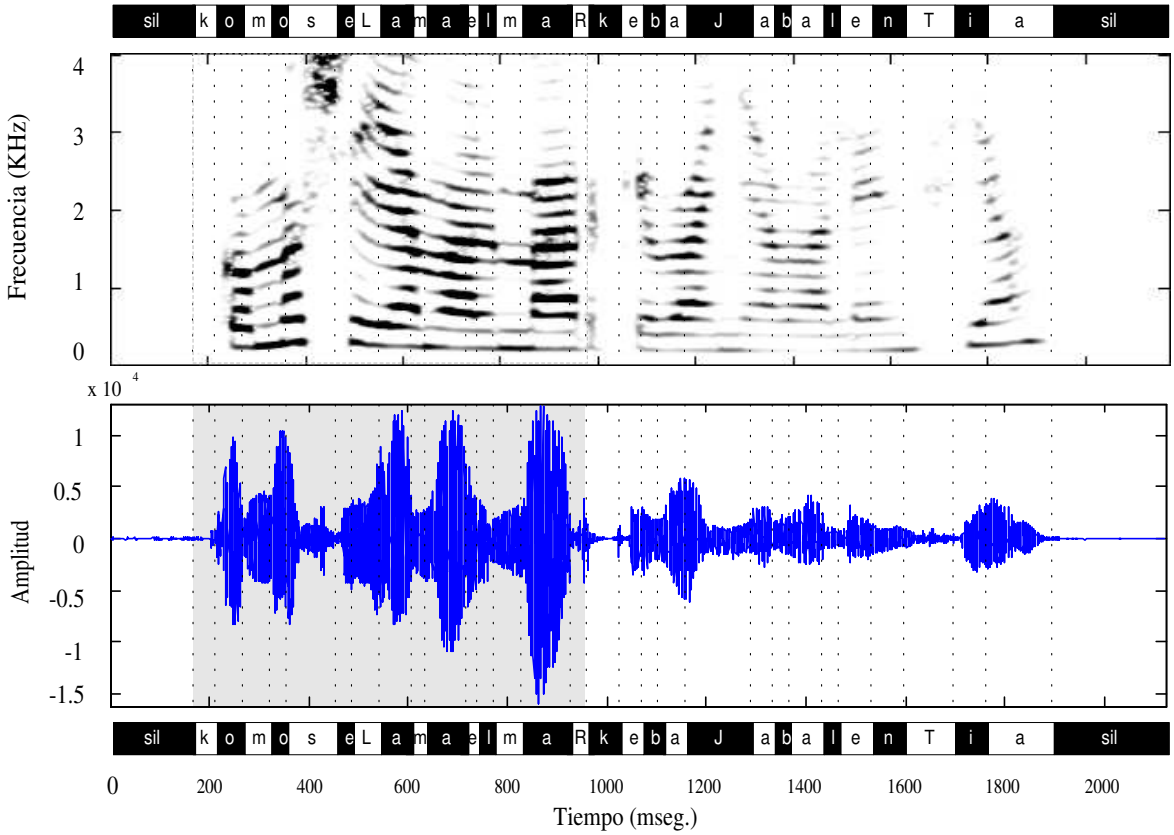

Figura 3.10: *Sonograma y espectrograma de la oración "¿Cómo se llama el mar que baña Valencia?", segmentada y etiquetada (etiquetas de acuerdo al alfabeto fonético* Worldbet*). El espectrograma es de banda angosta (Ver Sección 4.4.1). El trozo resaltado se amplía en la figura siguiente (3.11). La frase ha sido tomada de la base de datos de habla española Albayzin [17].*

horizontal, producto de su carácter sonoro cuasiperiódico. Se puede observar también el contenido de alta frecuencia de la */s/* y la ausencia de sonoridad.

Existen algunas características de la señal de voz que se pueden manifestar mediante análisis relativamente simples como ser la *energía de corta duración* y la *cantidad de cruces por cero* (Cx0). Estos análisis tienen la ventaja de ser sencillos en su implementación digital y muy rápidos. La energía da una idea de la intensidad de la señal en función del tiempo y constituye un parámetro de suma importancia ya que permite diferenciar entre varios tipos de fonemas. Es también una parte esencial de la entonación (junto con F_0). Los cruces por cero constituyen una medida indirecta del contenido frecuencial de la señal. En la Figura 3.11 se observa una sección ampliada de la frase mostrada en la figura anterior correspondiente al trozo "¿Cómo se llama el mar...". En ella se muestran junto con el espectrograma y las formantes, las curvas derivadas de estos análisis temporales. En general otro rasgo distintivo de los fonemas sonoros consiste en que poseen una menor cantidad relativa de Cx0 que de energía (ver por ejemplo */o/* y */a/*). La situación inversa puede apreciarse en los fonemas sordos (no sonoros), como la */s/*, debido a que poseen poca energía y distribuida en las frecuencias altas. De esta manera es posible distinguir rápidamente entre ambas clases. En el caso de los fonemas sordos puede apreciarse también la pérdida de la sonoridad por la anulación de F_0 (otra vez como en */s/*).

Pueden destacarse también otras pistas acústicas que permiten discriminar entre los diferentes fonemas, generalmente visibles en su representación espectral. En la Figura 3.12 pueden observarse algunos ejemplos de estas pistas que permiten discriminar entre */s/*, */f/*, */m/*, */n/*, */l/* y */r/* [15]. La */s/* suele ser fácil de reconocer y distinguir de la */f/*. Ninguna de la dos posee componente glótica. En el caso de la */s/* aparece un área de fricación de mayor energía en la zona de las altas frecuencias (entre los 3000 y los 8000 Hz) En el caso de la */f/* el área de mayor energía suele ser un triángulo alrededor de los 1200 Hz. También puede existir alguna coarticulación con los fonemas adyacentes. De forma similar pueden establecerse algunas pistas para discriminar entre los fonemas sonoros */m/* y */n/*. En el caso de */m/* las formantes generalmente se "sumergen" dentro del fonema y luego se elevan cuando este termina, excepto cuando las frecuencias de las mismas ya son bajas. En */n/* el cambio suele ser más abrupto. El nivel de frecuencia al que tiende F_2 para */m/* está entre 900 a 1400 Hz, mientras que para */n/* está entre 1650 a 1800 Hz. Para el fonema */l/* es posible notar un "hueco" (cero o anti-resonancia) en el espectro, aproximadamente entre 1500 y 2000 Hz. A ambos lados de este hueco F_2 y F_3 al principio divergen y posteriormente se juntan. En algunos casos */l/* sólo se puede distinguir como una disminución en la energía de F_2 y F_3. En el caso del fonema */r/* se puede apreciar que F_3 y F_2 se acercan, o inclusive se combinan, siempre se fuerza F_3 por debajo de 2000 Hz.

Podrían llenarse muchas páginas con gráficos y análisis de los distintos fonemas. Sin embargo el interés aquí no es presentar este material de manera exhaustiva sino más bien, y como ya se mencionó, mostrar unos pocos ejemplos que permitan comprender mejor la naturaleza de la señal de voz y sus rasgos más significativos.

Como consideraciones finales de esta sección se debe remarcar el hecho ya discutido acerca de que la realización acústica de un fonema depende mucho de su contexto inme-

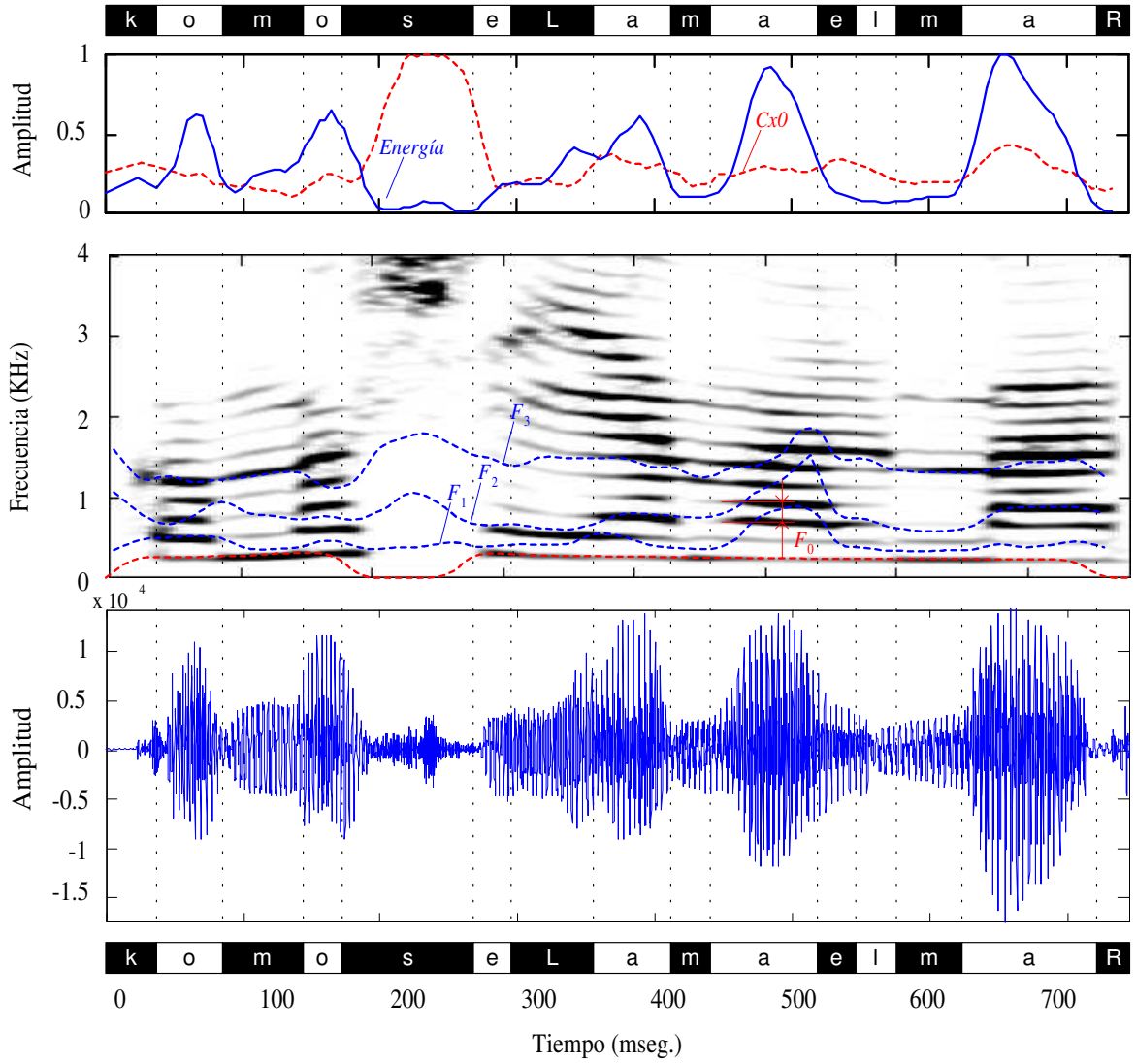

Figura 3.11: *Sonograma, espectrograma, formantes, energía y cruces por cero simultáneos de un trozo de la oración de la Figura 3.10 "¿Cómo se llama el mar...?". La combinación simultánea de estos análisis permite la rápida caracterización de los diferentes fonemas.*

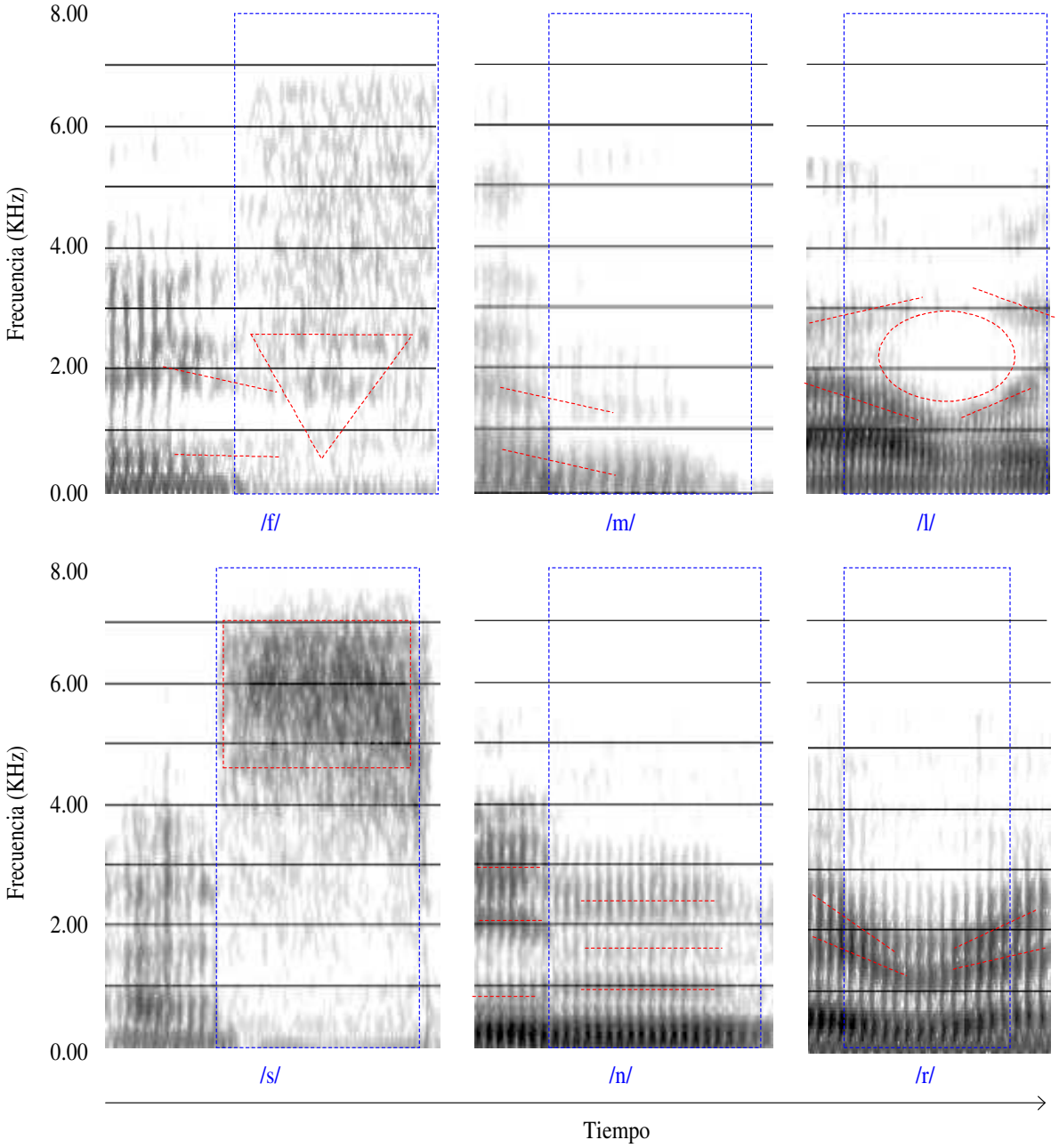

Figura 3.12: *Pistas acústicas correspondientes a ejemplos típicos de varios de los fonemas explicados en el texto resaltadas en los correspondientes espectrogramas de banda ancha (espectrogramas tomados de [15]). Estas pistas o rasgos acústicos permiten discriminar entre los diferentes fonemas (o alófonos de los mismos).*

diato. Por otra parte muchas veces, especialmente en el caso del habla espontánea, los fonemas no están articulados adecuadamente o no se parecen tanto a lo que se esperaba idealmente. El hecho que el habla sea una secuencia continua de fonemas sin pausas acústicas explícitas entre las palabras constituye un problema adicional.

3.4. Fisiología de la audición

En este trabajo, resulta de interés comprender cómo se realiza el procesamiento de la señal de habla en el sistema auditivo. Se debe tener en cuenta que este sistema realiza una enorme cantidad de procesamiento para que la señal llegue hasta nuestro cerebro, pero es realmente allí donde se produce el fenómeno de la audición. Se podría decir entonces que en realidad "escuchamos" con el cerebro. Por ello es importante comprender que rasgos significativos se preservan en las representaciones internas de la corteza cerebral, y cuales son los principios que orientan la formación de estas representaciones. Se podría realizar la siguiente pregunta: ¿Que características del sistema auditivo son particularmente apropiadas para codificar la voz?. La respuesta, en parte, se encuentra en la magnífica capacidad de este sistema para resolver simultáneamente tanto las características espectrales como temporales de los estímulos de banda ancha que constituyen el habla humana. Por otra parte esta capacidad se mantiene aún en condiciones acústicas muy desfavorables, con relativa independencia de cambios en el canal (presencia de ruido o ambiente reverberante) o la fuente del mensaje (velocidad de pronunciación o identidad del hablante).

En la Figura 3.13 puede apreciarse un corte transversal del oído, junto con un diagrama esquemático que ilustra su funcionamiento. En el mismo se observan sus tres secciones principales: el oído externo, el medio y el interno. Se podría decir que las dos primeras partes se encargan de la recepción y adecuación del sonido para su posterior procesamiento en la sección siguiente. Las funciones más importantes, como la transducción del sonido a impulsos nerviosos, se realizan en el oído interno. Se describirán a continuación estas partes del oído y sus funciones con mayor detalle.

3.4.1. Recepción y adecuación acústica

El oído humano funciona en un medio aéreo y por ello necesita cierta eficiencia para la recepción de sonidos transmitidos por el aire. La parte más externa es el *pabellón auditivo* que está encargado de captar el sonido y enfocarlo hacia el *conducto auditivo*. Las ondas de presión siguen el conducto auditivo hasta el *tímpano* que separa oído externo del oído medio. Este último está constituido por una cámara ocupada por aire (que se comunica con la faringe a través de la *trompa de Estaquio*) y un conjunto de huesecillos: el *martillo*, el *yunque* y el *estribo*. El sonido se transmite entonces desde la membrana del tímpano a través de la cadena de huesecillos, cuya función principal es la de adaptación de impedancias acústicas [100]. El *estribo*, el más interno de estos huesecillos, establece contacto con la *ventana oval* que está ubicada en la base de la *cóclea*, en lo que constituye el oído interno. La amplificación de las vibraciones producidas en el tímpano

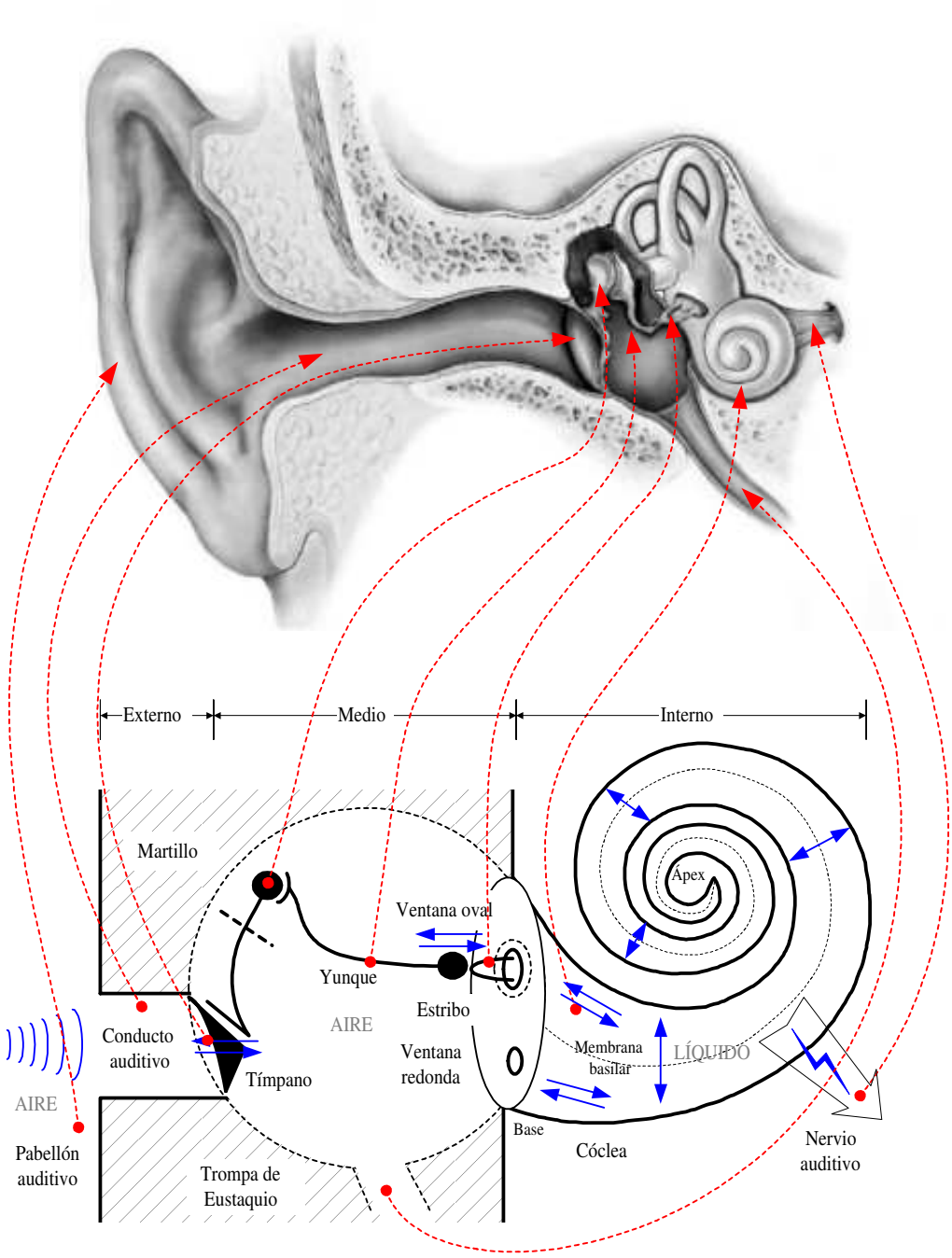

Figura 3.13: *Corte sagital anatómico del oído (arriba) y diagrama esquemático que ilustra su funcionamiento (abajo). El oído es el encargado de la recepción y adecuación del sonido y de su transducción a impulsos nerviosos. En el diagrama se resaltan sus secciones principales: el oído externo, el medio y el interno, que son las que realizan cada una de estas tareas.*

está limitada, en condiciones de cambios abruptos, por el *reflejo estapedial* para proteger al oído interno[7].

3.4.2. Transducción mecánico-eléctrica

El órgano principal del oído interno es la cóclea. La cóclea puede describirse como un tubo cónico lleno de líquido (*perilinfa*) y enrollado en forma de caracol. En la Figura 3.14 puede apreciarse una versión aislada y ampliada de la misma con su correspondiente diagrama esquemático. En este diagrama la cóclea se muestra desenrollada para mayor claridad. Una vez excitada la ventana oval el sonido se transmite a través del líquido de la *rampa vestibular* en la cóclea, atraviesa el *helicotrema* y sigue su recorrido en la *rampa timpánica* hasta la *ventana redonda*. La ventana oval y la redonda trabajan de forma tal que cuando una se comba hacia adentro la otra se comba hacia afuera y viceversa. El movimiento hacia adentro y afuera se repite con la misma frecuencia del estímulo sonoro.

Es en la *membrana basilar* donde tiene lugar la transducción, de manera selectiva, en base a la relación de las características del estímulo y la zona de vibración de la misma [73, 9, 51]. La membrana basilar varía sus propiedades mecánicas de forma continua a lo largo de su eje longitudinal. La membrana es más rígida en su base, cerca de la ventana oval, donde su ancho es mínimo. Por lo tanto tiene allí menor cantidad de masa por unidad de longitud. Esto hace que la región de la base región vibre con preferencia ante un estímulo de alta frecuencia. De esta forma, las vibraciones de frecuencias altas tienen su máxima amplitud cerca del lugar donde las ondas comienzan a desplazarse, luego disipan la mayor parte de su energía y se desvanecen en el camino, no alcanzando nunca el ápex. Las vibraciones de baja frecuencia, por el contrario, comienzan con una amplitud pequeña cerca de la base y la aumentan a medida que se acercan al ápex. De esta manera están representadas todas las frecuencias audibles a lo largo de toda la cóclea. A esta característica se la denomina *tonotopía de la membrana*.

Se han registrado las excursiones máximas de la membrana basilar en función de la distancia al estribo (envolventes de la onda de desplazamiento), para tonos de igual intensidad pero distintas frecuencias. Empleando estos datos se pueden dibujar las curvas de resonancia o sintonía mecánica, esto es las amplitudes relativas de las excursiones para los distintos puntos sobre la membrana basilar como una función de la frecuencia del estímulo (Figura 3.15). De estas curvas de sintonía resulta ser que la relación entre la distancia al estribo y la frecuencia de vibración máxima no es lineal, sino más bien de tipo logarítmica. Esta es una de las causas por las que la resolución frecuencial y la percepción de las frecuencias no es uniforme en toda la cóclea. A la escala psicoacústica que da cuenta de la relación entre la frecuencia física del sonido y la percibida se la denomina *escala de mel* (Ver Figura 4.20 más adelante). Los experimentos psicofísicos demuestran también una escala similar de carácter logarítmico en la percepción de la intensidad de los sonidos, cuya unidad es el *fono*[8].

[7] Esto funciona en la práctica como un *control automático de ganancia mecánica.*

[8] Por ejemplo, si un sonido complejo con muchas componentes, parece igualmente intenso que un tono puro de 1000 Hz con un nivel de presión de 80 dB (SPL), aquel tendrá un nivel de sonoridad de 80 fonos, independientemente del nivel de presión "real" que tenga.

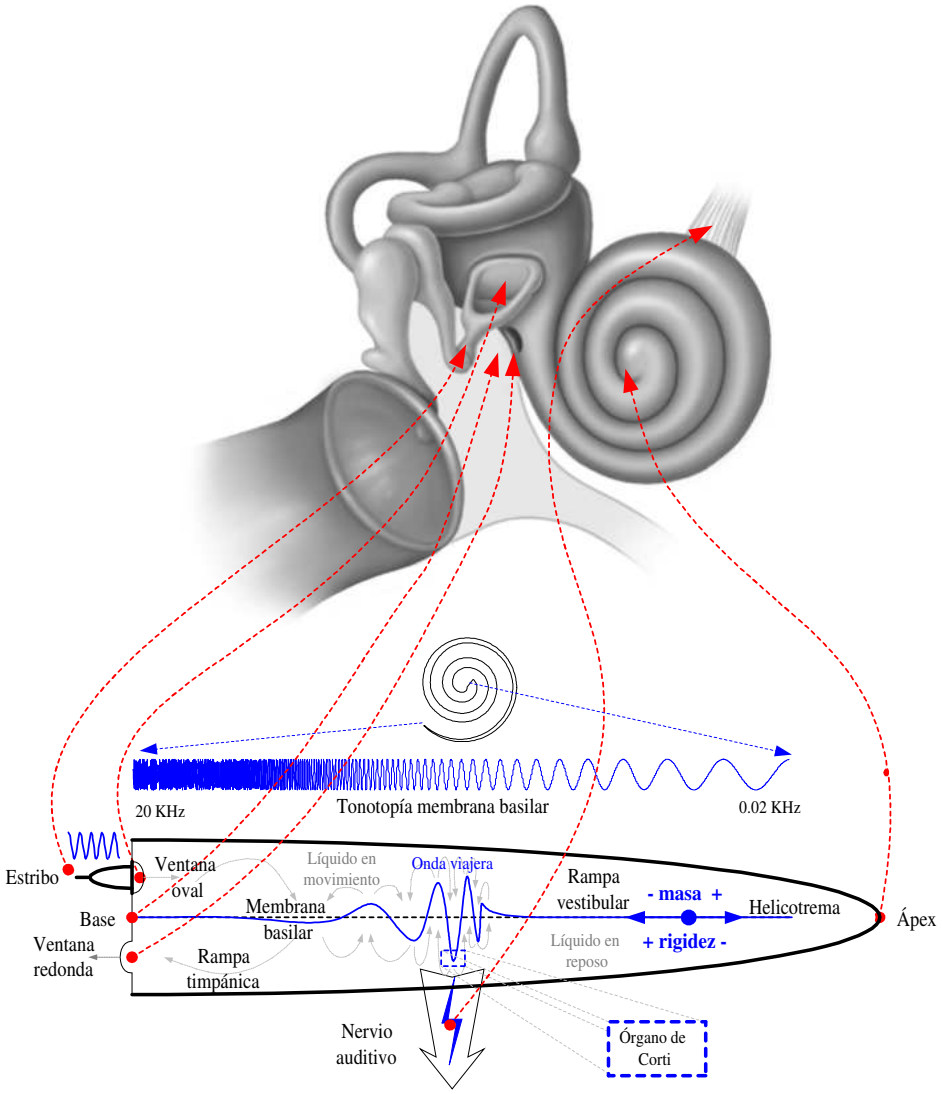

Figura 3.14: *Cóclea aislada (arriba) y diagrama esquemático que ilustra su funcionamiento (abajo). En el diagrama la cóclea se halla desplegada para mayor claridad. Se muestra también la forma de una* onda viajera *típica (cuya amplitud se ha exagerado) y se resaltan los aspectos relativos a la tonotopía de la membrana.*

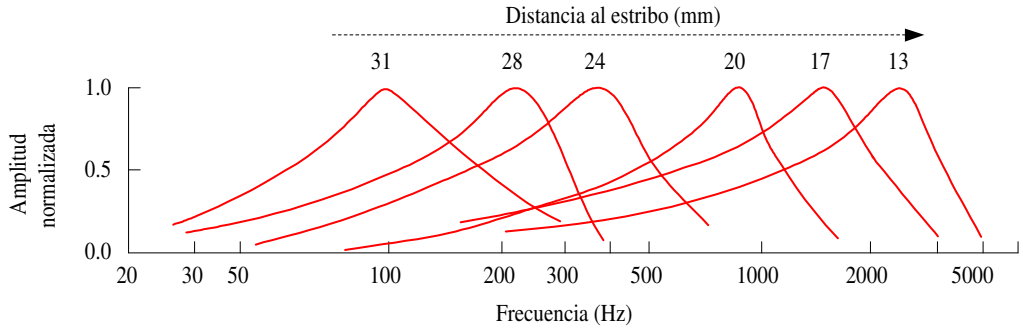

Figura 3.15: *Curvas de Resonancia: amplitudes relativas de las excursiones de la membrana basilar como función de la frecuencia de estimulación, para seis puntos a lo largo de la membrana. El estudio se realizó con cadáveres por lo cual algunos mecanismos activos no están presentes (adaptado de [9]).*

La transducción mecánico-eléctrica se produce en el denominado *órgano de Corti* ubicado a lo largo de toda la membrana basilar (Ver Figura 3.16). Ésta tiene lugar como respuesta a una curvatura de las *cilias* de las *células ciliadas*. Esta curvatura produce una variación en el potencial de membrana de las células; si las cilias se curvan hacia el cuerpo basal se produce una despolarización, mientras que si se curvan en el otro sentido se produce una hiper-polarización.

La excitación de las células ciliadas está determinada, en gran medida, por las excursiones de la membrana basilar. Sobre ella actúan las ondas de presión oscilatorias resultantes de la transmisión del sonido en las rampas vestibular y timpánica. De esta manera –dado que la amplitud de las vibraciones en distintos puntos de la cóclea varía con la frecuencia del estímulo– el grado en el cual es excitada una determinada célula ciliada es una función conjunta de su posición en la membrana basilar y de la amplitud del estímulo.

La curva de resonancia de la membrana basilar de la Figura 3.15 describiría con precisión la excitación de las células ciliadas en función de la frecuencia, si éste fuera el único factor que influyera en la vibración de las células ciliadas. Sin embargo, las propiedades mecánicas de las cilias y de la *membrana tectoria* que las cubre también influyen en la vibración de las células ciliadas. De hecho, la rigidez de las cilias, la masa y la elasticidad de la membrana tectoria también varían de un extremo al otro de la cóclea. Se ha registrado también cierto comportamiento "activo" de algunas células ciliadas[9]. Además de ello el penacho ciliar posee propiedades mecánicas especiales que derivan en un comportamiento no-lineal. Ésto parece explicar el conocido efecto de "oír" un tercer tono cuando solo se estimulo con dos [165]. Estas características del complejo célula-membrana tectoria tiene el efecto de limitar la sintonía de las células ciliadas a un ancho de banda de frecuencias más estrecho que el del punto de la membrana basilar donde se encuentra la célula. Se debe mencionar también que las células ciliadas se despolarizan solo durante la fase positiva de los estímulos sonoros produciendo un efecto de *rectificación de media onda* sobre las respuestas del nervio auditivo [155].

[9]Cuando éstas son estimuladas eléctricamente cambian su longitud.

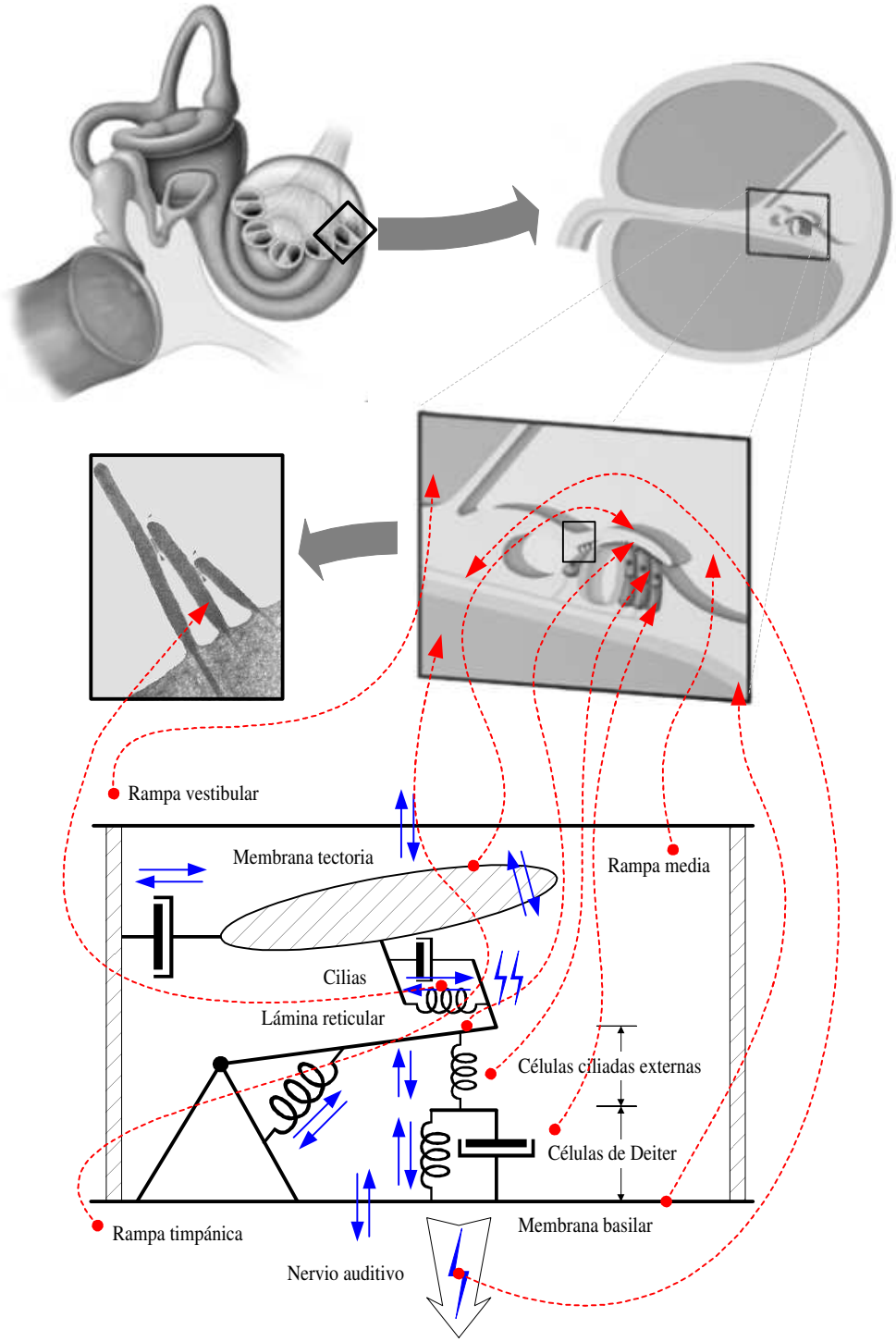

Figura 3.16: *Detalle del órgano de Corti y las células ciliadas (arriba). Diagrama esquemático que ilustra su funcionamiento (abajo). En el órgano de Corti, ubicado a lo largo de toda la membrana basilar, se produce la transducción mecánico-eléctrica.*

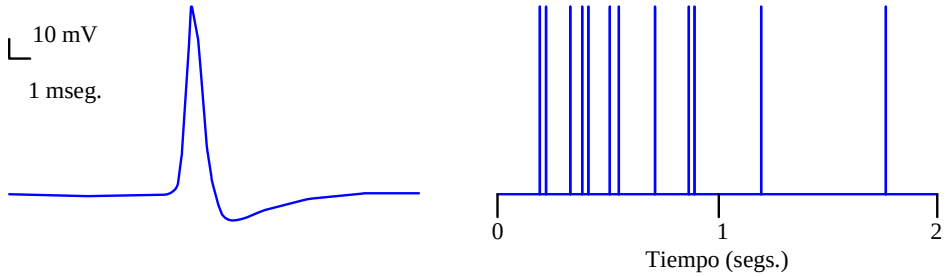

Figura 3.17: *Potencial de acción o pulso típico producido por la despolarización de una célula nerviosa o neurona (izquierda). Tren de pulsos característico producido por la despolarización repetida de una neurona como respuesta ante distintos estímulos de entrada (cada tiempo de disparo esta señalado con una barra vertical, derecha). Todo el código de comunicación neuronal está basado en estos trenes de pulsos (adaptado de [122]).*

3.4.3. Nervio auditivo y codificación nerviosa

Una pregunta fundamental en neurociencias está relacionada con la comprensión del código neuronal que se utiliza para organizar las distintas señales dentro del sistema nervioso. Este código está basado en la utilización de trenes de pulsos como el mostrado en la Figura 3.17. Estos trenes de pulsos se encuentran a todos los niveles del sistema, desde los transductores sensoriales hasta la corteza cerebral. En la Figura 3.18 es posible apreciar el registro simultáneo de un conjunto de neuronas corticales. En esta sección se discuten distintos aspectos que permiten explicar la codificación de los sonidos a nivel del nervio auditivo, mientras que en las secciones siguientes se describe lo ocurrido a lo largo del resto de la vía auditiva hasta llegar a la corteza.

El nervio auditivo está formado por la colección de axones periféricos correspondientes a las neuronas aferentes y eferentes que inervan a las células ciliadas. Aquí el interés principal se pondrá en la parte aferente, es decir aquellas fibras que llevan información desde la periferia auditiva en la dirección del sistema nervioso central. La respuesta de una fibra aislada puede describirse en términos de la frecuencia del correspondiente tren de pulsos, su fase y su patrón temporal de activación. Se considera que la respuesta de una fibra es estocástica, en el sentido que el patrón de disparo está relacionado de manera probabilística con las características del estímulo [193]. Aún sin estimulación acústica muchas fibras poseen respuesta espontánea, y ésta varía de fibra a fibra. Para el caso de tonos puros es posible suponer que existen tres características del estímulo que se deberían codificar a nivel nervioso: la intensidad, la frecuencia y la fase. La codificación de la fase es directa y tiene importancia principalmente en cuestiones de ubicación espacial de la fuente sonora. De acuerdo con lo presentado en la sección anterior se podría pensar que la frecuencia se codifica en términos de cuál es la fibra individual que dispara, y la intensidad en la tasa de disparo de los pulsos. Sin embargo, aunque ésto puede representar una primera aproximación, la codificación de los diferentes sonidos puede ser bastante más compleja y utilizar estrategias “mixtas” como se discutirá a continuación.

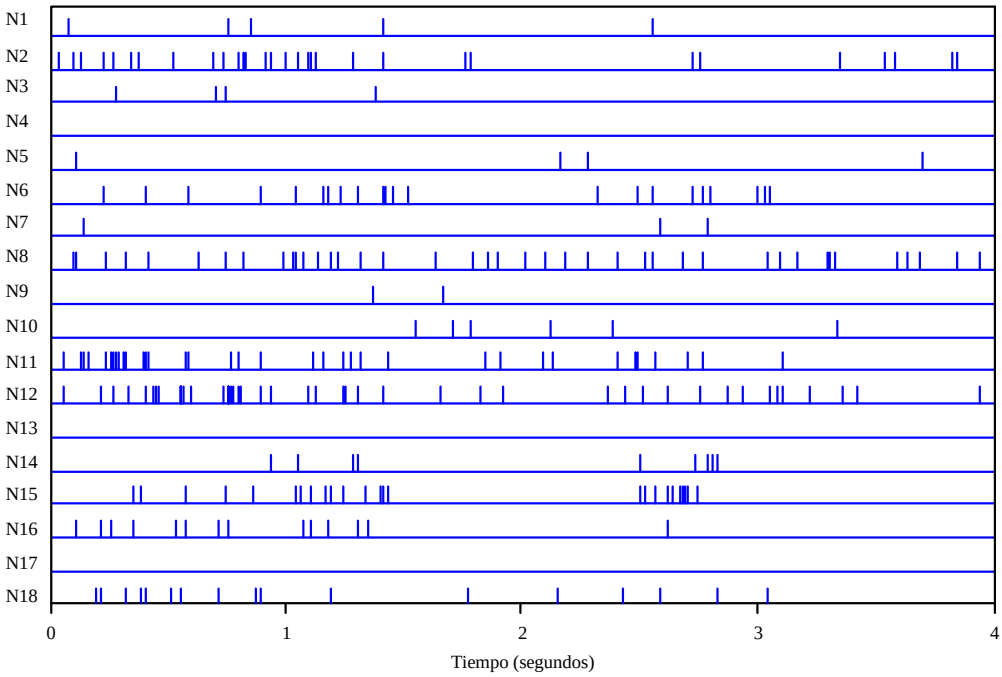

Figura 3.18: *Registro simultáneo de los tiempos de disparo de diferentes neuronas en la corteza estriada del mono durante cuatro segundos. La información se codifica en términos de cuál es la neurona que se activa, en conjunto con su frecuencia de disparo y su fase correspondiente (adaptado de [122]).*

Respuesta a estímulos simples

Como se ha visto, la membrana basilar está mecánicamente sintonizada con la frecuencia del sonido aplicado; por esta razón se puede pensar que las descargas nerviosas provenientes de zonas determinadas de la membrana basilar ya poseen la información de la frecuencia del estímulo. A esta forma de codificación de la frecuencia del estímulo se la denomina *mecanismo de la localización*. Los estudios fisiológicos iniciales de trenes de pulsos en fibras únicas del nervio auditivo brindaron información importante acerca de estos aspectos [101]. Estos estudios se realizaron en animales, principalmente en gatos, debido a la dificultad para realizarlos en humanos. Para ello se utilizaron fundamentalmente tonos puros. Una vez aislada una fibra se pudieron registrar impulsos de esa fibra única. De esta forma se obtenía una *curva de sintonía nerviosa* que trazaba los umbrales de respuesta en función de la frecuencia (Ver Figura 3.19). El mínimo de esta curva de sintonía (*frecuencia característica* o FC) indica el lugar a lo largo de la cóclea que ocupa la célula ciliada que excita la fibra. Ésto quiere decir que FC es la frecuencia para la cual la intensidad de estímulo necesaria para excitar la fibra es la mínima. Para estas fibras si estimulamos a la FC la intensidad del estímulo se codifica en la frecuencia o tasa de disparo (siempre por encima de su frecuencia espontánea). Se debe recalcar el hecho de que las fibras no responden a una única frecuencia, aunque requieren una mayor intensidad para ser excitadas fuera de su FC[10]. Ésto también sirve para codificar

[10]Esto se conoce como el problema del rango dinámico de una fibra nerviosa auditiva.

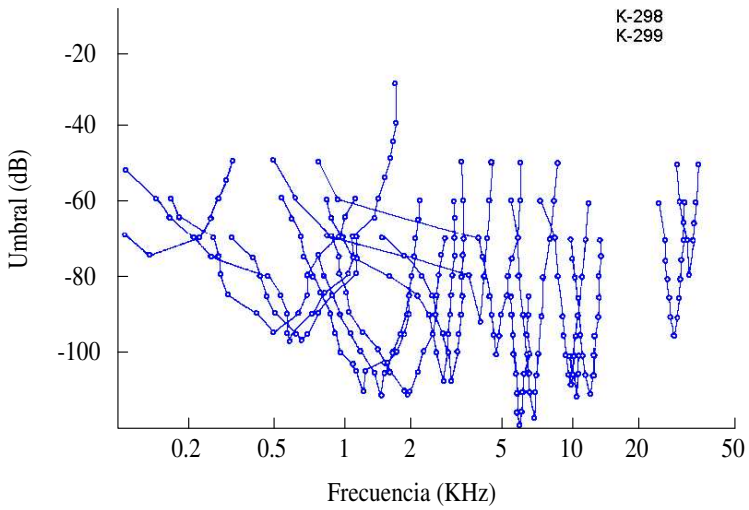

Figura 3.19: *Curvas de sintonía nerviosa: umbral de respuesta en función de la frecuencia de estimulación para varias fibras individuales del nervio auditivo de gato (adaptado de [101]). La frecuencia característica de una fibra es el mínimo de esta curva de sintonía y está relacionada con el lugar a lo largo de la cóclea que ocupa la célula ciliada que excita la fibra en cuestión.*

información acerca de la intensidad del estímulo (de acuerdo a la cantidad de fibras que responden).

En la Figura 3.20 se observa la curva de resonancia mecánica en un punto de la membrana basilar y la curva de sintonía de una fibra nerviosa que inerva a la célula ciliada en ese punto. La curva de resonancia muestra los niveles de presión sonora relativos requeridos para hacer vibrar la membrana en ese punto a una amplitud dada para varias frecuencias de sonido. La curva de sintonía muestra el umbral de la fibra nerviosa en función de la frecuencia del estímulo sonoro. Nótese que ambas curvas tienen frecuencias de corte similares, pero del lado de las bajas frecuencias la curva de sintonía posee una subida mucho más abrupta que la de resonancia. Se propusieron varios mecanismos para explicar esta aparente discrepancia entre las curvas de sintonía mecánicas y nerviosas. Estudios de la mecánica de la membrana basilar utilizando métodos más refinados mostraron una agudeza de sintonía mecánica bastante parecida a la de la sintonía neural [176].

Además de la percepción de la frecuencia de acuerdo a la posición de la fibra, para tonos de baja frecuencia (< 1 KHz) e intensidad moderada, las descargas nerviosas de una fibra determinada pueden "seguir" a los estímulos en frecuencia con una relación uno a uno. Ésto quiere decir que la información de la frecuencia se codifica también en la tasa de disparos. Sin embargo para tonos de frecuencias mayores ya no es posible seguir el "ritmo" tan de cerca. Entonces se recurre al fenómeno de excitación de varias fibras simultáneas, cada una con una fase diferente pero invariante. Este fenómeno, denominado respuesta *enganchada en fase*, permite la codificación de la frecuencia del estímulo en forma "distribuida" entre varias fibras. Este mecanismo funciona de manera confiable aproximadamente hasta los 3 KHz [155]. Este último modelo para la codificación de la frecuencia del estímulo se denomina *mecanismo temporal*.

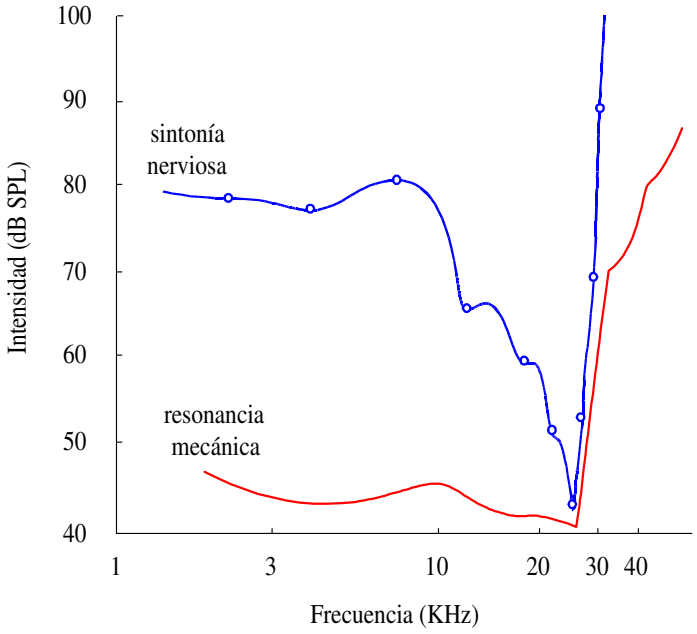

Figura 3.20: *Comparación entre resonancia mecánica y sintonía nerviosa en un punto de la membrana basilar (adaptado de [176]). Es posible observar que ambas curvas poseen frecuencias de corte similares, pero del lado de las bajas frecuencias la curva de sintonía nerviosa posee una subida mucho más abrupta que la de resonancia mecánica.*

Como resumen podríamos decir que existe acuerdo de que para la codificación de la frecuencia coexisten los dos mecanismos expuestos. Ésto es que para las bajas frecuencias se utiliza principalmente el temporal y para altas frecuencias principalmente el de localización. Sin embargo hay discrepancia acerca de la frecuencia a la cual comienza a reemplazarse uno por el otro [36]. Para la codificación de la intensidad también existe coincidencia acerca de un mecanismo mixto entre las tasas de disparo individuales y la cantidad de fibras que responden, según se ha explicado en esta sección.

Respuesta a estímulos complejos

La distinción entre estímulos simples y complejos es algo arbitraria. Se puede hablar de "complejo" en el sentido espectral cuando se tiene más de un tono puro. Complejidad también puede referirse al caso de señales no periódicas o aleatorias. A veces puede relacionarse con la cantidad de parámetros necesarios para una descripción matemática completa. Bajo este punto de vista todas las señales "naturales" son complejas. El interés aquí está puesto en aquellos estímulos sonoros similares al habla humana.

El estudio detallado de la respuesta del nervio auditivo a este tipo de estímulos requirió de algún tiempo. Los estudios iniciales con tonos puros daban solo una aproximación lineal para el estudio de un sistema "bastante" no lineal. Estas no linealidades no solo se dan a nivel nervioso sino inclusive a nivel de la mecánica coclear[11](como ya se ha

[11]A pesar de ello muchas de las pruebas clínicas para valorar la audición de uso habitual en la

visto en la Sección 3.4.2). Por ello no es posible comprender el comportamiento frente a estímulos complejos por la simple adición de los efectos producidos por sus componentes sinusoidales.

La continuación "natural" en este sentido de los estudios con tonos puros fue la utilización de tonos múltiples y señales de voz sintéticas basadas en modelos de producción del habla [133]. Con posterioridad se comenzó a trabajar con señales de voz reales [16]. Los estudios se continuaron realizando en animales[12].

En general se asume que la representación de la señal del habla en el nervio auditivo está compuesta por un numero finito de elementos (aproximadamente 30.000 fibras del nervio auditivo en el hombre) y las respuestas de cada elemento están determinadas por una secuencia compleja de estados distribuidos e iterativos que preceden la iniciación de los pulsos de descarga. El nervio auditivo puede considerarse una disposición ordenada de elementos arreglados de acuerdo a la FC. Las fibras en esta disposición responderán incrementando su probabilidad de descarga cuando el nivel del estimulo supera el umbral. El neurograma [184] es una representación directa de la información experimental de la estimulación del nervio auditivo, ordenada de acuerdo con la FC de las fibras individuales. En la Figura 3.21 se muestra un neurograma basado en las respuestas fisiológicas al sonido */ba/* sintetizado. Cada línea del neurograma representa tasa de disparo instantánea promedio de una fibra nerviosa. La FC de la fibra está dada a la izquierda. A pesar del parecido con el clásico espectrograma, el neurograma presenta información de manera distinta, utilizando otra forma de codificar los patrones generados "más a la medida del sistema auditivo". En todos estos trabajos se pudo encontrar las fibras nerviosas respondían como detectores de características sencillas, como ser la ubicación y seguimiento de las frecuencias formantes o la detección del *tiempo de ataque de la sonoridad*[13] (TAS, en inglés *voiced onset time* o VOT) [16]. Como se ha visto éstas constituyen pistas acústicas importantes para la discriminación de los fonemas. El examen fino de los patrones temporales de descarga de las fibras reveló además la codificación de otras características espectrales simples (como la representación directa de F_0). El efecto de enganche de fase discutido anteriormente resalta los picos espectrales en los sonidos complejos. La redundancia asociada a este mecanismo provee cierta robustez en la codificación y ésta es una de las razones por las cuales la información más importante del habla se concentra en las bajas frecuencias [63]. También se corroboraron algunos efectos de enmascaramiento de frecuencias en la presencia de estímulos simultáneos y no simultáneos. Para tener una idea del tipo de representación a nivel del nervio auditivo[14] se han desarrollado varios modelos que incluyen los principales aspectos discutidos en esta sección y las anteriores, y que se han validado mediante experimentos fisiológicos

actualidad continúan utilizando principios lineales debido a su simplicidad.

[12]Aunque es posible realizar extrapolaciones al caso del hombre, debe tenerse en cuenta que el procesamiento de sonidos como el habla puede ser diferente ya que se trata de criaturas que no poseen un lenguaje hablado (aunque pueden reconocer palabras). El problema parece ser mayor a medida que avanzamos en la vía auditiva hacia centros más especializados.

[13]Se denomina así al tiempo transcurrido entre la liberación de la presión sonora posterior a una consonante plosiva, es decir el momento de apertura de los labios, y el comienzo de la vibración de las cuerdas vocales en el fonema sonoro subsiguiente.

[14]A este tipo de representaciones se las refiere como *representaciones auditivas tempranas*.

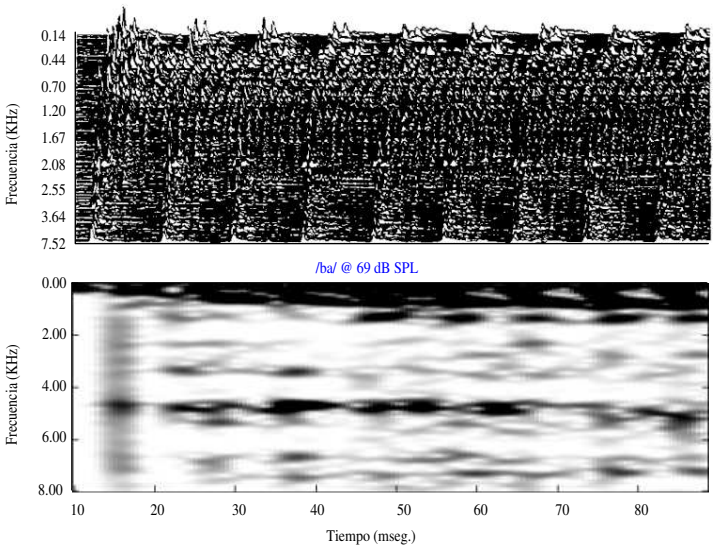

Figura 3.21: *Neurograma: tasas de disparo instantáneas promedio de las fibras del nervio auditivo del gato como respuesta a la estimulación acústica mediante la sílaba /ba/ sintética (arriba, tomado de [184]). Espectrograma de la misma sílaba pronunciada por un hablante masculino (abajo, nótese que el eje de frecuencias está invertido para facilitar la comparación).*

[211, 186]. La salida de estos modelos sería equivalente al neurograma ya descripto y suele denominarse *espectrograma auditivo*. En la Figura 3.22 puede apreciarse un espectrograma auditivo para un trozo de una oración, en comparación con el sonograma y el espectrograma tradicional correspondientes. Se puede notar fácilmente la mayor resolución frecuencial en la zona de las bajas frecuencias.

3.4.4. Vía auditiva

El nervio auditivo constituye sólo la primera parte de la denominada *via auditiva* (ver Figura 3.23). A lo largo de este camino, que lleva a la *corteza auditiva*, las señales nerviosas atraviesan una serie compleja de etapas de procesamiento en el tronco cerebral a través del *núcleo coclear*, el *núcleo olivar superior*, el *colículo inferior* y el *núcleo geniculado medio*. Las 30.000 fibras del nervio auditivo humano, se convierten en unos 100 millones de neuronas en cada lado de la corteza auditiva[15]. La organización tonotópica de la cóclea se mantiene en diversas partes de la vía auditiva, incluyendo la propia corteza. En el núcleo coclear se detectan algunos eventos acústicos simples, como comienzos y finales de fonemas y algunas transiciones. Ésto ha llevado a conjeturar que juega el papel de un modelo articulatorio inverso aproximado [137]. En el núcleo olivar superior se realiza la integración de la información proveniente de ambos oídos, cuyo objetivo principal es el de proveer la localización espacial de las fuentes de sonido. A partir de allí se continua en forma ascendente principalmente con información biaural, aunque existen centros que continúan procesando en forma monoaural [155]. La integración de las

[15]Se verá más adelante que esta sobre-representación es una característica importante para la robustez del sistema.

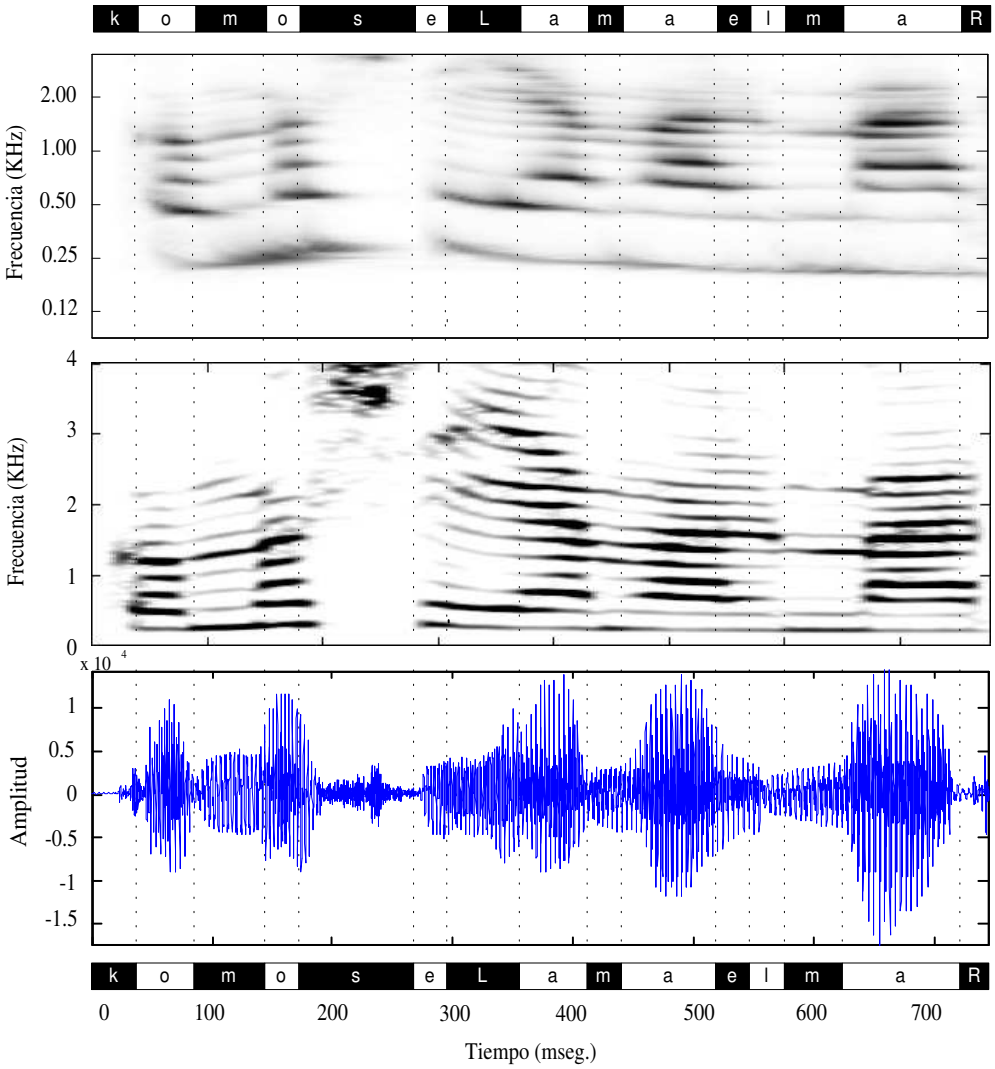

Figura 3.22: *Sonograma (abajo), espectrograma (centro) y espectrograma auditivo (arriba) de un trozo de la oración de la Figura 3.10. Los diferentes tonos de gris expresan la actividad neuronal de cada fibra del nervio auditivo ordenadas de acuerdo a su frecuencia característica.*

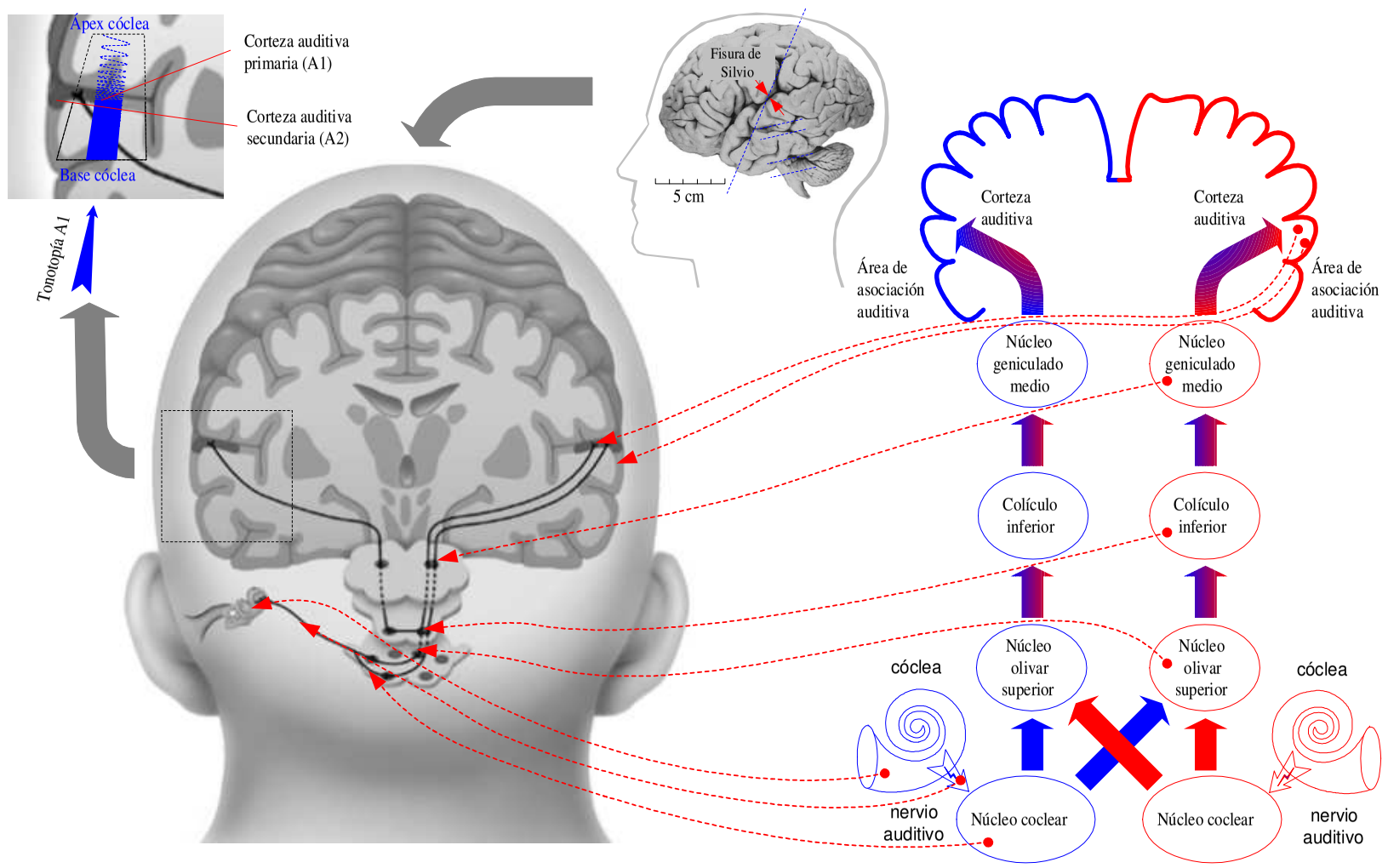

Figura 3.23: *Diferentes secciones de la vía auditiva y detalle de la corteza auditiva (izquierda). Diagrama esquemático que ilustra las conexiones y el trayecto seguido por la información en la vía, desde el nervio auditivo hasta la corteza (derecha).*

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
H. L. Rufiner; "Análisis y representación de la voz mediante técnicas no convencionales"
Universidad de Buenos Aires, Argentina, 2005.

diferentes vías continúa en el colículo inferior. Allí se procesan y analizan principalmente aquellos con patrones temporales especiales, como ser los modulados en frecuencia o con una duración específica. Antes de llegar a la corteza la información auditiva pasa por el núcleo geniculado medio, que es el primer lugar donde se generan respuestas específicas para ciertas combinaciones espectrales. Estas respuestas incluyen no solo la detección de combinaciones de frecuencias simultáneas, sino también de intervalos específicos entre dichas frecuencias. De esta manera a medida que se avanza hacia la corteza aparecen detectores de características cada vez más complejas [196].

3.4.5. Corteza auditiva

La corteza auditiva es la encargada de procesar los estímulos nerviosos para convertirlos en diferentes representaciones internas. Un dato neurobiológico importante es la arquitectura neuronal de la corteza auditiva. La corteza está formada por varias capas de células nerviosas, cada una de las cuales está constituida por tipos específicos de neuronas. Las capas corticales superiores, en el caso del hombre, poseen una gran proporción de la totalidad de las neuronas [185, 99]. La actividad neuronal sigue, en general, un patrón vertical que da lugar a la formación de columnas que a su vez están relacionadas lateralmente entre sí. Dentro de cada columna, una neurona perteneciente a una capa hace sinapsis directas sobre neuronas de la siguiente capa, o bien indirectamente, a través de interneuronas [99]. Ésto da lugar –teniendo en cuenta los retardos sinápticos– a que una neurona cualquiera de las capas más altas reciba simultáneamente información que fue generada en instantes distintos en la periferia, lo que permite establecer relaciones temporales complejas.

Gracias a las técnicas relativamente recientes de generación de imágenes funcionales como la *resonancia magnética funcional* [10] o la localización de dipolos mediante *potenciales evocados auditivos* (PEA) [18], es posible el estudio no invasivo de algunas funciones corticales en el hombre. Ésto ha permitido la identificación de las zonas que intervienen en el procesamiento del habla (ver Figura 3.2). En la Figura 3.24 se ha representado la localización promedio del dipolo de la componente P1 del PEA de latencia larga para diferentes frecuencias del estímulo. Ésto permite corroborar por ejemplo la tonotopía de la corteza. A pesar de ésto solo se conocen unas pocas características organizacionales de la corteza auditiva [186]. Ésta puede dividirse principalmente en dos áreas funcionales: la *corteza auditiva primaria* (AI) y la *corteza auditiva secundaria* (AII) (Ver detalle en Figura 3.23). La zona AI recibe información directa del núcleo geniculado medio y por lo tanto posee un mapa tonotópico preciso [155]. Se puede decir que AI posee un mapa topográfico de la cóclea, por lo que a veces se lo denomina también mapa *cocleotópico*. En disposición ortogonal al mapa tonotópico existe una organización en bandas de las propiedades biaurales. La zona AII posee una organización tonotópica menos precisa y posiblemente analiza sonidos más complejos. El área de Wernicke (Ver Sección 3.2) se ubica en el interior de AII [155].

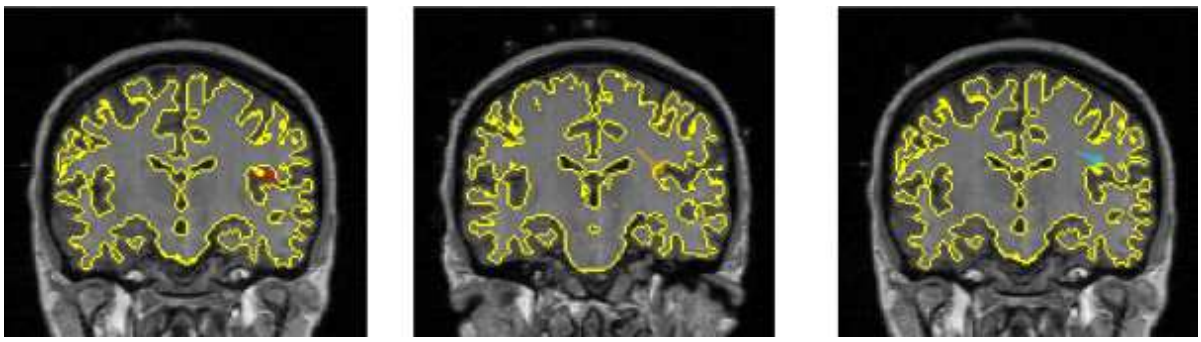

Figura 3.24: *Representación neuro-anatómica de la localización promedio del dipolo de la componente P1 del PEA de latencia larga en función de la frecuencia del estímulo (tomado de [18]). A partir del cambio de localización del dipolo es posible corroborar la tonotopía de la corteza auditiva.*

Representación cortical

En la Sección 3.4.3 se discutió acerca de la codificación neuronal a nivel del nervio auditivo. En esta sección se pretende introducir algunos conceptos que permitan comprender los aspectos sobresalientes de la codificación más complejos que se da a nivel de la corteza. En particular es de interés conocer como se codifican los rasgos distintivos del habla a nivel de la corteza auditiva. Se cree que el sistema auditivo ha aplicado principios de codificación eficiente para procesar a los sonidos naturales, especialmente el habla. Ésto parece muy razonable si se piensa que éstos son los sonidos "más importantes" de nuestro entorno. La teoría de la información provee conceptos generales que permiten abordar el tratamiento de los problemas de comunicación mediante señales. Entre estos conceptos aparece el de eficiencia de la codificación. Hace ya un tiempo que estos principios se han tratado de aplicar al código neuronal, pero es más recientemente cuando se ha obtenido cierto éxito [63]. Una versión neuronal de esta hipótesis de eficiencia en la codificación establece que el rol de los sistemas sensoriales "tempranos" es remover la redundancia estadística o aumentar la independencia entre las respuestas neuronales a estímulos naturales. A esta hipótesis suele agregarse otra que asegura que estos sistemas tienden a crear representaciones internas sumamente ralas, es decir teniendo en cuenta una cantidad importante de rasgos significativos de manera explícita (esto tiene su correlato en la sobre-representación de características a nivel cortical). De esta forma el cerebro crea un código eficiente mediante una representación rala e independiente de la señal, consistente principalmente en detectores de cambios en los picos espectrales y en los parámetros temporales (representaciones tiempo-frecuencia). Para llegar a validar estas hipótesis un posible camino consiste en armar un modelo sensorial que se base en ellas y tratar de contrastar las predicciones realizadas mediante este modelo con las respuestas reales. Entre las predicciones que han logrado validarse mediante estos modelos se puede mencionar la representación sensorial interna a nivel cortical a partir de los

denominados *campos receptivos espectro-temporales* (STRF)[16].

Campos receptivos espectro-temporales Como se mencionó anteriormente el enfoque tradicional para caracterizar la respuesta a nivel cortical basada en la utilización de tonos puros es inaplicable para un sistema como éste. Para que ésto funcione adecuadamente el sistema, con entrada a nivel sensorial y salida en la corteza, debería ser lineal e invariante en el tiempo. Por ello la respuesta frente a tonos puros constituye solo una primera aproximación al problema. A pesar de ello la mayoría de los estudios y experimentos tradicionales utilizan este tipo de estímulos (incluyendo por supuestos aquellos que permitieron caracterizar las diversas organizaciones tonotópicas) [186]. Ésto se agrava si se tiene en cuenta que la no-linealidad intrínseca de todo este sistema no es un mero accidente de la implementación biológica, sino que constituye un aspecto fundamental que le otorga características funcionales especiales (como su robustez al ruido, entre otras) [183]. La mayoría de las neuronas sensoriales de los niveles superiores poseen respuestas no lineales con propiedades complejas por lo que la caracterización completa de las mismas constituye un desafío importante aún sin resolver. Varios estudios recientes utilizando estímulos complejos combinados con análisis lineal y no-lineal han provisto una nueva visión acerca de las propiedades de estas respuestas en varios circuitos neuronales [199].

Entre estos estudios se encuentran aquellos que tratan de determinar el estímulo óptimo para una determinada neurona mediante la exposición de la misma a una batería de estímulos complejos. La mayoría de estos estudios utilizan para ello el denominado *método de la correlación inversa*. Queda claro, a partir de la discusión anterior, que los estímulos requeridos deben ser de banda ancha. El primer candidato con estas características es, por supuesto, el ruido blanco. Sin embargo los primeros experimentos que lo utilizaron presentaron problemas debido a que las respuestas de las neuronas corticales al ruido blanco eran particularmente bajas. Por ello a partir de allí se ensayaron diferentes estímulos que fueron mejorando las estimaciones. Entre los diferentes estímulos utilizados se pueden mencionar: ruido modulado espectro-temporalmente, secuencias de "cuerdas" (consistentes en varios tonos de frecuencias seleccionadas al azar), "ondas móviles" (en inglés *moving ripples*, constituidos por la superposición de varios tonos que varían su frecuencia de manera lineal) y sus combinaciones (lineales o temporalmente ortogonales (TORC)) o estímulos naturales (por ejemplo vocalizaciones) (Ver Figura 3.25).

Por ejemplo, para determinar el estímulo óptimo para una neurona mediante el método de la correlación inversa, es "tocada" a la neurona en cuestión. La neurona responde con un tren de pulsos. A continuación se promedian los patrones tiempo-frecuencia de los tonos durante un tiempo fijo antes de cada pulso para producir la estimación de los STRF. De esta forma se hacen evidentes las frecuencias que excitan o inhiben la neurona como función del tiempo anterior al disparo [215]. En la Figura 3.26 se puede apreciar un diagrama esquemático del proceso para estimar los STRF mediante la correlación inversa (o promediado sincronizado con los pulsos de descarga.). La motivación de este

[16]Esta predicción se ha validado inicialmente para el sentido de la visión y más recientemente para el caso de la audición.

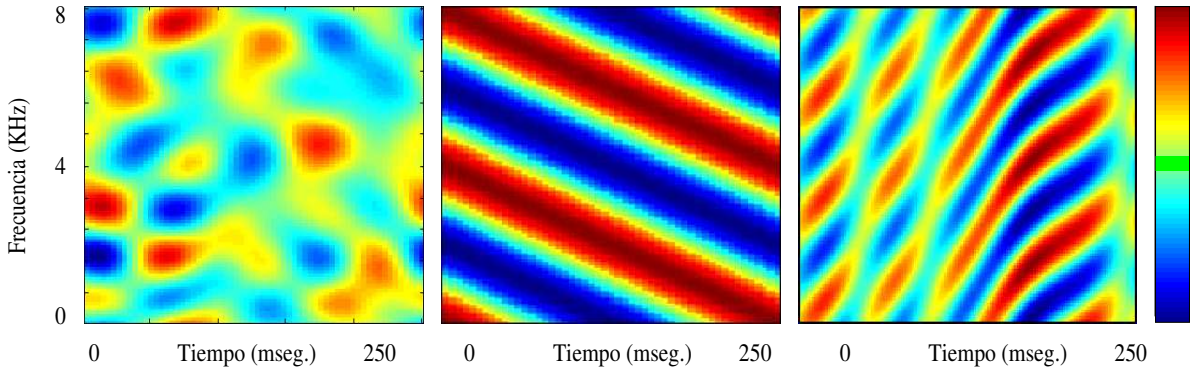

Figura 3.25: *Ejemplos de espectrogramas correspondientes a diferentes estímulos complejos utilizados para estimar los STRF: ruido modulado espectro-temporalmente (izquierda), ondas móviles (centro) y TORC (derecha). Tomados de [187, 102].*

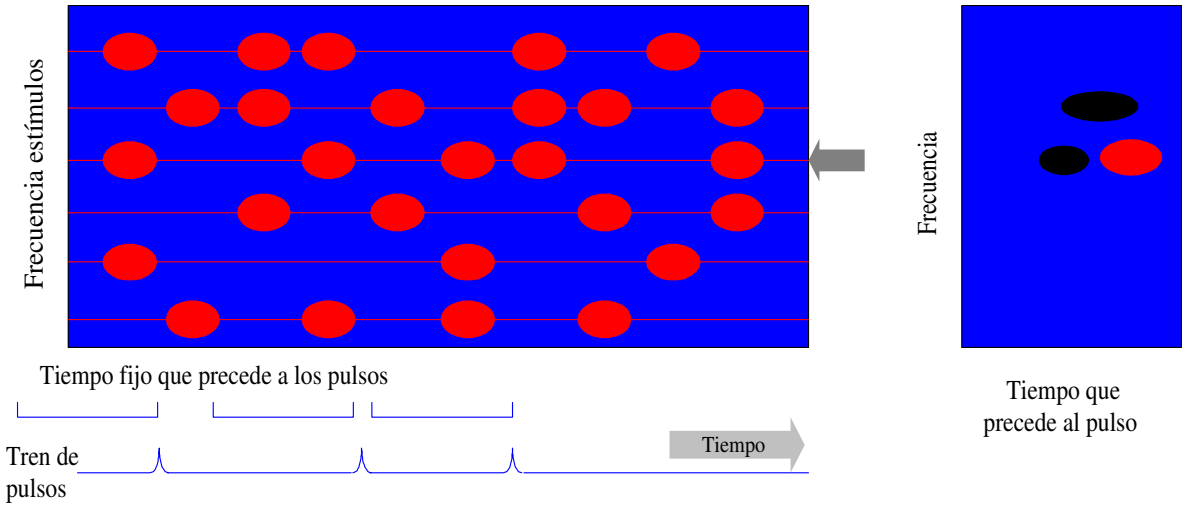

Figura 3.26: *Esquema simplificado del procedimiento para obtener los campos receptivos espectro-temporales de las neuronas de la corteza auditiva por medio de la correlación inversa estimulando mediante una secuencia de cuerdas (adaptado de [215]). En la figura la neurona es excitada por la frecuencia indicada con la flecha e inhibida por una frecuencia mayor.*

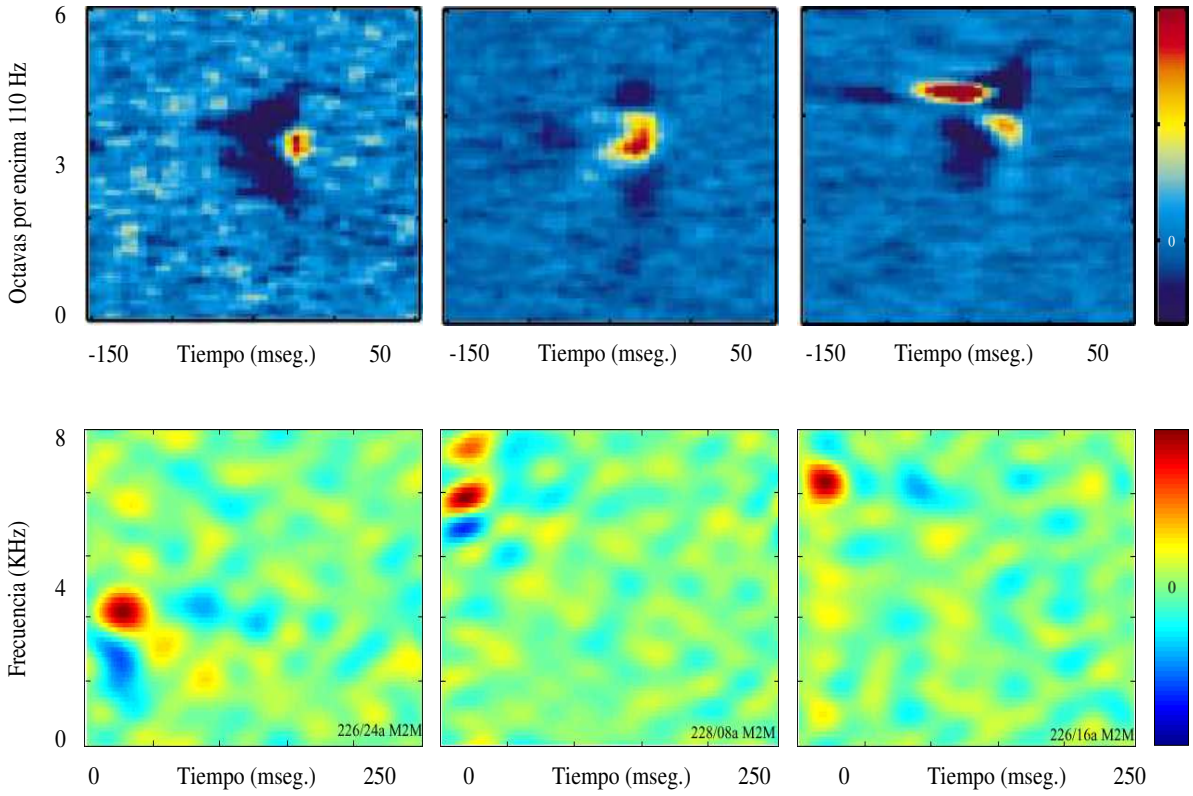

Figura 3.27: *Comparación entre los STRF de las células de la corteza auditiva AI obtenidos por correlación inversa para diferentes situaciones: Estímulo utilizado: cuerdas, Especie: mono aullador (arriba, tomados de [35]) ; Estímulo utilizado: TORCs, Especie: hurón (abajo, tomados de [187, 102]). Obsérvese la relativa similitud de las respuestas en cuanto su especificidad espectrotemporal.*

método es la de preservar aquellos patrones del estímulo que consistentemente provocan que la neurona dispare, mientras se logra disminuir en la promediación aquellos no correlacionados directamente con el disparo [102]. En la Figura 3.27 se puede apreciar los STRF estimados mediante este método para diferentes estímulos y especies animales [35, 187, 102]. Además del método de la correlación inversa, también se desarrollaron otros métodos que permitieron mejorar algunos aspectos y comparar diferentes estimaciones. Entre ellos se pueden mencionar la correlación polinomial de Laguerre y el análisis sinusoidal de estado estacionario.

Para los distintos métodos y estímulos se lograron estimaciones bastante consistentes, que permitieron inclusive la predicción de la respuesta de las neuronas a estímulos complejos no presentados previamente. Además se encontraron ciertos comportamientos cuasi-lineales frente a algunos estímulos complejos. Por ejemplo las neuronas se comportaban como si fueran lineales para combinaciones de ondas móviles.

A partir de estos estudios se ha podido determinar [187, 102] que las neuronas corticales responden bien solo a un pequeño conjunto de ondas móviles (como las del centro de la Figura 3.25) alrededor de un espaciado particular de picos espectrales y velocidades

temporales[17]. De esta forma codifican las características espectro-temporales como variaciones espectrales de hasta aproximadamente 1 ciclo/octava (en casos especiales hasta 4 ciclos/octava) y variaciones temporales de entre 2 y 20 Hz (en casos especiales hasta 100 Hz). Se encontraron células corticales para "todas" las frecuencias centrales, simetrías espectrales, anchos de banda, latencias y simetrías de respuesta al impulso.

Se puede decir entonces que AI descompone el espectro de entrada en diferentes "canales" sintonizados espectral y temporalmente. De manera equivalente, una población de células, sintonizadas alrededor de diferentes parámetros de ondas móviles, pueden representar efectivamente el espectro de entrada en múltiples escalas. A partir de este comportamiento se han simulado una serie de filtros de ondas "teóricos" [187, 102] para generar un modelo de la representación cortical. En la Figura 3.28 se puede observar el análisis realizado por diferentes células de la corteza basado en este modelo teórico para un trozo de una oración del español. Este tipo de modelo ha sido validado también con pruebas de inteligibilidad en humanos. Como comentario final acerca de las STRF se puede decir que están basadas en un concepto muy flexible que ha permitido avanzar en la comprensión de los mecanismos corticales bastante más allá de las curvas de sintonía y las tonotopías. Sin embargo constituyen todavía una estimación lineal y por lo tanto no explican todas las propiedades de las respuestas corticales.

3.5. Percepción

Hasta aquí se han descripto las representaciones corticales más básicas que nos dan una primer idea de los aspectos importantes a preservar e incluir en una representación "óptima" de la señal de voz. Sin embargo, la percepción acústica completa constituye en sí un fenómeno integrador complejo que va bastante más allá de estas representaciones internas. Para intentar comprender algunos otros aspectos significativos de este fenómeno se han realizado diversos experimentos perceptuales.

Está claro que para la decodificación completa del mensaje contenido en el habla se debe hacer uso de todos los niveles utilizados en su codificación (que se describieron en la Sección 3.2). Como ya se explicó el uso de información redundante en varios de estos niveles asegura algunas de las propiedades de robustez de la comunicación oral humana. Sin embargo hace falta todavía recorrer un largo camino para descubrir la forma exacta en la que este proceso se realiza en nuestro cerebro. Un concepto importante asociado a la percepción del habla es el de la inteligibilidad que se discutirá a continuación.

3.5.1. Inteligibilidad

Se podría definir a la inteligibilidad como un rasgo subjetivo medible que da cuenta de la correcta interpretación del habla. Está relacionada con las medidas de calidad del habla, aunque posee diferente significado. Ésto se debe a que para medir la calidad no se requiere interpretar o comprender el habla, como en el caso de la inteligibilidad. De todas maneras, en general alta calidad implica buena inteligbilidad pero lo contrario no

[17] Ésto es en lo que respecta al método explicado.

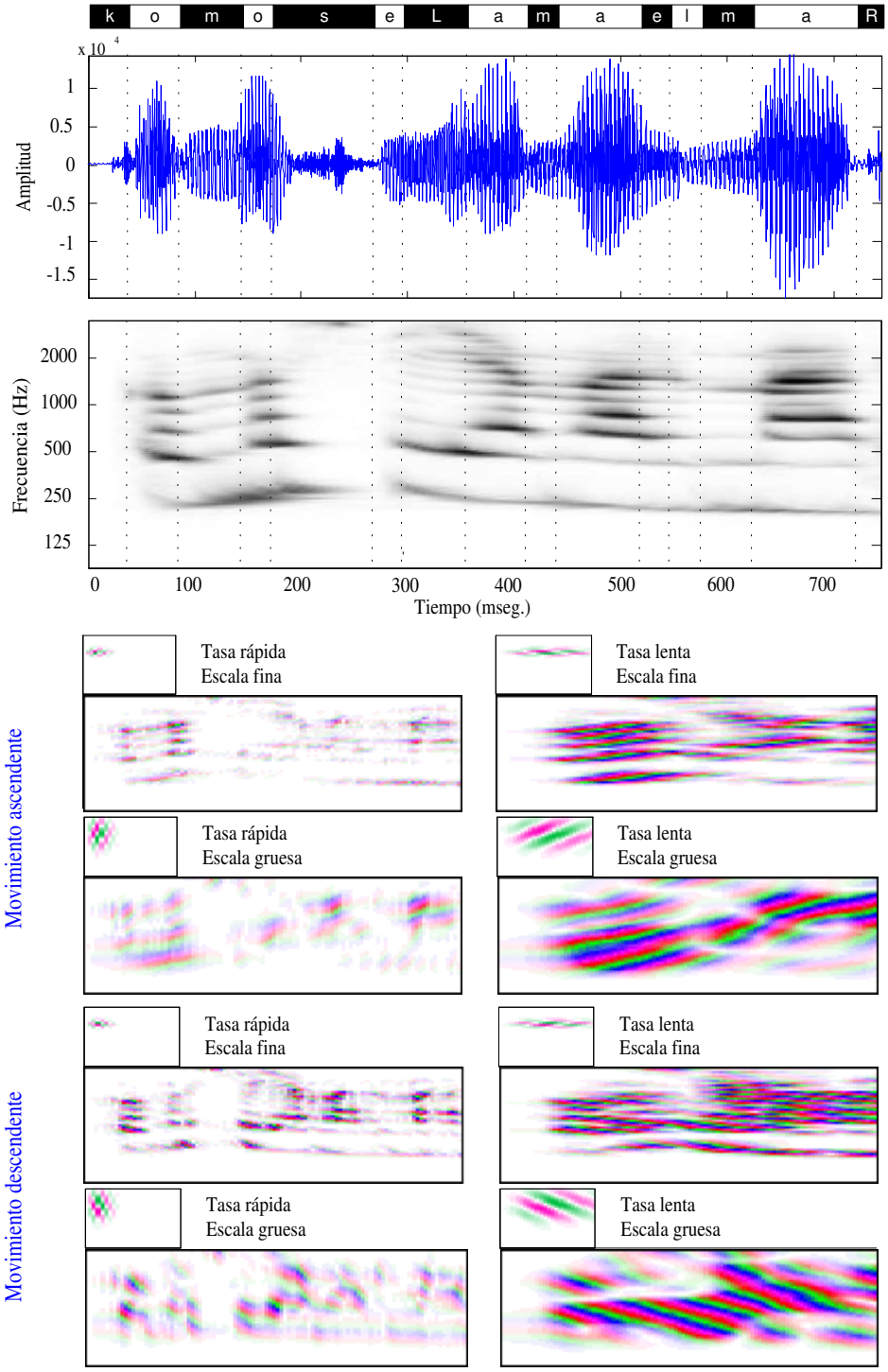

Figura 3.28: *Sonograma (arriba), espectrograma auditivo (centro), modelado de los filtros corticales "multiresolución" y sus salidas respectivas (abajo). La representación se obtuvo mediante el modelo cortical descripto en [187]. Este modelo se ha utilizado con éxito para predecir la inteligibilidad en sujetos normoyentes frente a diferentes tipos de distorsiones.*

siempre se cumple. De esta forma ambas medidas constituyen diferentes dimensiones para el análisis del habla.

Existen varias pruebas que permiten medir la inteligibilidad. En general se basan en la presentación de material hablado a un conjunto de sujetos y el análisis de la cantidad de aciertos o errores en la identificación o interpretación del mismo. El material puede estar compuesto por sonidos básicos, números, palabras de uso común, monosílabos, bisílabos o frases, dependiendo de los niveles que se pretende involucrar en el análisis. Resulta también de interés determinar los cambios debidos a variaciones en las condiciones del estudio, como por ejemplo la diferencia entre los resultados para habla limpia y para habla ruidosa a diferentes relaciones señal-ruido.

3.5.2. Algunos experimentos perceptuales

Se han realizado numerosos experimentos que permiten resaltar diversas características importantes para la percepción del habla. La mayoría de estos experimentos se han realizado para el idioma inglés, aunque en términos generales suelen ser de aplicación a otros idiomas. Para el caso del idioma español pueden mencionarse algunos estudios específicos de la interacción entre pistas acústicas o rasgos distintivos y su efecto en la percepción (como por ejemplo [67]).

Entre los estudios más significativos figuran los trabajos de Fletcher [51] sobre el denominado *índice de articulación*. Fletcher realizó pruebas de inteligibilidad con fonemas filtrados en diferentes bandas. De esta forma demostró que la probabilidad de error en la percepción de los fonemas cuando se utilizaba la señal sin filtrar era igual al producto de las probabilidades de error de las bandas individuales[18]. Ésto implica la independencia estadística entre los errores de una banda con respecto a los de otras bandas, lo que se interpretó en términos de que, a nivel perceptual, la información "temprana" se procesa en varios "canales" separados de manera independiente, integrándose recién en los niveles superiores para producir el reconocimiento. Este mecanismo de percepción posee la ventaja de que si la información de algún canal se corrompe, la probabilidad de error en ese canal aumenta mucho, y su influencia en el error total se vuelve casi despreciable. Por lo tanto ésto le confiere propiedades de robustez a través de la "eliminación" de aquella información poco confiable.

Los resultados de los estudios discutidos, y de otros similares, muestran que los humanos pueden reconocer el habla con una exactitud relativamente alta incluso en presencia de una cantidad limitada de pistas espectrales (debido al filtrado). Ésto sugiere que el espectro de la señal de voz contiene cantidades significativas de redundancia. Los estudios también apuntan a la habilidad de los oyentes humanos para integrar fácilmente las pistas acústicas provenientes de regiones de frecuencias diferentes (disjuntas) en el proceso de percepción del habla.

Es claro que la sincronía y temporización de los eventos acústicos es fundamental para la inteligibilidad del habla. Como ya se ha visto, características acústicas como el VOT, o la duración de períodos de silencio pueden ser signos o pistas para la percepción de las

[18]Con posterioridad se encontraron resultados similares para otros tipos de descomposiciones de la señal de voz, por ejemplo a diferentes escalas temporales.

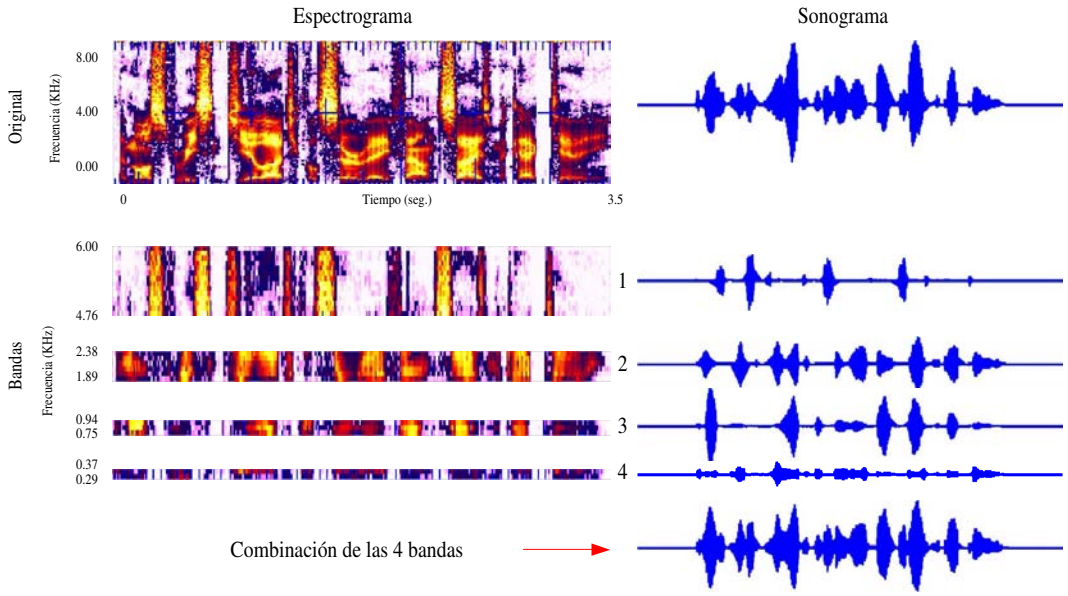

Figura 3.29: *Representaciones en el tiempo y en la frecuencia de las señales utilizadas en los experimentos perceptuales de Greenberg para demostrar el efecto sobre la inteligibilidad de diferentes cantidades de asincronía [65]. Mediante estos experimentos se demuestra que la sincronía entre los eventos acústicos resulta fundamental para la inteligibilidad del habla.*

distintas categorías fonéticas de los sonidos del habla. En escalas temporales mayores se sabe también que la caracterización de otros eventos como la duración de las sílabas o el seguimiento de la entonación, o la envolvente temporal, también resultan importantes en la discriminación de las palabras. La importancia de la sincronía en la percepción y discriminación de fonemas y sílabas se ha comprobado en enfermedades donde esta sincronía está comprometida [105]. Por otra parte Greenberg demostró [65] que para entender el lenguaje hablado no se requiere un análisis tiempo-frecuencia detallado de la señal de voz. Una representación extremadamente rala conformada solo por 4 canales de tan solo 1/3 de octava de ancho de banda (Ver Figura 3.29), puede en determinadas circunstancias dar una inteligibilidad casi perfecta. En este caso cantidades relativamente pequeñas de asincronía ($>$ 25 mseg) sobre las bandas de esta representación pueden resultar en una degradación significativa de la inteligibilidad[19]. Asincronías mayores a 50 mseg pueden causar un profundo impacto en la inteligibilidad.

3.6. Comunicación en condiciones adversas

Como se mencionó anteriormente el habla puede sufrir diferentes transformaciones antes de llegar al oyente. Las transformaciones que más afectan el proceso de la comunicación son aquellas relacionadas con distorsiones provocadas en el canal de transmisión,

[19]Se debe tener en cuenta que este tiempo es equivalente a la duración de algunos fonemas "cortos", por lo que desde el punto de vista perceptual no sería adecuado decir que constituye un tiempo muy pequeño.

entre las que principalmente se pueden mencionar la presencia de ruido y la reverberación.

Estas transformaciones degradan las pistas acústicas presentes en la señal y por ello se habla en este caso de percepción del habla en condiciones adversas. Ésto afecta directamente la inteligibilidad del habla. Se debe volver a resaltar aquí que los efectos de distorsión debidos al canal no serían los únicos que actúan dado que, como ya se explico, el propio hablante modifica su fonación en presencia de ruido. Sin embargo el objetivo de esta modificación es el de mejorar la inteligibilidad[20]. En esta sección se discutirán estos efectos y se mostraran algunos ejemplos de como afectan la comunicación tanto en humanos como en dispositivos artificiales (sistemas de ASR).

3.6.1. Ruido y reverberación

En la Figura 3.30 puede apreciarse un trozo de la oración "¿Cómo se llama el mar...?" ya presentada anteriormente y la misma señal "sucia" con ruido pseudoaleatorio con distribución gaussiana y relación señal-ruido de 10 dB. Se puede observar claramente como este tipo de ruido enmascara algunas pistas acústicas importantes (por ejemplo la zona de alta energía de la */s/*). En general el enmascaramiento se traduce en una especie de pérdida del contraste espectral que dificulta la percepción. A pesar de ello las representaciones internas en la vía auditiva poseen cierta robustez para este tipo de degradación.

La reverberación es el efecto producido por los rebotes de la señal en las superficies de las paredes y otros objetos de un recinto. Este efecto es el responsable del famoso "eco". En la Figura 3.31 se muestra la misma señal del ejemplo anterior, pero acompañada de una versión de la misma señal pronunciada en una cuarto de oficina con unos 200 mseg de tiempo de reverberación. Se observa claramente el "borroneado" horizontal producto de la convolución de la señal con la respuesta al impulso del cuarto. Ésto produce otra vez el enmascaramiento de algunas pistas acústicas y la aparición o sostenimiento de otras más allá de sus duraciones originales (lo que se aprecia principalmente para el caso de los formantes).

3.6.2. Humanos y máquinas

El oyente humano ha demostrado una importante robustez a las degradaciones de la señal de voz mencionadas anteriormente. Lippmann [119] realizó un estudio comparativo del desempeño de humanos y máquinas en tareas de reconocimiento del habla, con tamaños de vocabulario de entre 10 y más de 65,000 palabras. Los resultados de las comparaciones muestran que, incluso bajo las condiciones silenciosas, las tasas de error de los dispositivos artificiales son un orden de magnitud superiores que aquéllas de los humanos (Ver Figura 3.32). Esta diferencia de desempeño se acentúa aún más en la

[20] A pesar de ello ésto puede ser un problema para un sistema artificial ya que las modificaciones suelen ser importantes, de tipo no lineal y persisten aunque logre limpiarse la señal. Otros cambios en el hablante, como por ejemplo la velocidad de pronunciación, o cualquier alejamiento de las condiciones de entrenamiento (desapareamiento), pueden resultar perjudiciales para estos sistemas.

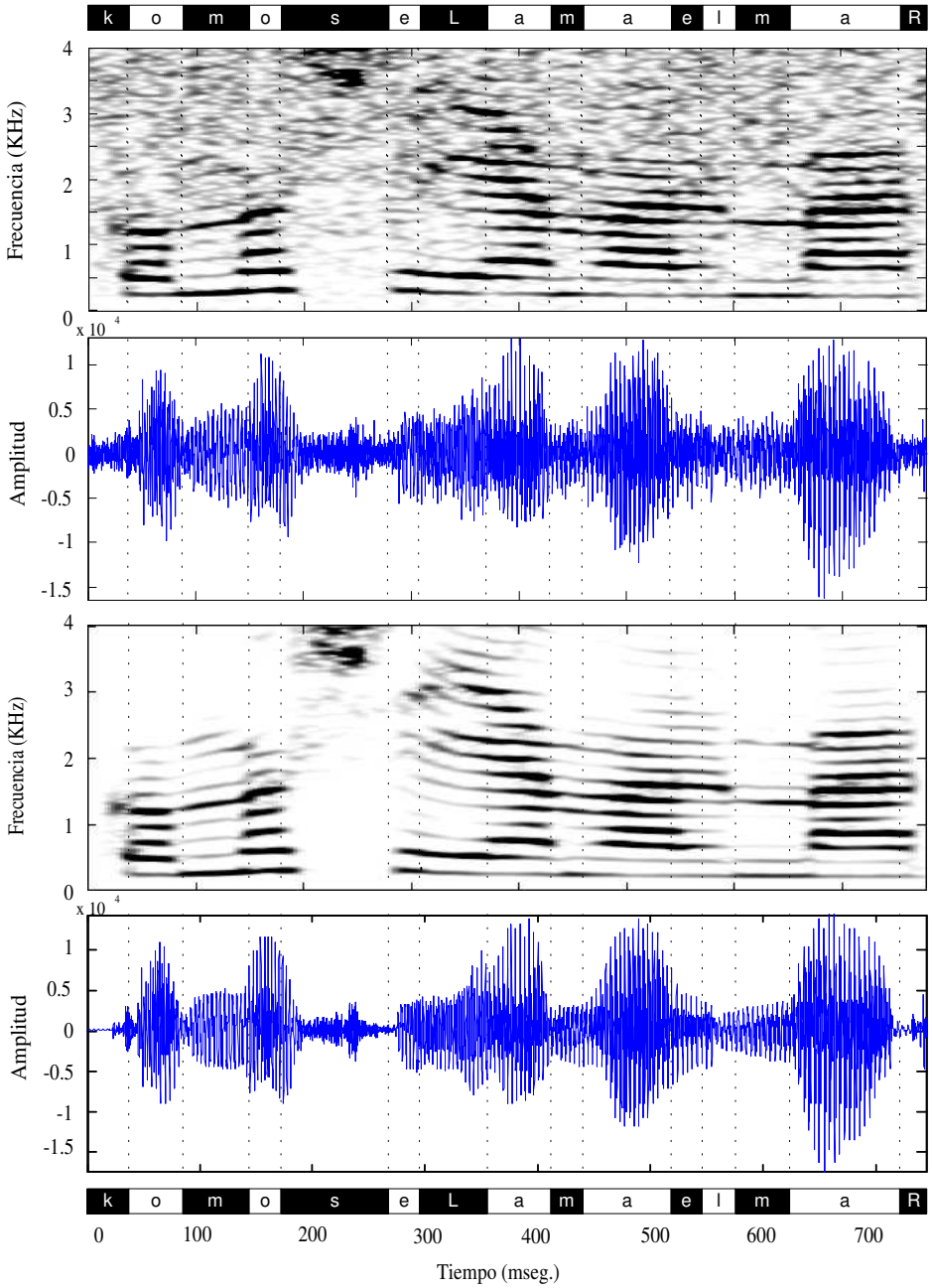

Figura 3.30: *Sonograma y espectrograma de un trozo de la oración de la Figura 3.10 "¿Cómo se llama el mar...?" (abajo) y la misma señal "sucia" con ruido pseudoaleatorio con distribución gaussiana y relación señal-ruido de 10 dB (arriba). Se puede apreciar como en el espectrograma ruidoso se ocultan rasgos acústicos importantes para una adecuada inteligibilidad.*

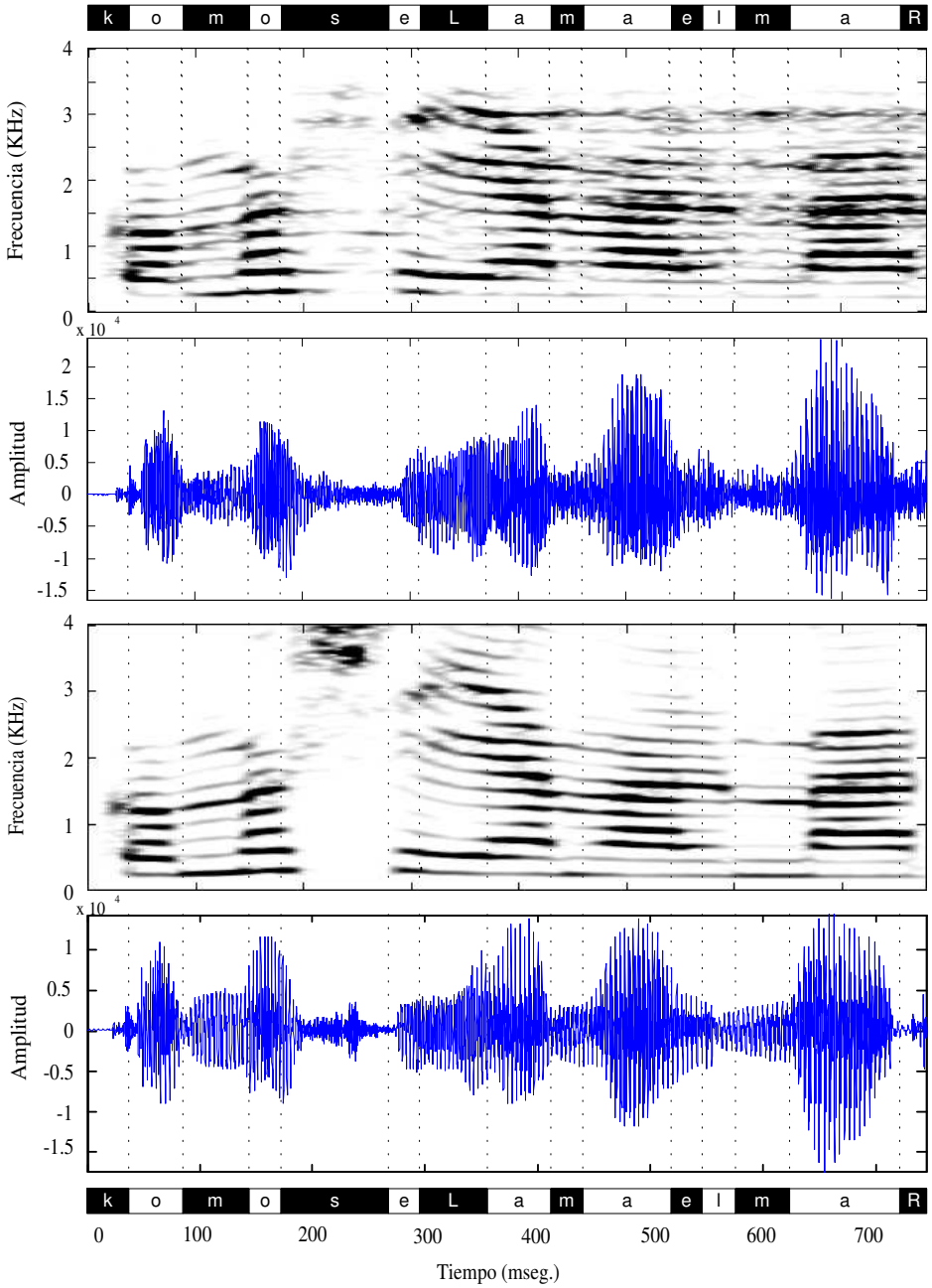

Figura 3.31: *Sonograma y espectrograma de un trozo de la oración de la Figura 3.10 "¿Cómo se llama el mar...?" (abajo) y la misma señal pronunciada en una cuarto de oficina con 200 mseg de tiempo de reverberacion (arriba). Es posible observar el efecto de la reverberación como un "borroneado" horizontal producto de la convolución de la señal original con la respuesta al impulso del cuarto.*

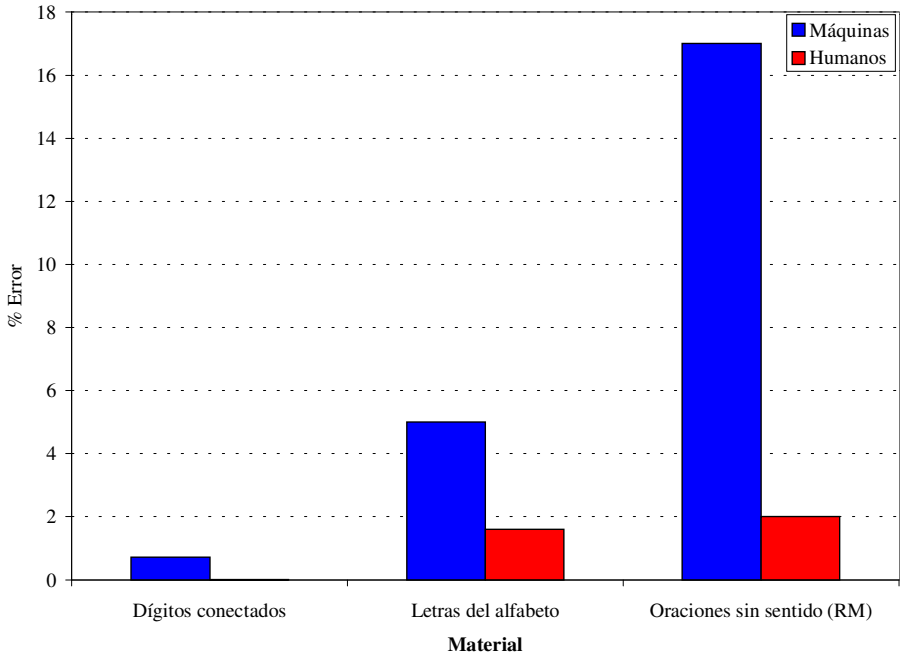

Figura 3.32: *Comparación entre los errores cometidos por oyentes humanos y máquinas para tareas de reconocimiento de material hablado "limpio" de diferente complejidad (adaptado de [119]). Obsérvese la gran diferencia de desempeño.*

presencia de ruido (Ver Figura 3.33) o cambios en el canal de transmisión (Ver Figura 3.34). Ésto ocurre a pesar de que las pruebas se realizaron con dispositivos que incluían algún mecanismo para compensar el ruido o adaptarse a modificaciones con respecto a las condiciones de entrenamiento originales. Estas diferencias sugieren que los mecanismos de percepción y reconocimiento humanos son todavía muy diferentes de los utilizados por las máquinas.

3.7. Comentarios de cierre del capítulo

En este capítulo se han presentado diferentes aspectos del proceso de la comunicación humana que podrían explicar, total o parcialmente, las diferencias de robustez y adaptación evidenciadas en las experiencias comparativas de los humanos con las máquinas. El desafío consiste en integrarlas adecuadamente en las diferentes etapas de los sistemas artificiales, comenzando por el procesamiento y la representación de la señal de voz.

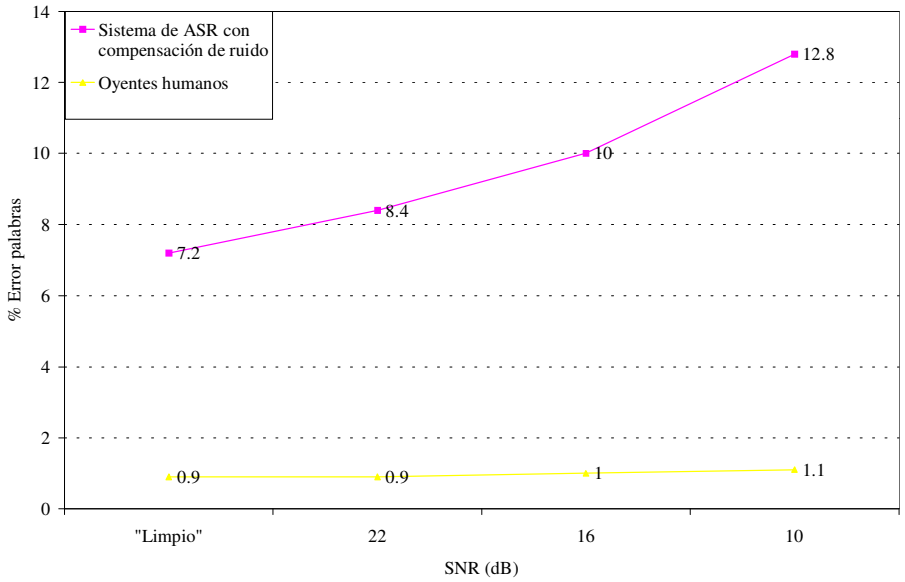

Figura 3.33: *Comparación entre los errores cometidos por oyentes humanos y un sistema de ASR con compensación de ruido, para tareas de reconocimiento de material hablado para diferentes SNR (Base de datos: Wall Street Journal, 5000 palabras, ruido aditivo de automóvil, adaptado de [119]). El desempeño en condiciones limpias es sustancialmente mejor para los humanos, y además resulta menor la degradación a medida que aumenta el ruido.*

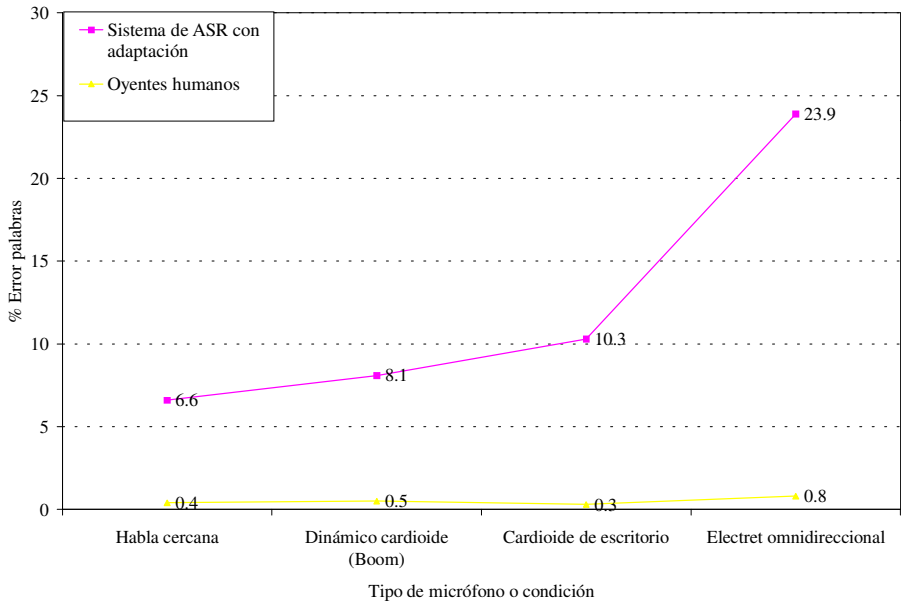

Figura 3.34: *Comparación entre los errores cometidos por oyentes humanos y un sistema de ASR con adaptación, para tareas de reconocimiento de material hablado en diferentes condiciones (Base de datos: NA Business News, habla leída, sala de cómputo, adaptado de [119]). El comportamiento es similar al de la Figura 3.33.*

Capítulo 4

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

4.1. 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 [71]. 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 3, 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.

Son precisamente estos enfoques no-convencionales a los que se les ha dado un mayor énfasis en el desarrollo de este trabajo. Algunos métodos no convencionales ya han sido presentados en el Capítulo 2 en las secciones de "Modelización de señales" y "Análisis estadístico de datos" (como por ejemplo la representación rala y el análisis de componentes independientes). 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

[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.

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 4.2). Las siguientes dos Secciones (4.3 y 4.4) 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 4.3.3) 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 4.4.2 (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 4.5) 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 3. En la Sección 4.6 se revisarán los aspectos relacionados con las características de robustez al ruido y a las distorsiones de los análisis presentados. Finalmente se realizarán algunos comentarios generales y se presentará el siguiente capítulo (Capítulo 5) donde se discutirán algunos ejemplos de aplicación de Onditas para el caso de la señal de voz.

4.2. 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 (por ejemplo [37] o [160]).

4.2.1. 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 [52].

Definición 4.1 *Sea $x(t) \in L^2(\mathbb{R})$ entonces su transformada de Fourier $X(f)$ existe y puede calcularse mediante:*

$$X(f) = \left\langle x(t),\ e^{j2\pi ft} \right\rangle = \int_{-\infty}^{\infty} x(t)\ e^{-j2\pi ft} dt. \tag{4.1}$$

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

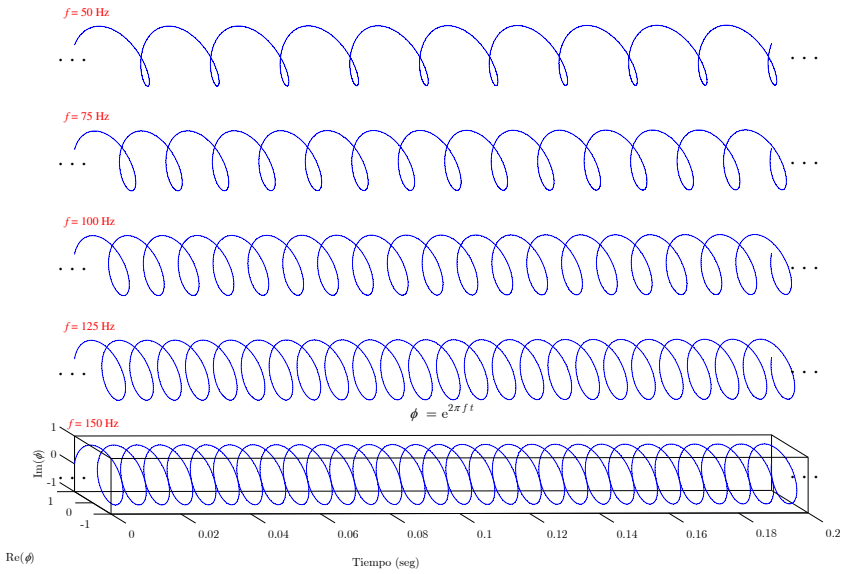

Figura 4.1: *Ejemplos de algunas exponenciales complejas de distintas frecuencias que forman la base de la FT. Estas exponenciales complejas "eternas" son las autofunciones de los sistemas LTI.*

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", ver Figura 4.1). 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:

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

4.3. 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 4.2 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 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 4.5 (aunque no todos resultan estrictamente lineales). Por ejemplo se ha utilizado ampliamente para los coeficientes LPC, el cepstrum, o incluso la energía [37]. Este "parche" al análisis estacionario se denomina también de *análisis de corta duración*. La idea general consiste

[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 $\tau \in \mathbb{R}$.

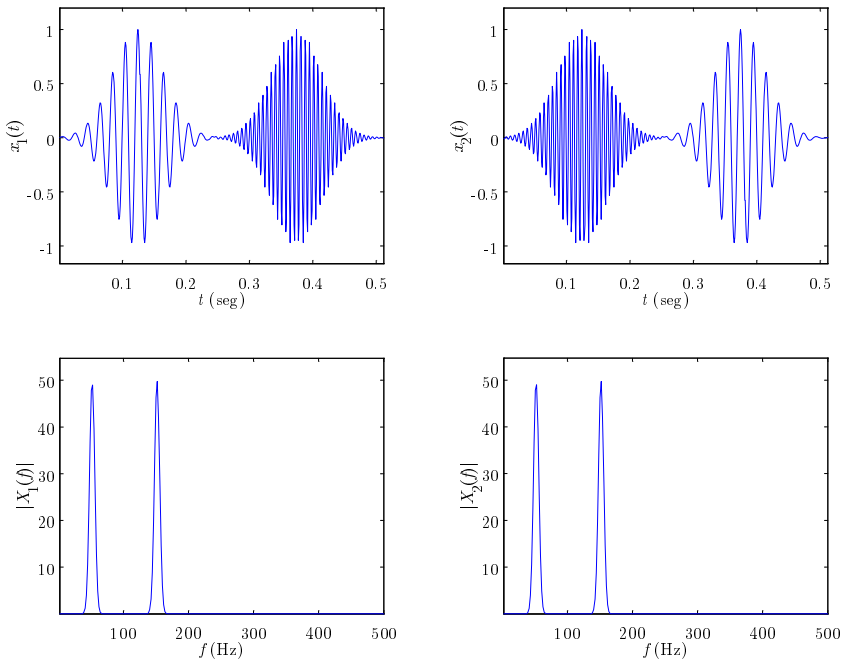

Figura 4.2: *Dos señales $x_1(t)$ y $x_2(t)$ formadas por combinaciones de dos tonos ventaneados de 10 Hz (A) y 30 Hz (B) (arriba) y sus respectivas magnitudes espectrales $|X_1(f)|$ y $|X_2(f)|$ calculadas mediante la FT (abajo). Es posible observar como el espectro no refleja los cambios en la secuencia de aparición de los tonos.*

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 [37]. Las ideas anteriores dan lugar a la siguiente definición .

Definición 4.2 *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:*

$$x_v(t;\tau) = x(t)g(t-\tau), \tag{4.2}$$

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 4.3).

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

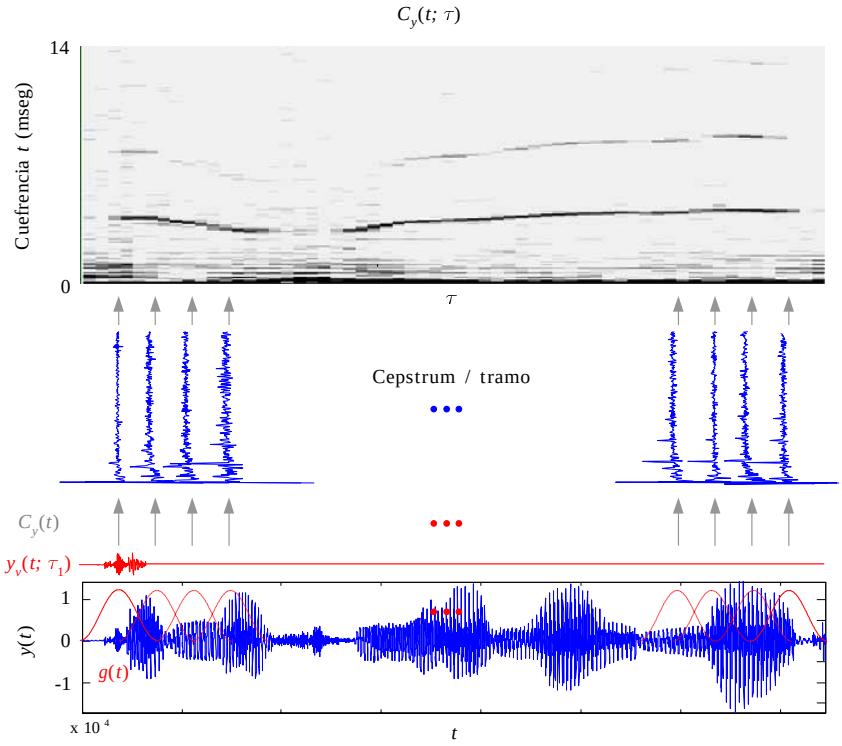

Figura 4.3: *Esquema del análisis por tramos ejemplificado para el caso del cepstrum real de corta duración de una señal de voz. Como se verá en la Sección 4.5 el primer pico del cepstrum está asociado a la frecuencia de entonación de la voz en los fonemas sonoros o F_0, cuya variación puede apreciarse claramente en este análisis.*

4.3.1. Señales analíticas y frecuencia instantánea

La *frecuencia instantánea* [49] 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.

Definición 4.3 *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:

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

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

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

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.

4.3.2. 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 (4.1), fue adaptada por primera vez por Gabor para definir $S(t, f)$ como sigue [54].

Definición 4.5 *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 (4.1) de las señales ventaneadas $x(t)\, g^*(t-\tau)$ constituye la transformada de Fourier de corta duración:*

$$S_F(\tau, f) = \int_{-\infty}^{\infty} x(t)\, g^*(t-\tau)\, e^{-j2\pi ft} dt. \tag{4.3}$$

Esta transformación mapea la señal $x(t)$ en una función bidimensional en el plano tiempo-frecuencia (τ, f). El parámetro f en (4.3) 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, (4.3) 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, 154].

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 4.4).

En la Figura 4.5 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.

[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.

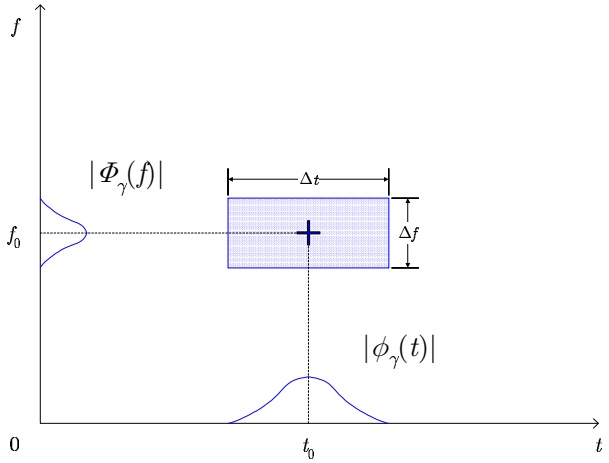

Figura 4.4: *Rectángulo de Heisenberg que representa al átomo $\phi_\gamma(t)$ en el plano tiempo-frecuencia (adaptado de [127]). La localización y las características de este rectángulo dependen de la dispersión tiempo-frecuencia de $\phi_\gamma(t)$, la que está condicionada por el principio de incertidumbre.*

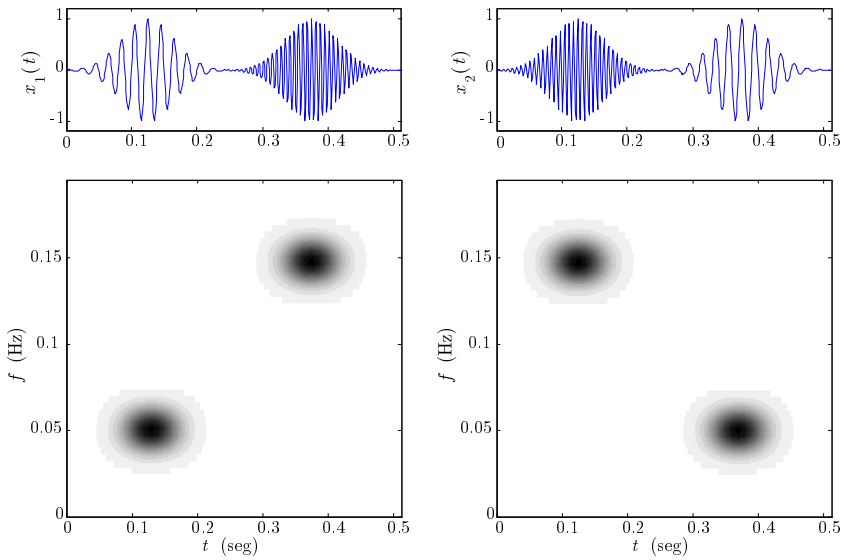

Figura 4.5: *Dos señales $x_1(t)$ y $x_2(t)$ formadas por combinaciones de dos tonos ventaneados de 10 Hz (A) y 30 Hz (B) (arriba) y sus respectivas magnitudes espectrales $|STFT_1(f,t)|$ y $|STFT_2(f,t)|$ calculadas mediante la STFT (abajo). A diferencia de lo que ocurría para el caso de la FT (Figura 4.2), en la representación obtenida a través del espectrograma es posible determinar la secuencia de aparición de ambos tonos.*

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:

$$\Delta f^2 \triangleq \frac{\int_{-\infty}^{\infty} f^2 \cdot |G(f)|^2 \, df}{\int_{-\infty}^{\infty} |G(f)|^2 df}. \tag{4.4}$$

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:

$$\Delta t^2 \triangleq \frac{\int_{-\infty}^{\infty} t^2 \cdot |g(t)|^2 \, dt}{\int_{-\infty}^{\infty} |g(t)|^2 dt}. \tag{4.5}$$

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* :

$$\Delta t \cdot \Delta f \geqslant \frac{1}{4\pi}. \tag{4.6}$$

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 (Ver Figura 4.6).

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 (4.6) (es decir la igualdad) se da cuando la ventana $g(t)$ es de tipo gaussiana. Para este caso (4.3) se denomina *transformada de Gabor*:

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

Otras posibles ventanas pueden ser las que se muestran en la Figura 4.7.

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

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} S_F(\tau, f) g(t-\tau) \, e^{j2\pi f\tau} df \, d\tau. \tag{4.7}$$

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 $g(t) \in L^2(\mathbb{R})$.

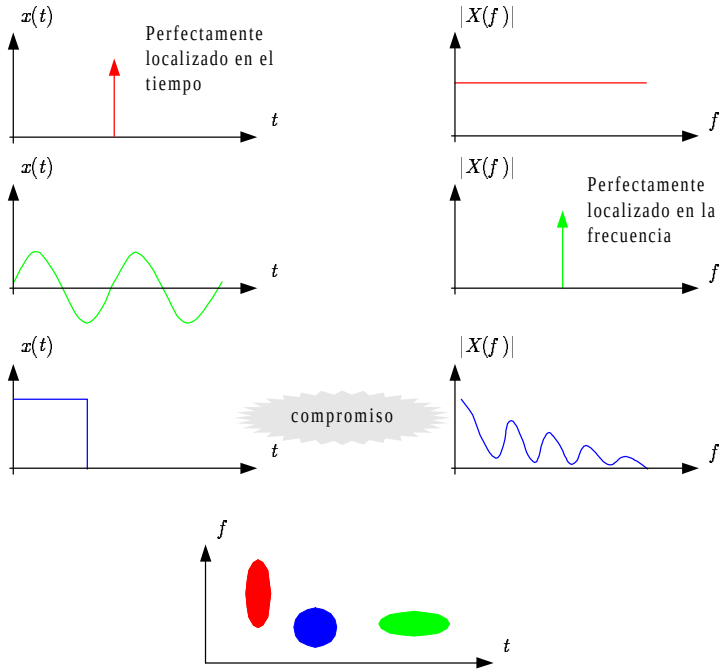

Figura 4.6: *Esquema ilustrativo del compromiso en la resolución tiempo-frecuencia "impuesto" por el principio de incertidumbre. De acuerdo con este principio no es posible obtener una resolución arbitrariamente buena en el tiempo y en la frecuencia simultáneamente.*

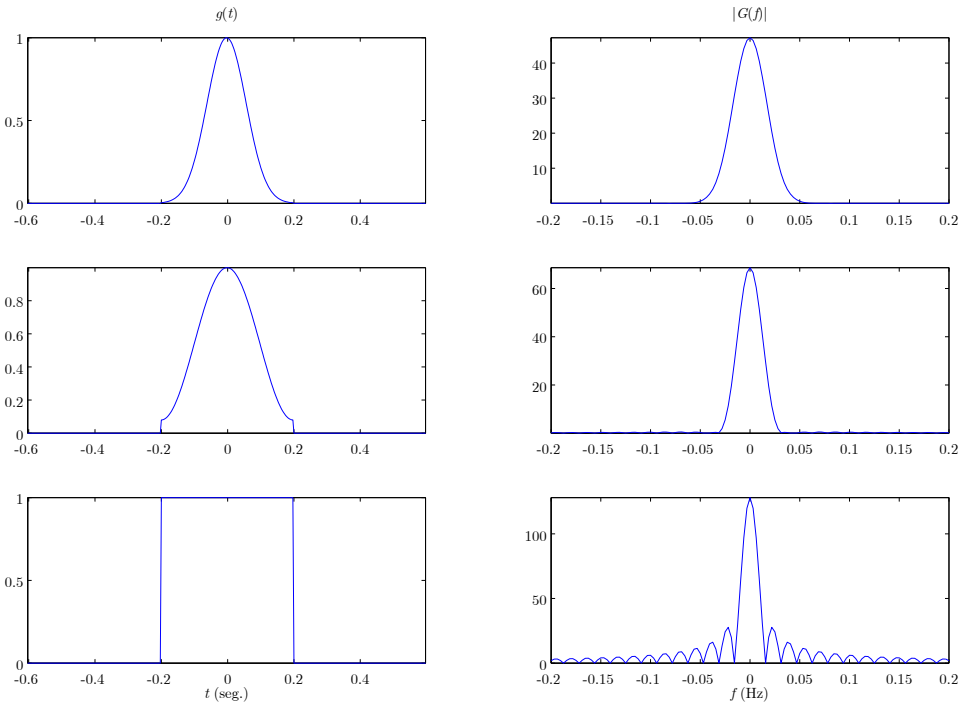

Figura 4.7: *Ventanas que pueden utilizarse en la STFT (izquierda) y sus respectivas magnitudes espectrales (derecha): Gabor (arriba), Hamming (centro) y cuadrada (abajo). Desde el punto de vista del principio de incertidumbre la ventana óptima sería la de Gabor (o gaussiana).*

4.3.3. 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 [130]. 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 [34], Wojtaszczyk [214], y Mallat [126, 125].

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 (4.6) es hacer que la resolución en el tiempo se incremente con la frecuencia central de los filtros de análisis. Más específicamente se impone que:

$$\frac{\Delta f}{f} = c, \tag{4.8}$$

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 4.6 *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:*

$$S_w(\tau, a) = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{|a|}} \psi^* \left(\frac{t-\tau}{a} \right) dt, \tag{4.9}$$

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:

$$\int_{-\infty}^{\infty} \psi(t)\, dt = 0. \tag{4.10}$$

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 4.8 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 (4.8) de la siguiente forma:

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

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 (4.9) y se hacen más evidentes si ahora se reescribe como:

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

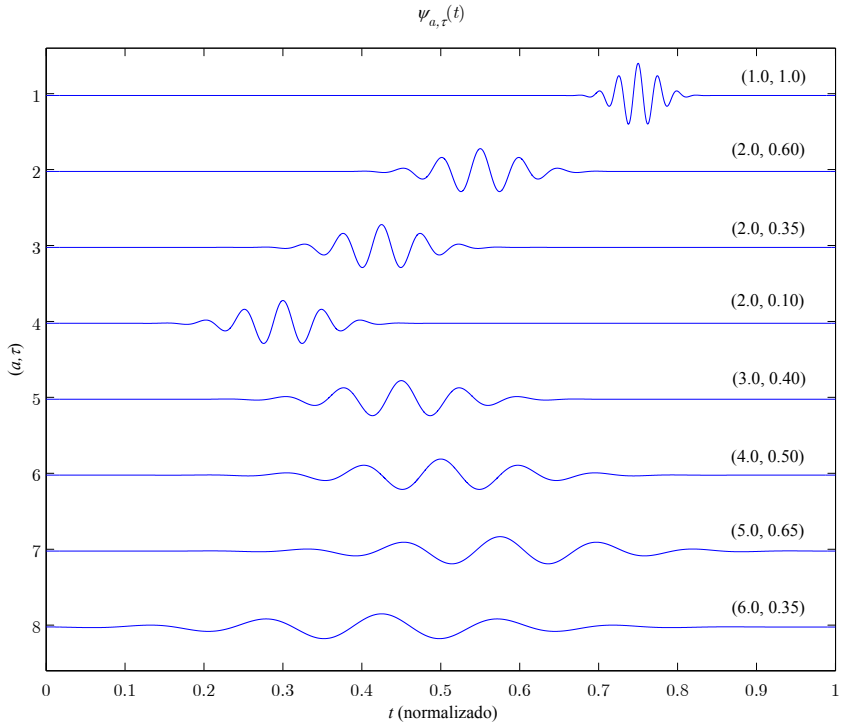

Figura 4.8: *Ejemplos de onditas de Morlet (parte real) a distintas escalas y localizaciones. Las gráficas se realizaron de acuerdo a los parámetros (a, τ), para una ondita $\psi_{a,\tau}(t)$ como en (4.11).*

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 $\psi(t)$:

$$\psi_{a,\tau}(t) = \frac{1}{\sqrt{|a|}}\,\psi\left(\frac{t-\tau}{a}\right). \tag{4.11}$$

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 [130]. 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:

$$x(t) = c\int_{-\infty}^{\infty}\int_{>0}^{\infty} S_W(\tau,a)\,\psi_{a,\tau}(t)\frac{da\,d\tau}{a^2}, \tag{4.12}$$

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.

4.4. 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).

4.4.1. 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 [127].

Definición 4.7 *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 [127]:*

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

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 la frecuencia. La WVD posee algunos inconvenientes, como la existencia de términos de interferencia y la no positividad (Ver Figura 4.9). En la Figura 4.10 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* [157]:

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

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 (Ver Figura 4.11) .

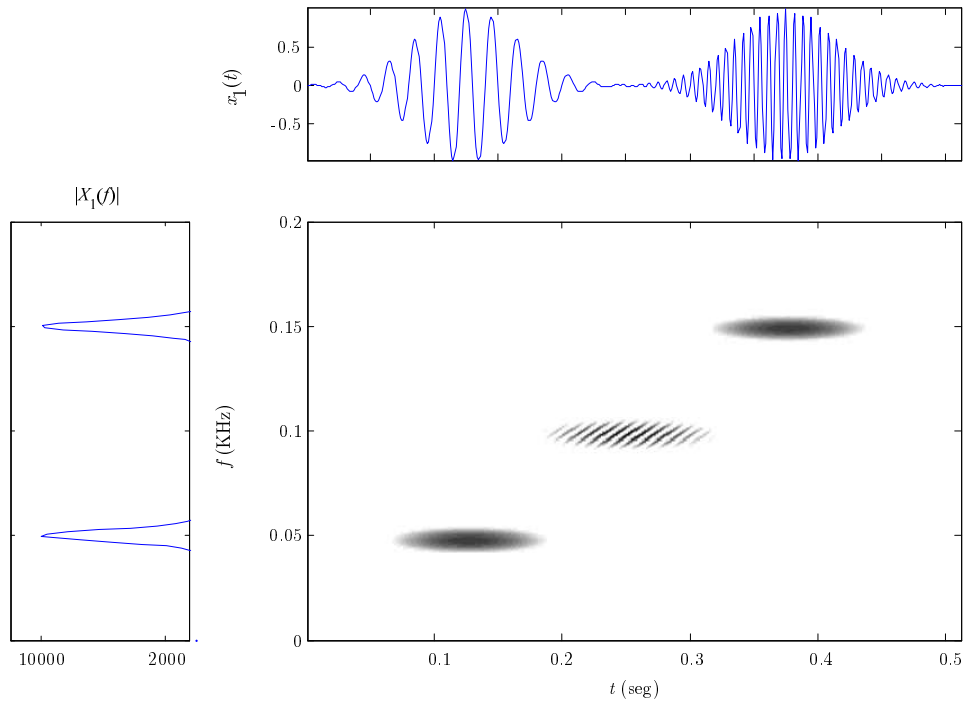

Figura 4.9: *Distribución de Wigner Ville (centro) de dos tonos (arriba) y su correspondiente espectro (izquierda). Es posible observar una mejora en la localización frecuencial con respecto al espectrograma pero a costa de la aparición de los términos cruzados (comparar con la Figura 4.5).*

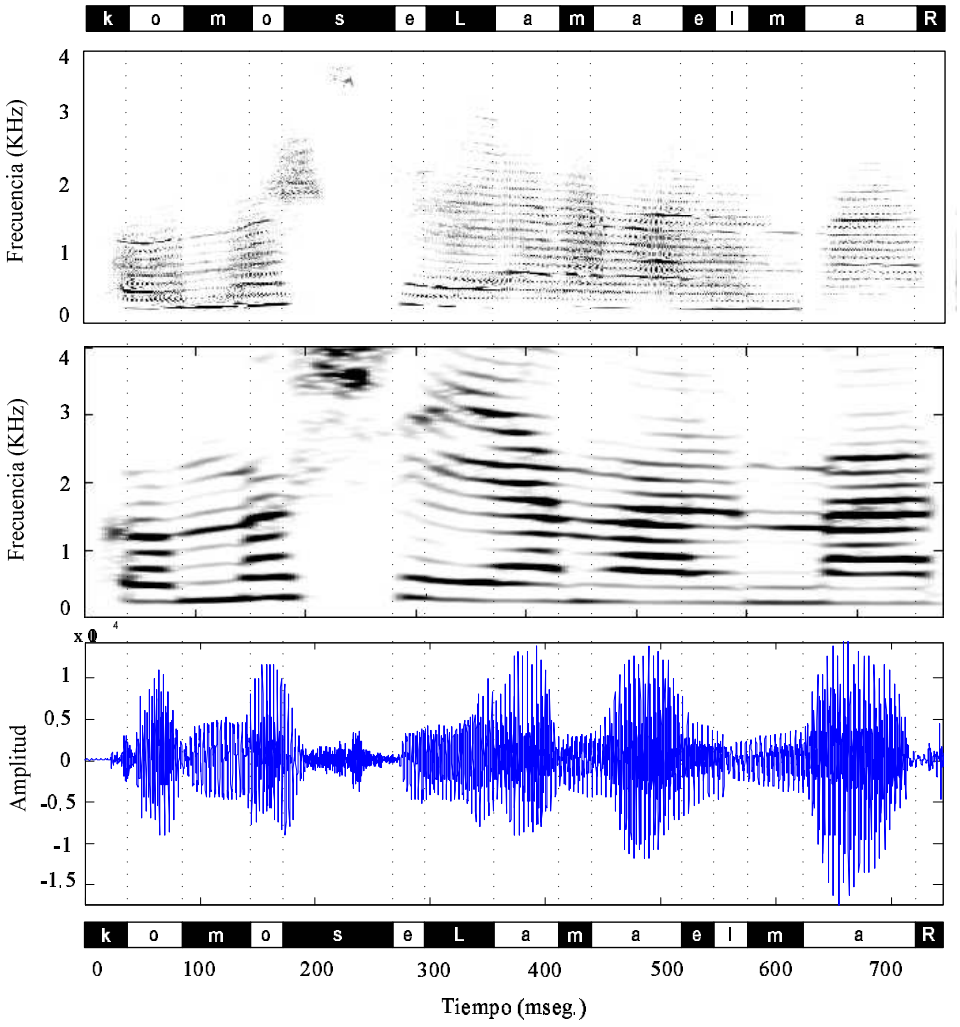

Figura 4.10: *Distribución de WV (arriba), comparada con el espectrograma de banda angosta (centro) de un trozo de voz (abajo). A pesar de que la localización frecuencial mejora notablemente con respecto al espectrograma (sin sacrificar resolución temporal), puede apreciarse también la aparición de términos cruzados que "oscurecen" el análisis mediante la introducción de elementos no presentes en la señal original.*

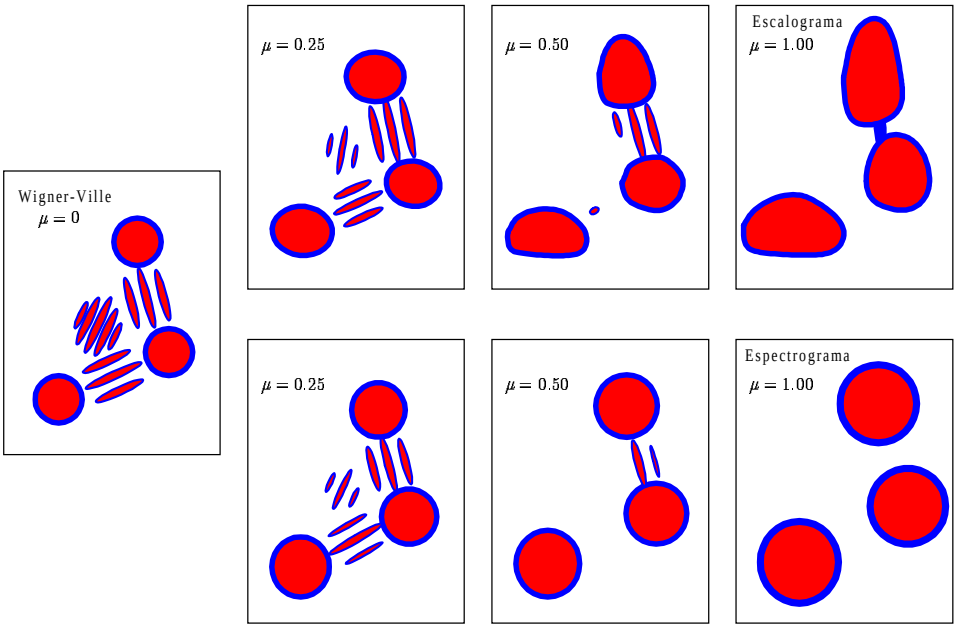

Figura 4.11: *Distribución de Wigner Ville (izquierda), escalograma (arriba), espectrograma (abajo) y casos intermedios obtenidos por convolución bidimensional con un núcleo adecuado. Es posible demostrar que todas las descomposiciones cuadráticas $t - f$ pueden escribirse de esta forma, lo que permite definir una familia general de distribuciones denominada clase de Cohen. El valor de μ en las gráficas permite regular la "dispersión" del núcleo de convolución.*

Distribución de Choi-William La distribución de Choi-Willian consiste en convolucionar la WVD con un núcleo exponencial bidimensional (cuasi cónico) [157]:

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

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 4.12).

Espectrograma

A partir de (4.3) es posible definir una densidad de energía que se denomina espectrograma:

$$P_F(\tau, f) = |S_F(\tau, f)|^2 = \left| \int_{-\infty}^{\infty} x(t) \cdot g^*(t - \tau) \cdot \ e^{-2j\pi f t} dt \right|^2 . \tag{4.13}$$

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 4.4 se generalizará el uso de este tipo de representaciones a través de la denominada distribución de Wigner-Ville.

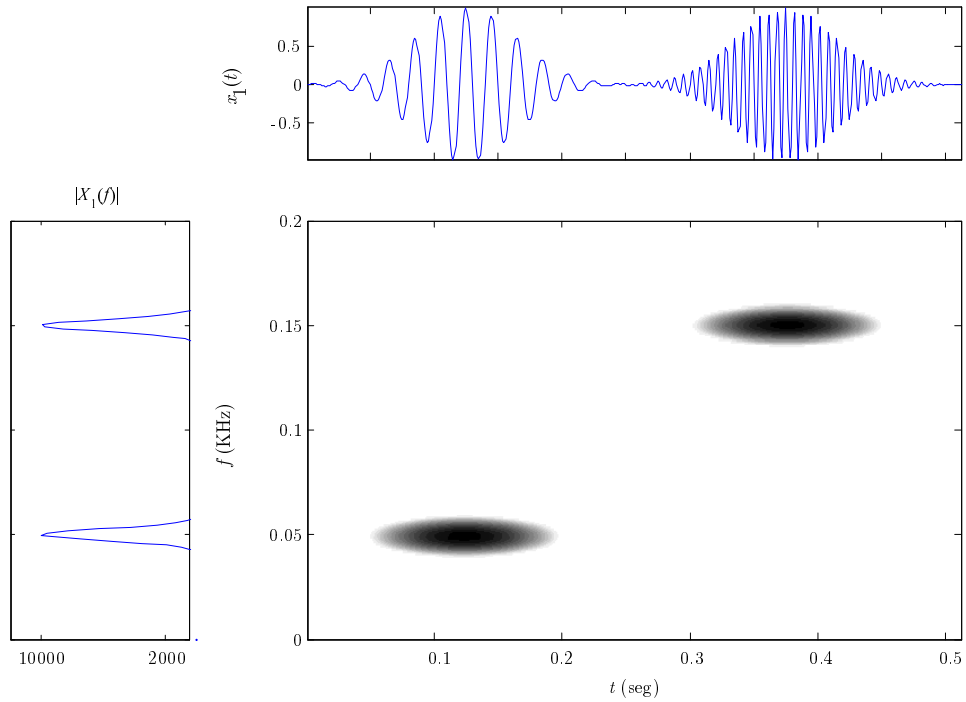

Figura 4.12: *Distribución de Choi-William de dos tonos donde pueden apreciarse la casi desaparición de los términos cruzados. Para esta distribución, que pertenece a la clase de Cohen, el núcleo empleado es exponencial.*

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 4.13). 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 $(f = \eta/a)$ [127]:

$$P_W(\tau, f) = |S_W(\tau, a)|^2 = |S_W(\tau, \eta/f)|^2 . \tag{4.14}$$

El escalograma "hereda" las mismas propiedades de la WT con respecto a la variación de la resolución en el plano $t-f$ (Ver Figura 4.14). En la Figura 4.15 se puede apreciar el escalograma de un trozo de señal de voz comparado con el correspondiente espectrograma de banda angosta.

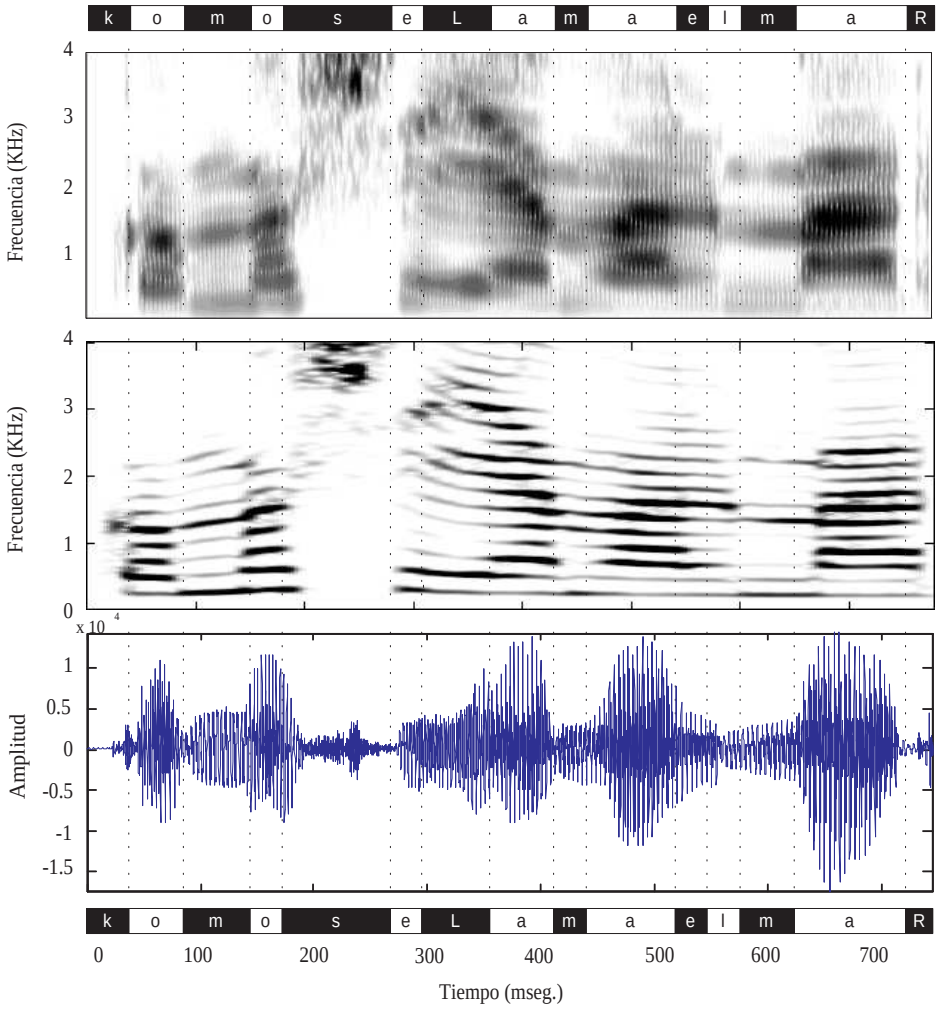

Figura 4.13: *Ejemplo de espectrograma de banda ancha (arriba), angosta (centro) y sonograma correspondiente (abajo). En los espectrogramas de banda angosta se logra una muy buena resolución en frecuencia, lo que resulta especialmente útil para la detección de formantes a partir de las "líneas" horizontales. 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.*

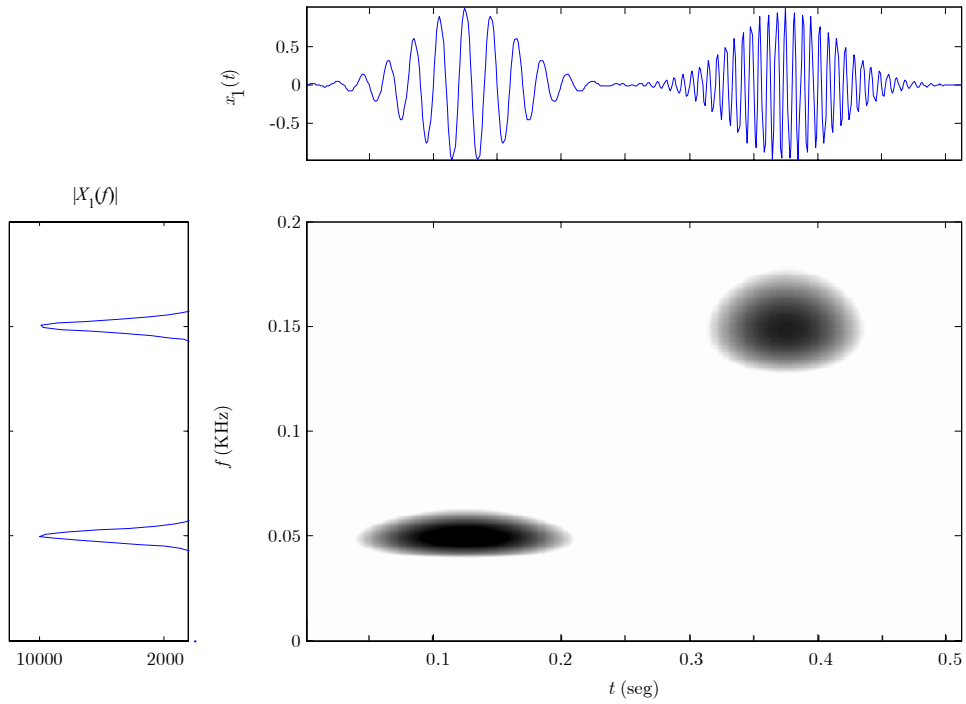

Figura 4.14: *Escalograma (centro, calculado con la ondita de Morlet) de una señal formada por dos tonos ventaneados (arriba, igual a la señal de la parte izquierda de la Figura 4.5) y su correspondiente espectro (izquierda). En el escalograma es posible observar fácilmente el cambio de resolución en las diferentes zonas del plano $t - f$.*

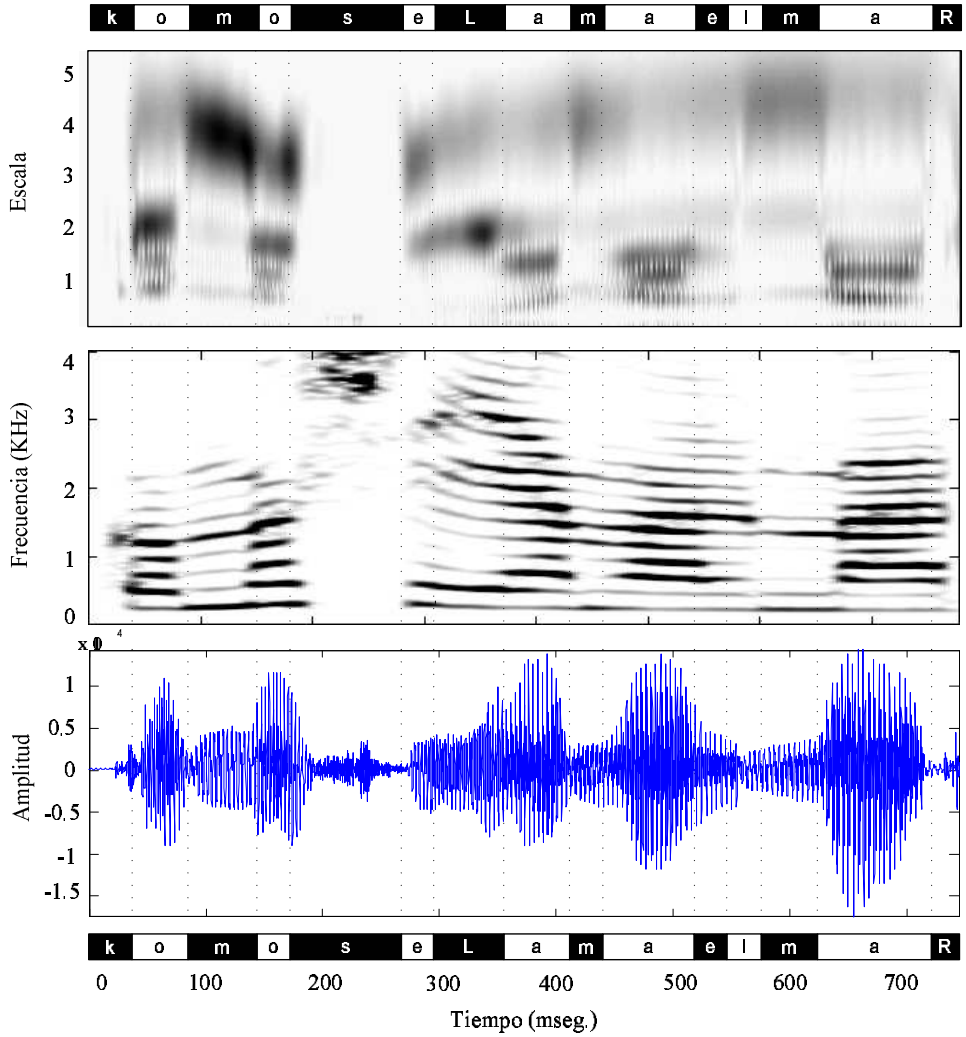

Figura 4.15: *Escalograma calculado utilizando la ondita de Morlet (arriba), comparado con el espectrograma de banda angosta (centro) de un trozo de voz (abajo). En el escalograma se puede observar un cambio en las estrías verticales de las escalas bajas (que tienen que ver con la F_0) hacia patrones 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 4.13).*

4.4.2. 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 [157]:

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

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.

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 [127]. Por supuesto que la aproximación puede ser completamente lineal si para ello se utiliza la Definición 2.3 de combinación lineal 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 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:

$$x(t) = \sum_{i \in \Gamma_{M_x}} c_i \phi_i(t). \tag{4.15}$$

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

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

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].

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* [145] que ya se ha introducido en la Sección 2.4.1. 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.

4.5. 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 3). 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.

4.5.1. Coeficientes de predicción lineal

Una de las técnicas paramétricas de análisis del habla más potentes es el método de *análisis predictivo lineal* (LPC) [159]. 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 4.16.

[7] A veces se denomina a este tipo de métodos como *análisis mediante diccionarios dependientes de los datos* o adaptativos.

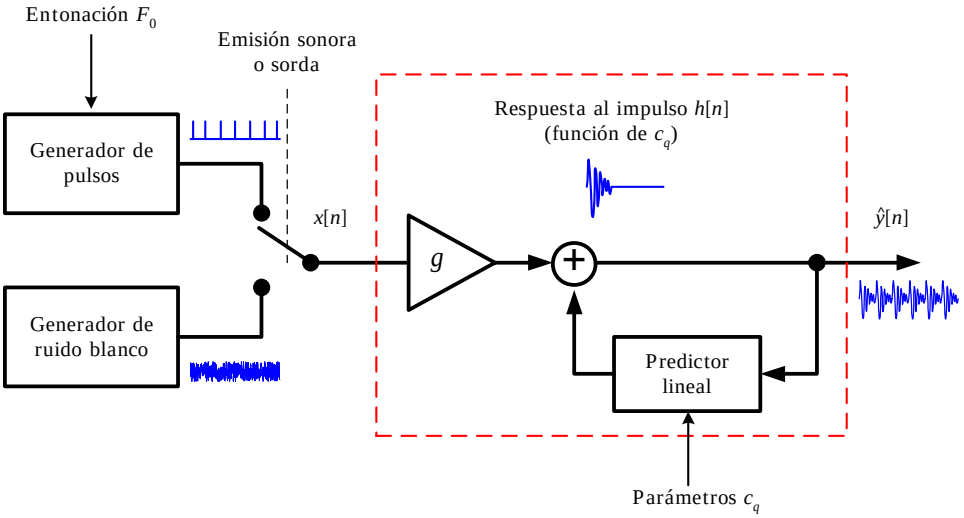

Figura 4.16: *Diagrama para el modelo AR del aparato fonador, donde la señal de voz $y[n]$ se aproxima mediante la salida $\hat{y}[n]$ de un sistema lineal de tiempo discreto frente a una excitación o entrada $x[n]$. Esta señal de excitación puede ser un tren de pulsos o ruido blanco dependiendo de si el fonema considerado es sonoro o sordo respectivamente.*

Definición 4.8 *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:*

$$\frac{\partial \; \mathcal{E}[(y[n] - \hat{y}[n])^2]}{\partial c_q} = 0. \tag{4.16}$$

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 4.16). 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 $v.a.$.

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 [37]. En la Figura 4.17 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.

4.5.2. 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 3.2.1). Ésto resulta similar al planteo anterior, salvo por el hecho de que ahora el modelo es de tiempo continuo:

$$y(t) = x(t) * h(t). \tag{4.17}$$

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 (4.17), entonces la ecuación en el dominio de la frecuencia es ahora un producto:

$$Y(f) = X(f)H(f). \tag{4.18}$$

Si luego se toman logaritmos en ambos miembros de (4.18), 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 4.9 *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.

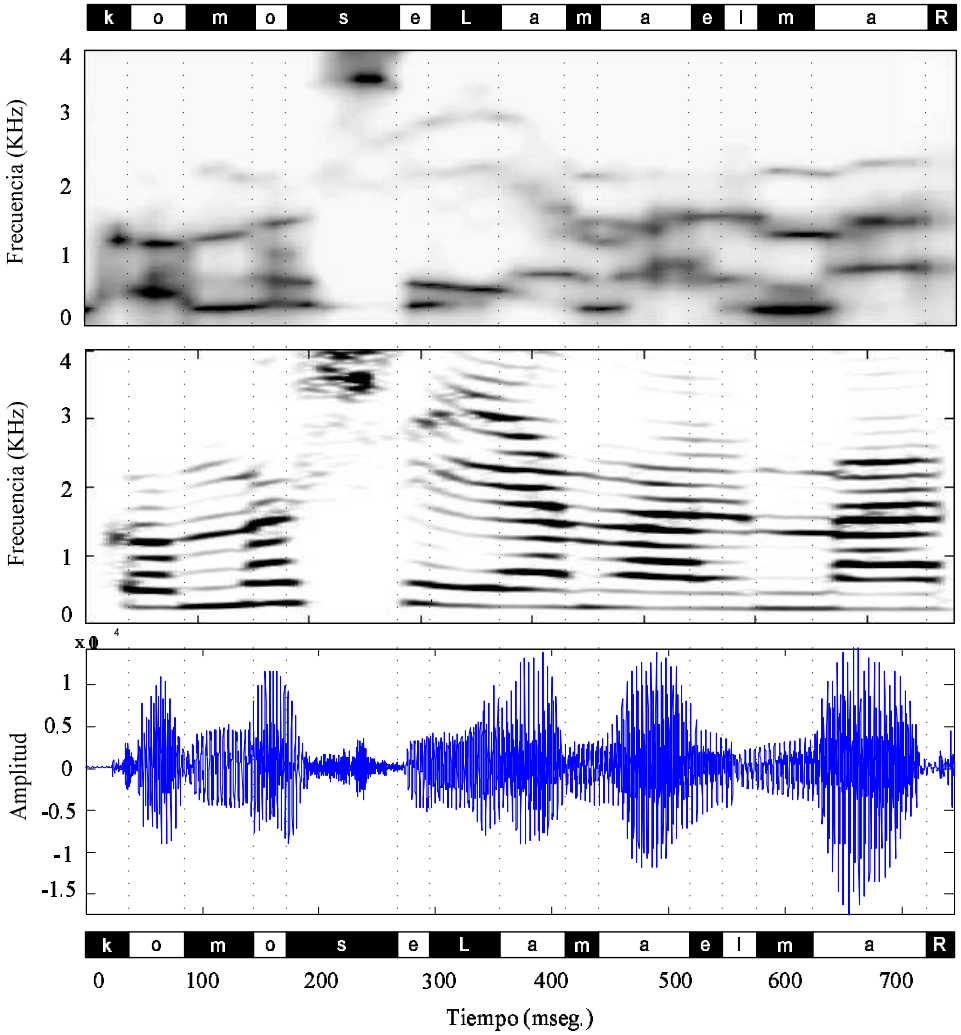

Figura 4.17: *Espectrograma "suavizado" estimado a partir de los coeficientes LPC (arriba), comparado con un espectrograma de banda angosta (centro) de un trozo de voz (abajo). Es posible observar como en este suavizado se pierden algunas de las componentes espectrales más finas pero se conservan rasgos importantes como las resonancias asociadas a las frecuencias formantes.*

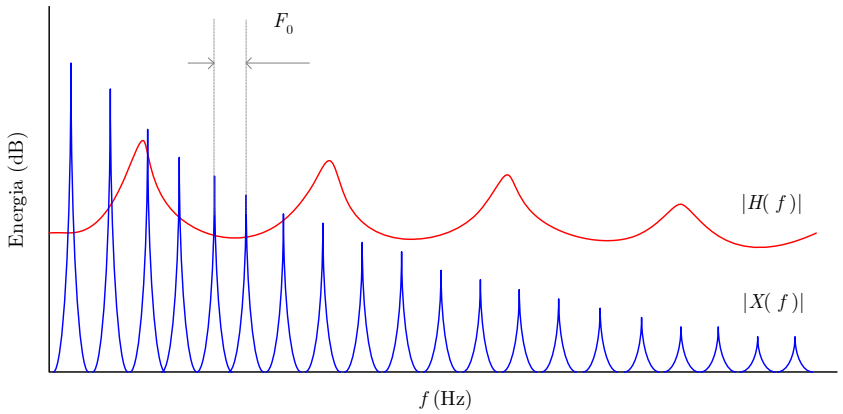

Figura 4.18: *Magnitud espectral de la excitación $X(f)$ y de la respuesta en frecuencia del tracto vocal $H(f)$ "simulado" para el caso de los fonemas sonoros. El espectro de la excitación $X(f)$ se ha representado mediante un tren de pulsos decrecientes, mientras que la respuesta en frecuencia del tracto vocal $H(f)$ mediante una función continua con varios picos correspondientes a las frecuencias formantes.*

El 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 4.18). En la Figura 4.19 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.

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 [37]. 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 3.4) 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,

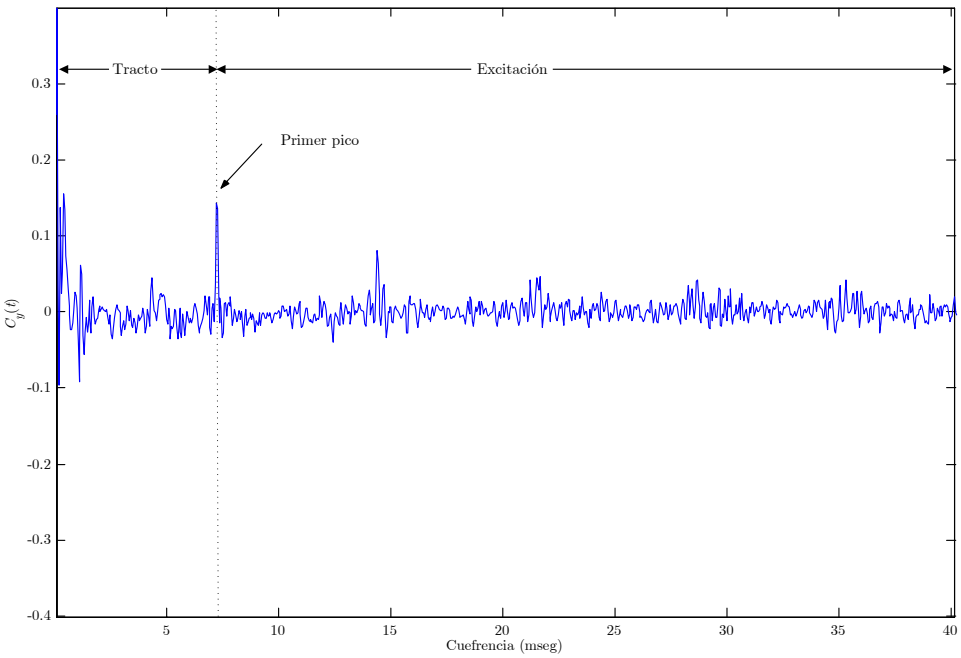

Figura 4.19: *Cepstrum real correspondiente a un trozo de una vocal /e/ sostenida de un hablante masculino. Se puede apreciar que la parte de bajas cuefrencias (antes del primer pico) corresponde a la componente de la respuesta del tracto vocal, mientras que las altas cuefrencias corresponden a la componente de la excitación.*

en la escala psicoacústica de mel. La relación entre la escala lineal en Hz y la escala de mel se muestra en la Figura 4.20. El mapeo es aproximadamente lineal por debajo de 1 KHz y logarítmico por encima, lo cual lleva a una aproximación comúnmente utilizada [37]:

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

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

4.5.3. Análisis predictivo lineal perceptual

El *análisis predictivo lineal perceptual* (PLP) fue introducido por Hermansky [74] 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*).

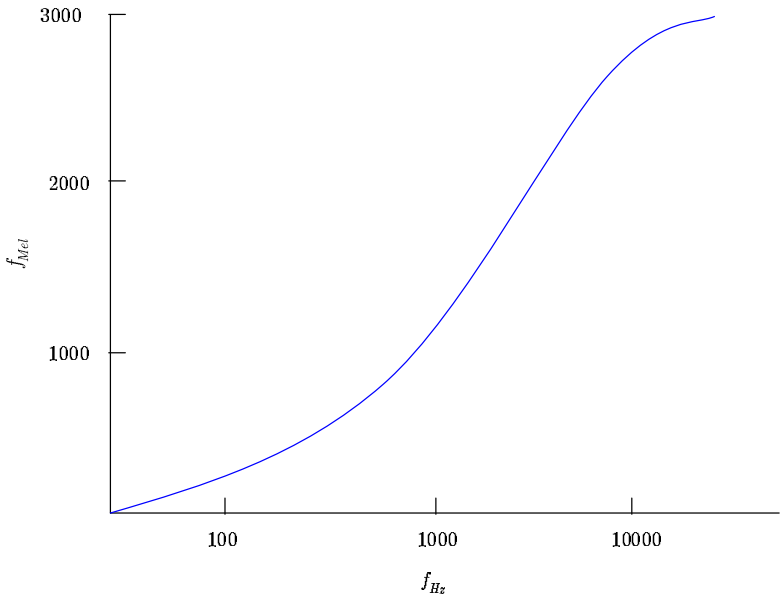

Figura 4.20: *Relación entre la escala frecuencial lineal en Hz y la escala frecuencial de mel. Esta escala esta dada por la relación entre la altura tonal percibida y la frecuencia "real" obtenida a partir de experimentos de proporcionalidad entre sensaciones.*

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.

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) [76]. En la Figura 4.21 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.

4.5.4. 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 [75] (véase también [197] y [156]).

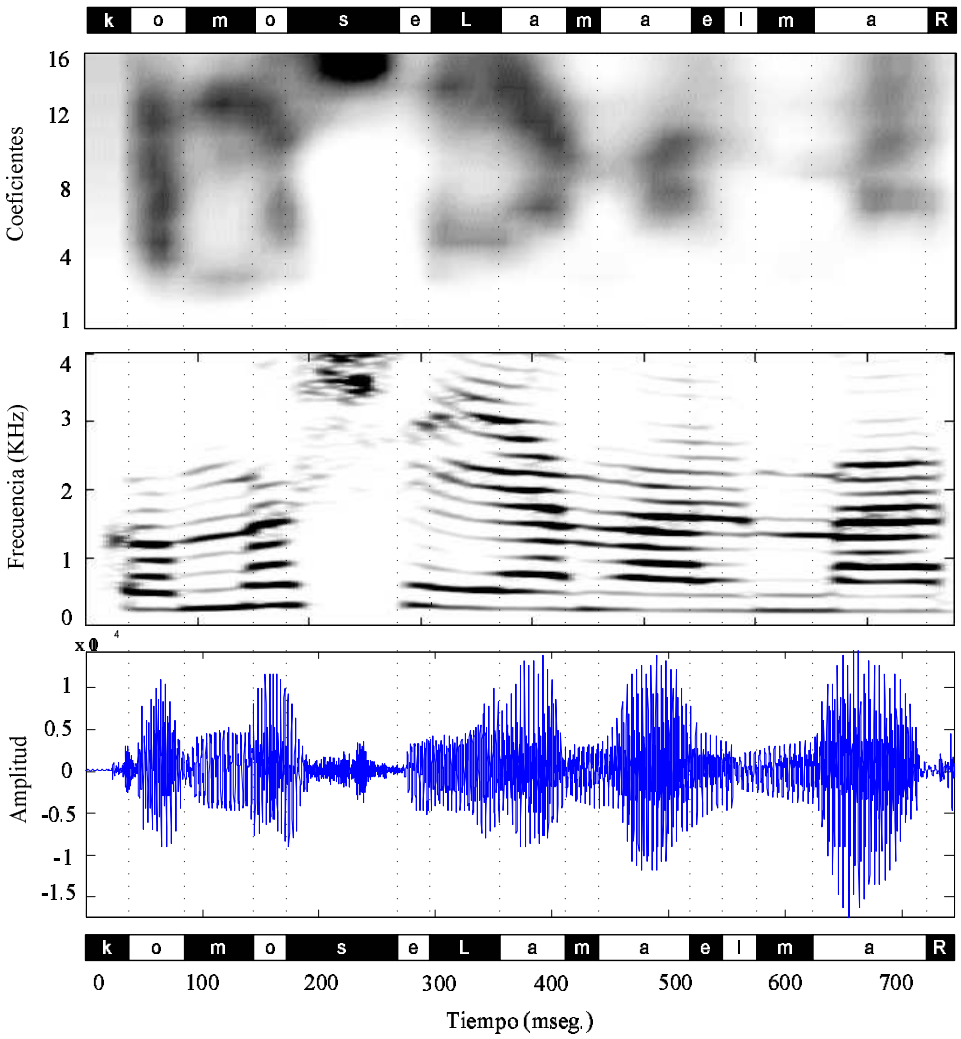

Figura 4.21: *Análisis RASTA-PLP (arriba), comparado con el espectrograma (centro) de un trozo de voz (abajo). Para facilitar la comparación se ha realizado una interpolación bidimensional del análisis RASTA-PLP que es de naturaleza discreta. Aquí se observa la pérdida de alguna información, que aparecía 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.*

Generalmente este enfoque requiere un mayor tiempo de cálculo, aunque se han reportado modelos bastante "exactos" que se han optimizado en este sentido [36]. 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.

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 3.22 ha sido realizado mediante un modelo auditivo [186].

4.6. 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) [37]. 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

[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 [117] y pruebas in vivo [35].

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 3). 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 [175].

4.7. 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 próximo capítulo 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 o el análisis estadístico de datos (ya revisadas en el Capítulo 2). Por ejemplo, a diferencia de la mayoría de las técnicas desarrolladas en el presente capítulo, la técnica de ICA surge desde una perspectiva principalmente estadística. 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 nos devuelve a la discusión presentada en el Capítulo 2. 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.

Capítulo 5

Representaciones basadas en diccionarios discretos

"Por esto les hablo en parábolas, porque viendo no ven y oyendo no oyen ni entienden; ..."

(Mateo 13,13)

Contenido

5.1. Introducción

HASTA aquí se han presentado los fundamentos generales y las técnicas disponibles para modelar las características relevantes de una señal y convertirlas en una representación adecuada de la misma. En este capítulo se revisarán con mayor detalle algunos aspectos relacionados con las representaciones "atómicas" de señales discretas basadas en diccionarios que constituyan marcos o bases ortogonales. La definiciones de átomo y diccionario se presentaron en el Capítulo 2. Aquí el interés principal reside en aquellas representaciones con propiedades interesantes para su aplicación a la señal de voz, como las basadas en onditas o en sus diferentes variantes. El caso de las representaciones ralas y/o independientes también resulta de interés y será tratado en detalle en el Capítulo 6).

En el capítulo anterior el enfoque ha estado fundamentalmente orientado al tiempo continuo, mostrándose resultados para tiempo discreto sólo cuando eran importantes para el problema considerado. Por otra parte la aplicación final de todos estos conceptos se realiza sobre secuencias discretas y finitas. La razón de utilizar el enfoque continuo en algunas secciones es que facilita la obtención de resultados teóricos. Puede tomarse el caso de las onditas donde ésto resulta más evidente. Como ya se ha visto, una base onditas en $L^2(\mathbb{R})$ se construye a partir de dilataciones y traslaciones de una única función madre. Pero la dilatación no está definida sobre secuencias discretas por lo que las bases de onditas discretas poseen una estructura bastante más complicada. De esta forma, una vez comprendidas las propiedades de las onditas como funciones continuas, es posible obtener resultados asintóticos para secuencias discretas cuando el intervalo de muestreo tiende a cero. Sin embargo esta transición debe realizarse con cuidado ya que, por ejemplo, el muestreo uniforme de una base de onditas de tiempo continuo no produce una base discreta ortonormal [127].

En este capítulo se continuarán empleando funciones continuas cuando faciliten los desarrollos, aunque el objetivo final consiste en concluir con las expresiones para secuencias discretas y finitas que constituyen el tipo de señales que se deberán manipular en las aplicaciones.

El capítulo está organizado de la siguiente manera. En la Sección 5.2 se retoma el caso del análisis clásico de Fourier pero ahora en un contexto discreto y finito. El resto del capítulo esta orientado a las representaciones discretas menos convencionales, basadas en onditas. En la Sección 5.3 se presenta la *transformada ondita discreta diádica* y en la Sección 5.4 se presenta la *transformada de paquetes de ondita discreta* como una generalización de estas ideas. Un desarrollo similar pero más relacionado con la familia de bases de Fourier es el de los *paquetes de cosenos* que se presenta en la Sección 5.5.

5.2. Transformada discreta de Fourier

Se comenzará retomando el análisis de Fourier, según se discutió en la Sección 4.2.1, pero ahora en un contexto discreto. Este tipo de análisis permite un abordaje discreto bastante más directo que para el caso de onditas. El principal cuidado que debe tenerse para mantenerse suficientemente "cerca" de las propiedades de las señales de tiempo continuo, consiste en cumplir con las condiciones del *Teorema del muestreo*. Ésto significa muestrear a la señal al menos al doble de la frecuencia de Nyquist.

Definición 5.1 *Se define la* transformada discreta de Fourier *(DFT) de la señal* $x[n] \in \mathbb{R}^N$ *como:*

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j\frac{2\pi kn}{N}}, \tag{5.1}$$

con $0 \leq k \leq N$.

La correspondiente transformada inversa que devuelve a $x[n]$ *es ahora:*

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k]\, e^{j\frac{2\pi kn}{N}}, \tag{5.2}$$

con $0 \leq n \leq N$.

Obsérvese que es posible reescribir la ecuación de síntesis (5.2) como una sencilla multiplicación matricial de la forma:

$$\mathbf{x} = \mathbf{\Phi}\mathbf{X},$$

donde $\mathbf{\Phi} = \{\phi_{n,k}\}$ y sus columnas constituyen los átomos de la base de Fourier discreta, por lo que sus muestras están dadas por[1]:

$$\phi_{n,k} = \frac{1}{N}\, e^{j\frac{2\pi kn}{N}}.$$

En las Figuras 5.1 y 5.2 es posible apreciar la "estructura" de la matriz $\mathbf{\Phi}$ junto con algunas columnas o átomos de la misma para $N = 256$.

La matriz $\mathbf{\Phi}$ tiene la propiedad de que su inversa es igual al conjugado de su traspuesta. Ésto convierte a la DFT en una *transformación unitaria* y permite reescribir la ecuación de análisis (5.1) como:

$$\mathbf{X} = \mathbf{\Phi}^{*\mathrm{T}}\mathbf{x}.$$

En la aplicaciones esta propiedad permite ahorrar bastante tiempo de cálculo, debido a que se está reemplazando la inversión de la matriz por una trasposición.

Además, dado que las columnas de $\mathbf{\Phi}$ constituyen funciones exponenciales periódicas muestreadas, es posible escribir:

$$e^{j\frac{2\pi}{N}kn} = e^{j\frac{2\pi}{N}(nk \mod N)}. \tag{5.3}$$

Esta propiedad, junto con algunos "trucos" de factorización en submatrices adecuadas, permite reducir el orden de la cantidad de operaciones necesarias para realizar la multiplicación matricial de $\mathcal{O}(N^2)$ a $\mathcal{O}(N \log_2(N))$. Esto constituye el fundamento para la implementación del algoritmo de cálculo rápido de la DFT, denominado *transformada rápida de Fourier* (FFT). La DFT hereda las propiedades de un análisis estacionario como el de la FT. Sin embargo ahora la estacionariedad queda acotada a secuencias discretas finitas de longitud N. En el caso en que las propiedades de esta señal varíen sustancialmente dentro de estas N muestras es necesario considerar la versión de corta duración de la DFT.

5.2.1. Transformada discreta de Fourier de corta duración

En la Sección 4.3.2 se definió la transformada de Fourier de corta duración. Para el caso discreto tenemos ahora las siguientes expresiones "equivalentes".

Definición 5.2 *Sea* $x[n] \in \mathbb{R}^N$ *una señal discreta y* $g[n] \in \mathbb{R}^N$ *una ventana simétrica con norma unitaria y soporte compacto* $P \leq N$. *Se define entonces a la* transformada

[1] Salvo aviso en contrario los vectores utilizados son vectores columna

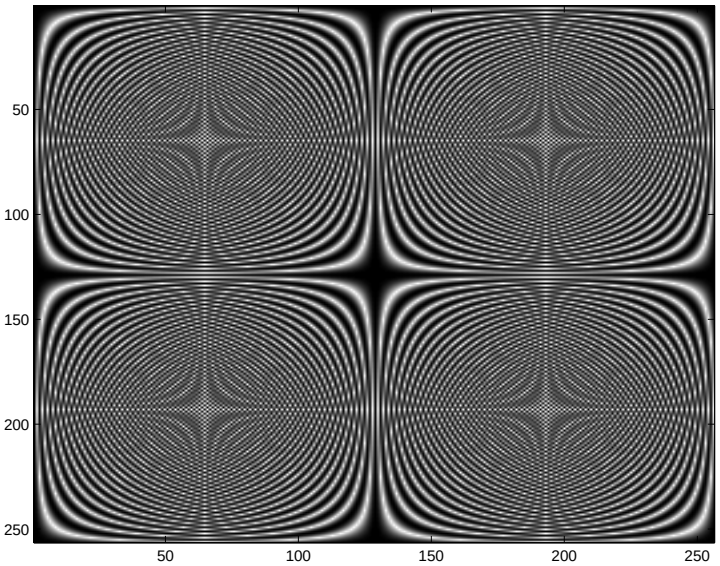

Figura 5.1: *Representación de la estructura interna del diccionario* $\mathbf{\Phi}$ *de la DFT para* $N = 256$ *(parte real). El eje horizontal corresponde al índice de* k *de las columnas que está relacionado con la frecuencia de las exponenciales complejas en (5.3), mientras que el eje vertical corresponde al índice temporal* n*. Obsérvese como el cambio en la periodicidad de las exponenciales, junto con las restricciones en cuanto a que éstas deben poseer un número entero de períodos son las responsables de los patrones característicos que forman parte de esta estructura.*

discreta de Fourier de corta duración *(STDFT) de la señal* $x[n]$ *como:*

$$S_F[m,k] = \langle x[n], g_{m,k}[n] \rangle = \sum_{n=0}^{N-1} x[n] g[n-m]\, e^{\frac{-j2\pi kn}{N}}, \tag{5.4}$$

$$g_{m,k}[n] = g[n-m]\, e^{\frac{-j2\pi kn}{N}}. \tag{5.5}$$

donde $0 \leq m < N$, $0 \leq k < N$.

La correspondiente transformación inversa es:

$$x[n] = \frac{1}{N} \sum_{m=0}^{N-1} \sum_{k=0}^{N-1} S_F[m,k] g[n-m]\, e^{\frac{j2\pi kn}{N}}. \tag{5.6}$$

El caso de la reconstrucción discreta (5.6) es otra vez muy redundante. Sin embargo es posible construir un marco o inclusive una base ortogonal de acuerdo con los parámetros de la discretización (haciendo por ejemplo que el paso de la ventana de análisis sea mayor que 1) [34].

De forma análoga al caso anterior es posible escribir la ecuación de síntesis (5.6) en forma matricial. Para ello se definen $\mathbf{\Phi} \in \mathbb{R}^{N \times N^2}$, $\mathbf{F}^l \in \mathbb{R}^{P \times N}$ y $\mathbf{S}'_F \in \mathbb{R}^{N^2}$ de la siguiente

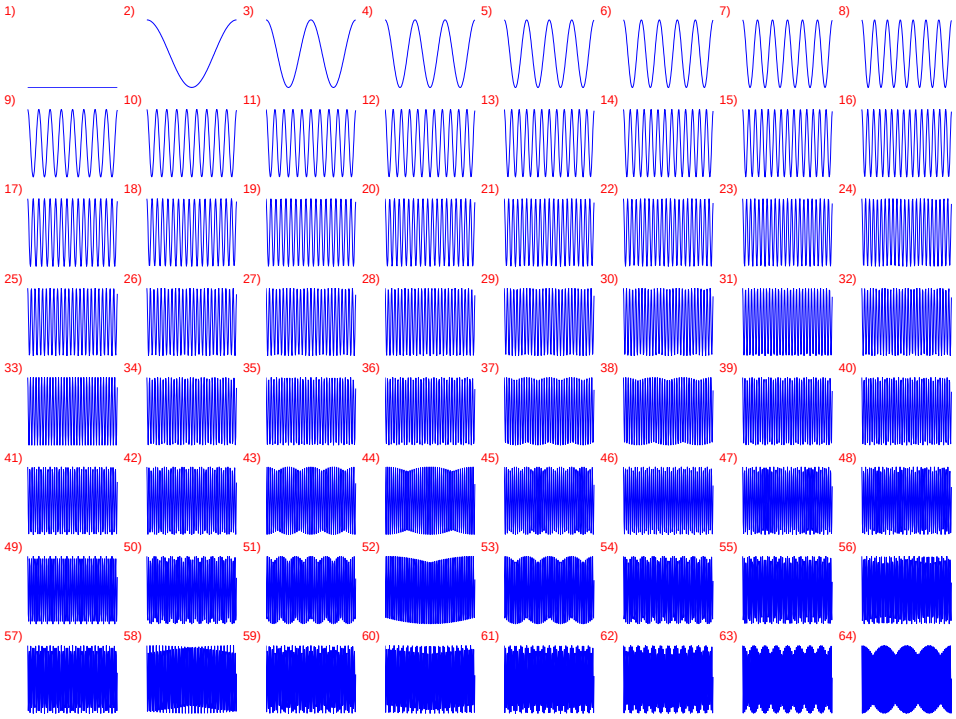

Figura 5.2: *Algunos átomos o columnas del diccionario* $\mathbf{\Phi}$ *de la DFT para* $N = 256$ *(parte real). El valor indicado en la parte superior izquierda de cada átomos corresponde al índice* k *de cada columna. Se puede apreciar más claramente el aumento de la frecuencia de las exponenciales con este índice.*

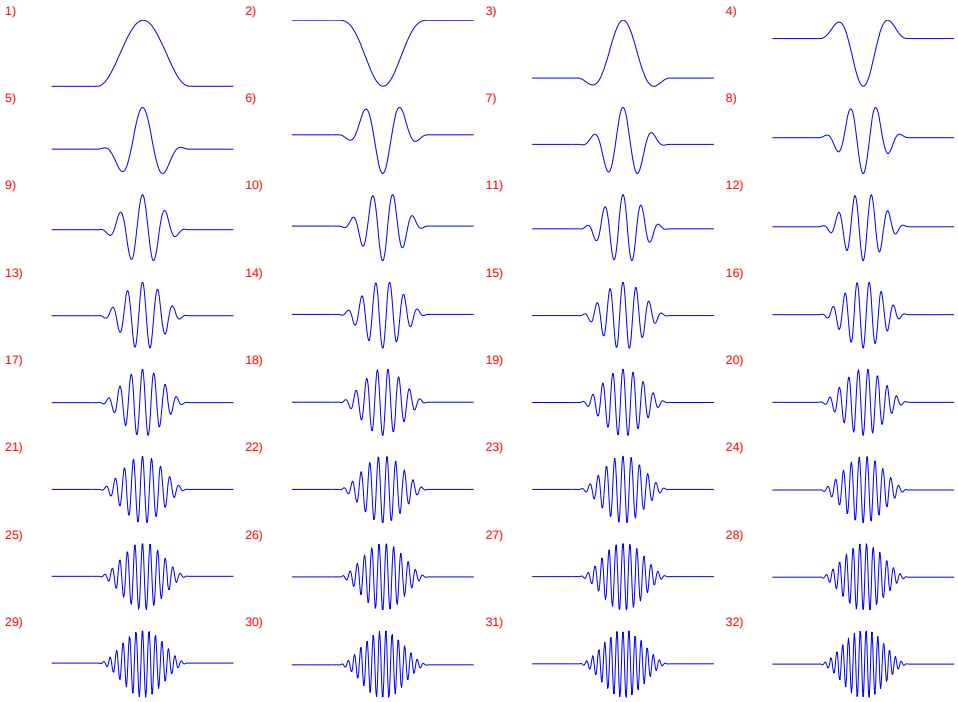

Figura 5.3: *Algunos átomos del diccionario* $\mathbf{\Phi}$ *de la STDFT para* $N = 256$ *(parte real) y* $P = 128$ *con una ventana de Hanning, que corresponden a las columnas (completas) que contienen a una de las submatrices* $\mathbf{F}^l$.

forma:

$$\mathbf{\Phi} = \begin{bmatrix} \ddots & \boxed{\begin{matrix} \mathbf{F}_1^{l-1} \\ \vdots \\ \mathbf{F}_P^{l-1} \end{matrix}} & \boxed{\begin{matrix} \mathbf{F}_1^{l} \\ \vdots \\ \mathbf{F}_P^{l} \end{matrix}} & \boxed{\begin{matrix} \mathbf{F}_1^{l+1} \\ \vdots \\ \mathbf{F}_P^{l+1} \end{matrix}} & \ddots \\ \ddots & & & & \ddots \\ \ddots & & & & \ddots \end{bmatrix}, \mathbf{F}^l = \frac{1}{N} \begin{bmatrix} \vdots \\ g[p]\ e^{j\frac{2\pi k(P-p)+(l \div N)}{N}} \\ \vdots \end{bmatrix},$$

$$S'_F[mN + k] = S_F[m, k],$$

donde $0 \leq l < N$. De allí la ecuación matricial queda escrita finalmente como:

$$\mathbf{x} = \mathbf{\Phi}\mathbf{S}'_F.$$

La forma de los átomos de este diccionario puede apreciarse en la Figura 5.3. La matriz $\mathbf{\Phi}$ posee una estructura particular donde las submatrices $\mathbf{F}$ resultan ser casi iguales (salvo por la fase de sus columnas). Esta estructura se evidencia claramente en la Figura 5.4. Es posible también plantear un ordenamiento alternativo para las columnas de $\mathbf{\Phi}$ (con el correspondiente cambio de orden en los coeficientes $\mathbf{S}'_F$). Para ello puede mantenerse ahora la frecuencia fija por bloques (de una columna a la siguiente), cambiando sólo el

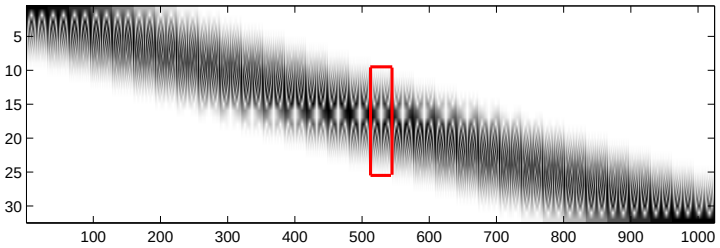

Figura 5.4: *Estructura del diccionario* $\mathbf{\Phi}$ *de la STDFT para* $N = 32$ *y* $P = 16$ *con una ventana de Hanning (parte real). El eje horizontal corresponde al índice de las columnas mientras que el eje vertical corresponde al índice de las filas. El diccionario* $\mathbf{\Phi}$ *posee una estructura particular donde las submatrices* $\mathbf{F}$ *resultan ser casi iguales (salvo por la fase de sus columnas). En la figura se ha remarcado la zona correspondiente a una de las submatrices* $\mathbf{F}^l$.

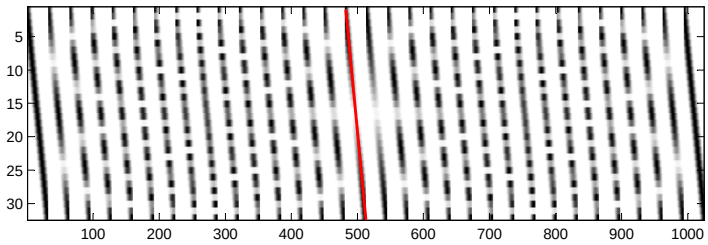

Figura 5.5: *Estructura del diccionario* $\mathbf{\Phi}$ *de la STDFT para* $N = 32$ *y* $P = 16$ *con una ventana de Hanning (parte real) y las columnas ordenadas de forma alternativa a las de la Figura 5.4. El eje horizontal corresponde al índice de las columnas mientras que el eje vertical corresponde al índice de las filas. Al compararla con la Figura 5.4 es posible ver los cambios en la estructura debidos al reordenamiento de las columnas. Se ha remarcado la zona correspondiente a una misma frecuencia desplazada en el tiempo.*

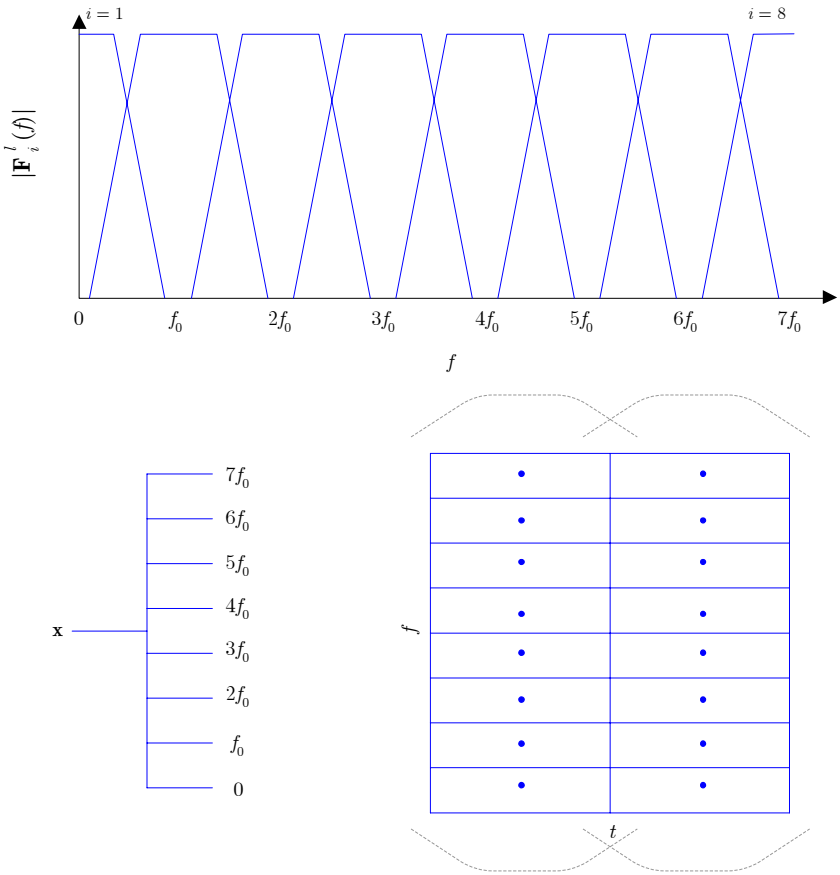

Figura 5.6: *Tipo de banco de filtros utilizado por la STDFT (arriba), representación esquemática del mismo (abajo izquierda) y partición $t - f$ asociada (abajo derecha).*

desplazamiento de la ventana dentro del bloque. En la Figura 5.5 se muestra la estructura de $\mathbf{\Phi}$ con este nuevo ordenamiento. Se debe hacer notar que en este caso los átomos no cambian, sólo lo hace el orden de las columnas dentro de la matriz.

En la práctica es posible realizar la multiplicación matricial en forma "rápida" mediante N FFTs de tamaño N, requiriendo por lo tanto un total de $\mathcal{O}(N^2 \log_2(N))$ operaciones. En forma análoga al caso continuo, es posible interpretar a la STDFT en términos de un banco de filtros como el mostrado en la Figura 5.6.

5.3. Transformada ondita discreta

En la Sección 4.3.3 se ha visto que el conjunto de onditas $\{\psi_{a,\tau}(t)\}_{a,\tau\in\mathbb{R}}$ se comporta en forma similar a una base ortogonal, tanto para análisis como para síntesis. Tal como en el caso de la STDFT es posible también construir un marco o inclusive una base verdaderamente ortogonal discretizando a y τ adecuadamente (por ejemplo con $a = {a_0}^k$ y $\tau = nb_0{a_0}^k$). Ésto constituye el caso de la *transformada ondita discreta* (DWT) y su existencia dependerá fundamentalmente de como se elija la función $\psi(t)$. En particular resulta de interés la discretización de a y τ en forma diádica, o *transformada ondita dis-*

creta diádica (DDWT), debido su implementación computacional sencilla. Existen otras discretizaciones posibles que dan lugar a la denominada *transformada ondita continua muestreada* (SCWT) y a la *transformada ondita semicontinua* (SWT) que no se analizaran en este trabajo, pero que pueden resultar de interés en algunas aplicaciones[2]. Con todas estas ideas presentes se procederá a definir los fundamentos teóricos del denominado *análisis multiresolución*, que se basa en una descomposición en onditas de tipo diádica. A continuación se definen estas onditas fijando la discretización tiempo-escala utilizada de aquí en adelante.

Definición 5.3 *Una ondita diádica es una función $\psi(t) \in L^2(\mathbb{R})$ que tiene la propiedad de que la familia de funciones $\psi_{k,n}(t) = 2^{k/2}\psi(2^k t - n)$ para $k, n \in \mathbb{Z}$, es una base ortonormal del espacio de Hilbert $L^2(\mathbb{R})$.*

La definición habla sobre la existencia de una *sola* función cuyas traslaciones y dilataciones apropiadas forman una base ortonormal. En la Figura 5.7 se observa como ejemplo el caso de la familia de onditas diádicas Symlets a distintas escalas y localizaciones. Es posible ver esta base ortonormal como un banco de filtros en octavas. En la Figura 5.8 se muestran la respuesta en frecuencia de estos filtros, comparada con la de la STDFT. Ésto genera una partición tiempo-frecuencia determinada que puede también apreciarse en la misma figura.

Aunque no resulta obvio, generalmente las "buenas" onditas (por tener propiedades adicionales como regularidad) se construyen a través de un *análisis multiresolución* (MRA), definido de la siguiente manera:

Definición 5.4 *Un análisis multiresolución es una secuencia $(V_k)_{k\in\mathbb{Z}}$ de subespacios de $L^2(\mathbb{R})$ tales que :*

1. *$V_k \subset V_{k+1}$ para cualquier $k \in \mathbb{Z}$,*
2. *$\overline{\bigcup_{k\in\mathbb{Z}} V_k}$ es denso en $L^2(\mathbb{R})$ y $\bigcap_{k\in\mathbb{Z}} V_k = \{0\}$,*
3. *$x(t) \in V_0 \Leftrightarrow x(2^{-k}t) \in V_k$ para cualquier $k \in \mathbb{Z}$,*
4. *$\exists$ una función $\phi(t) \in V_0$, llamada la función de escala, tal que la familia $\{\phi(t - n)\}_{n\in\mathbb{Z}}$ es una base ortonormal para V_0.*

Por la Definición 5.4 es posible aproximar cualquier función o señal en $L^2(\mathbb{R})$ por una función en alguno de los V_k. Se dice que ésto constituye una "aproximación" a la resolución o escala k. Como $\{2^{k/2}\phi(2^k t - n)\}_{n\in\mathbb{Z}}$ es una base ortonormal para V_k, ésto da la posibilidad de analizar una señal con "ventanas" de diferentes tamaños, a diferencia del análisis basado en la FT. Las condiciones sobre los subespacios V_k implican que por estar en el espacio de Hilbert $L^2(\mathbb{R})$ existen subespacios $(W_k)_{n\in\mathbb{Z}}$ tal que cada V_{k+1} es la suma directa de V_k con W_k. Las W_k tienen la interpretación de que representan la información

[2]Por ejemplo para el caso de necesitarse una transformación invariante al desplazamiento o que provea información $t - f$ de manera arbitraria

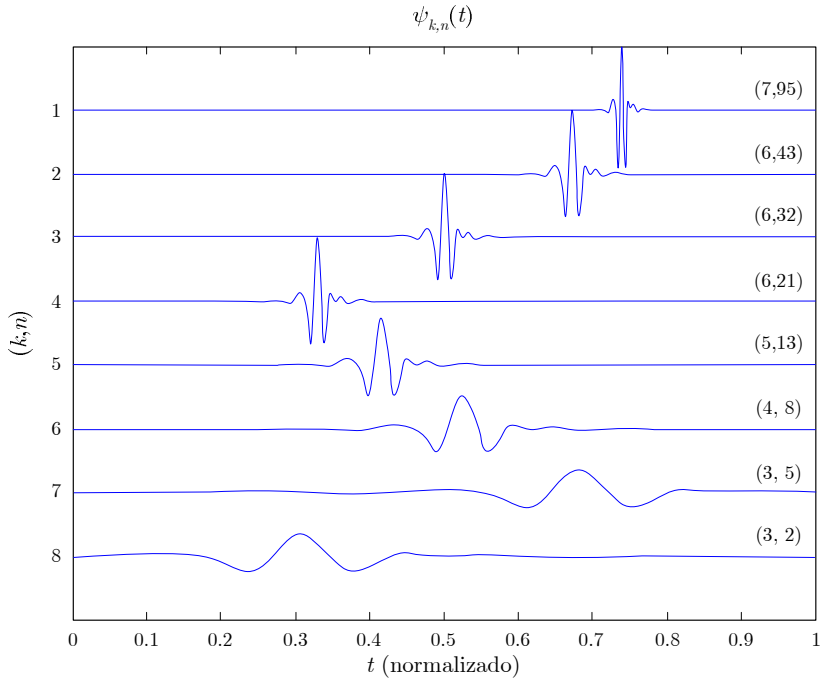

Figura 5.7: *Ejemplos de onditas Symlets diádicas a distintas escalas y localizaciones (con 6 momentos nulos). Las gráficas se realizaron de acuerdo a los parámetros* (k, n)*, para una ondita de ancho aproximado* 2^{-k} *y localización en* $n/2^k$ *en el intervalo unitario.*

de detalle que se requiere cuando se pasa de una aproximación a la resolución k, a una aproximación a la resolución $k+1$. En este caso $V_{k+1} = V_k \oplus W_k$, y continuando el proceso obtenemos $V_{k+1} = V_l \oplus W_l \oplus W_{l+1} \oplus \ldots\ldots \oplus W_{k-1} \oplus W_k$. Entonces una aproximación a una resolución más alta puede ser representada en términos de una resolución más baja con los detalles adicionales. Además, $L^2(\mathbb{R})$ es la suma directa de los W_k.

Por las condiciones de la Definición 5.4 es posible realizar el siguiente razonamiento ([34], capitulo 5). Si $\phi(t)$ es la función de escala, entonces $\phi(t) \in V_0$ y $\phi(t/2) \in V_1$. Dado que $V_0 \subset V_1$, ésto significa que $\phi(t)$ se encuentra también en V_1. Por lo tanto $\phi(t)$ debe poder expresarse en términos de la base $\phi(2t - n)$ para el espacio V_1:

$$\phi(t) = 2 \sum_{n=-\infty}^{+\infty} h[n]\phi(2t - n), \tag{5.7}$$

donde $h[n] = 1/2 \langle \phi(t), \phi_{1,n}(t) \rangle = \int \phi(t)\phi^*(2t - n)dt$ y $\sum_n |h[n]|^2 = 1/2$. A (5.7) se la llama *ecuación de escala*.

Tomando la TF de esta ecuación se obtiene:

$$\Phi(f) = H(f/2)\Phi(f/2), \tag{5.8}$$

donde $H(f) = 2\sum_n h[n]\ e^{-j2\pi fn}$, que es una función de período unitario. Resulta que H es un filtro pasabajos que además determina a la función de escala (si se aplica (5.8) de forma recursiva, ver Sección 6.4 [194]).

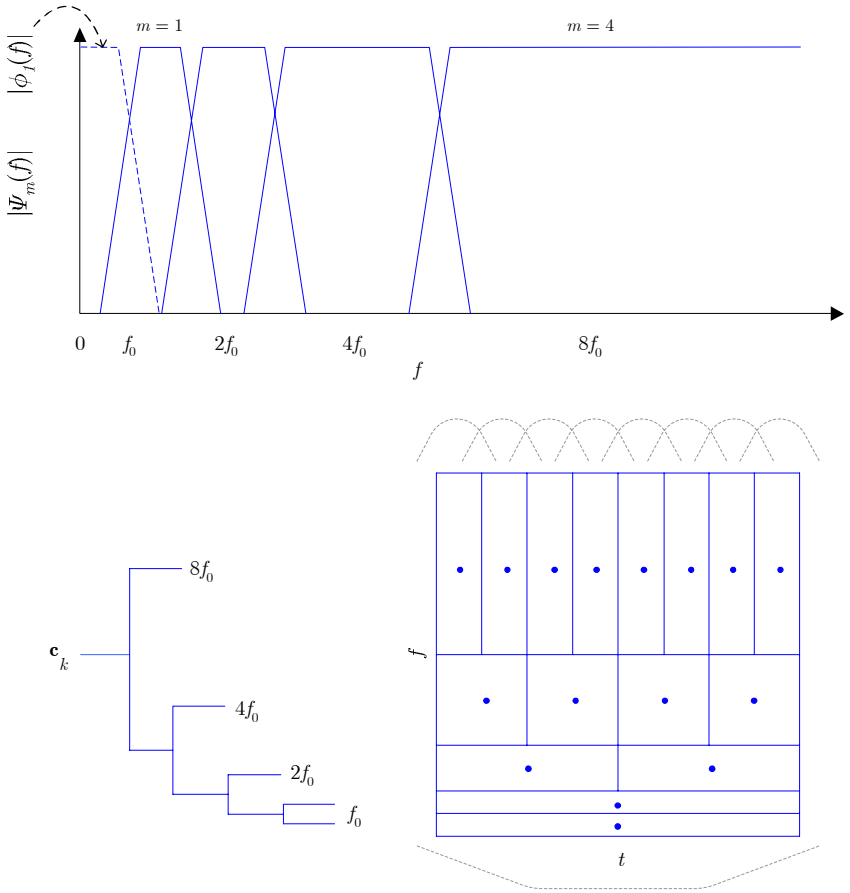

Figura 5.8: *Tipo de banco de filtros utilizado por la DDWT (arriba), representación esquemática del mismo (abajo izquierda, aprovechando la descomposición jerárquica del MRA) y partición $t-f$ asociada (abajo derecha). Compárese con el correspondiente a la STDFT (Fig. 5.6).*

Ejemplo 5.1 *La ondita más sencilla es la de Haar que tiene el valor 1 en el intervalo* $[0, 1/2)$*, -1 en el intervalo* $[1/2, 1]$*, y 0 para otros valores reales (Ver Figura 5.14) Para este caso la función de escala es la denominada función característica del intervalo [0,1]. Además se tiene la siguiente ecuación de escala:*

$$\phi(t) = \phi(2t) - \phi(2t-1), \tag{5.9}$$

y a partir de allí se puede obtener que $H(f) = 1/2\,(1 + \ e^{-j2\pi f})$*.*

Con esta notación, se presenta el siguiente teorema (c.f. Teorema 5.1.1 [34]) acerca de la existencia de onditas asociadas con un MRA:

Teorema 5.1 *Supóngase que* $(V_k)_{k\in\mathbb{Z}}$ *es un MRA con la función de escala* $\phi(t) \in V_0$*. Entonces la función* $\psi(t) \in W_0$ *es una ondita si y sólo si* $\Psi(f) = \ e^{j2\pi f/2} v(f) H^*(f/2 + 1/2)\Phi(f/2)$ *para alguna función de período unitario* $v(f)$ *con* $|v(f)| = 1$ *para casi todo punto. Cada ondita* ψ *tiene la propiedad que* $\{\psi_{k,n}(t)\}_{n\in Z}$ *es una base ortonormal de* W_k*.*

Si se quiere construir una ondita utilizando el teorema, es posible tomar $v(f) = 1$. Utilizando las propiedades del MRA se tiene la siguiente *ecuación de ondita*:

$$\psi(t) = 2 \sum_{n=-\infty}^{+\infty} g[n]\phi(2t-n). \tag{5.10}$$

Si se calcula la TF de esta ecuación como en (5.7) se obtiene:

$$\Psi(f) = G(f/2)\Phi(f/2), \tag{5.11}$$

donde $G(f) = 2\sum_n g[n]\ e^{-j2\pi fn}$. Aplicando el Teorema 5.1 se tiene que $G(f) = -\ e^{-j2\pi f} H(f + 1/2)$ y $g[n] = (-1)^{1-n} h[1-n]$. De esta forma G es un filtro pasa-altos, y junto con H forman una pareja de *filtros espejo conjugados*. En la Figura 5.9 puede observarse un ejemplo de la forma de los filtros $h[n]$ y $g[n]$ y las respectivas funciones escala y ondita, para el caso de la familia Symlets con 8 momentos nulos. En la Figura 5.10 se muestra la respuesta en frecuencia de ambos filtros, donde puede comprobarse la relación entre ellos.

Hasta aquí se ha revisado la forma en la que se puede diseñar una ondita, y su correspondiente función de escala, tal que permitan construir un MRA. La base de este desarrollo es la relación existente entre la ondita (función de escala) y el filtro pasa-alto (pasa-bajo) correspondiente dada por la ecuación de ondita (escala). A partir de allí es posible obtener un algoritmo de cálculo rápido que se discutirá a continuación.

5.3.1. Transformada ondita rápida

La DDWT posee además una implementación rápida denominada *transformada rápida ondita* (FWT). La FWT se implementa con un árbol de filtros diádicos y submuestreos por 2 según se verá a continuación. Generalmente no existen formas analíticas cerradas

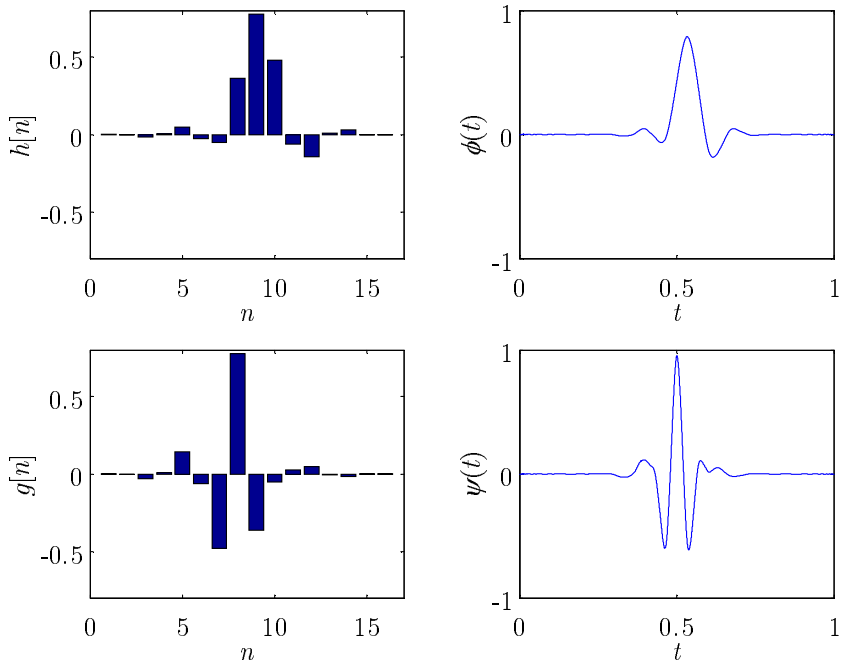

Figura 5.9: *Filtros $h[n]$ (arriba izquierda) y $g[n]$ (abajo izquierda) y las respectivas funciones escala (arriba derecha) y ondita (abajo derecha), para el caso de la familia Symlets con 8 momentos nulos.*

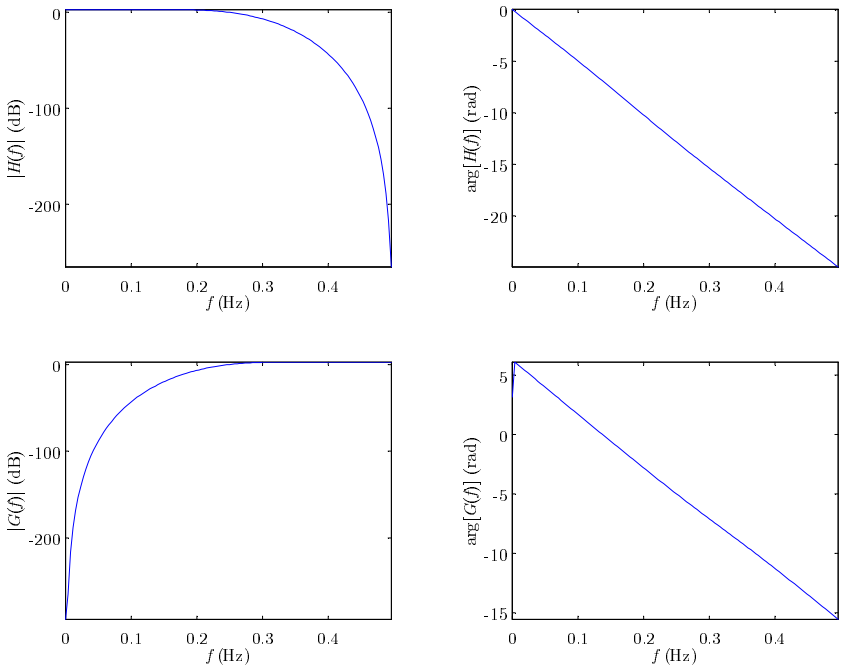

Figura 5.10: *Magnitud (derecha) y fase (izquierda) de los filtros $h[n]$ (arriba) y $g[n]$ (abajo), para el caso de la familia Symlets con 8 momentos nulos.*

para las onditas, y para describir la DDWT de una señal se pueden obtener las ecuaciones de análisis y síntesis sin necesidad de escribir explícitamente la ondita [157]. Sea $x(t) \in V_k$. Para el caso de la descomposición de la señal $x(t)$, como $V_k = V_{k-1} \oplus W_{k-1}$, se puede escribir $x(t)$ de las siguientes maneras:

$$x(t) = \sum_{n=-\infty}^{+\infty} c_k[n]\phi_{k,n}(t), \tag{5.12}$$

$$x(t) = \sum_{n=-\infty}^{+\infty} c_l[n]\phi_{l,n}(t) + \sum_{m=l}^{k-1}\sum_{n=-\infty}^{+\infty} d_m[n]\psi_{m,n}(t) \quad \text{para} \quad k > l, \tag{5.13}$$

donde los c y d constituyen los coeficientes de aproximación y detalle respectivamente.

En la práctica, el interés en general no está en $x(t)$, si no más bien en la señal muestreada. En ese caso es posible mantener las ideas anteriores, reemplazando ahora la TF por la *transformada Z*, y la señal muestreada pasa a ser "equivalente" a los coeficientes c a la escala k.

Esta última equivalencia no debe tomarse en el sentido tradicional de que $x[n]$ constituye un muestreo uniforme sobre $x(t)$ [127]. Más bien se supone que $x[n]$ se obtiene mediante un dispositivo de resolución finita que muestrea y promedia la señal $x(t)$. Si la distancia entre muestras es N^{-1}, se necesita asociar $x[n]$ a una función $x(t) \in V_k$ aproximada a la escala $2^k = N^{-1}$, y calcular $c_k[n] = \langle x(t), \phi_{k,n}(t)\rangle$.

Un posible camino consiste en escribir $x(t)$ de la siguiente forma [127]:

$$x(t) = \sum_{n=-\infty}^{+\infty} x[n]\phi\left(\frac{t - 2^k n}{2^k}\right) \quad \in \quad V_k. \tag{5.14}$$

Dado que la familia $\left\{\phi_{k,n}(t) = 2^{-k/2}\phi(2^{-k}t - n)\right\}_{n\in\mathbb{Z}}$ es ortonormal y $2^k = N^{-1}$ se tiene que:

$$x[n] = N^{1/2}\left\langle x(t), \phi_{k,n}(t)\right\rangle = N^{1/2}c_k[n]. \tag{5.15}$$

A partir de allí, teniendo en cuenta que $\int_{-\infty}^{+\infty} \phi(t)dt = 1$, y si $x(t)$ es regular es posible escribir:

$$x[n] = N^{1/2}c_k[n] \approx x(N^{-1}n), \tag{5.16}$$

lo que demuestra que la equivalencia planteada suele ser, en general, bastante buena.

A partir de (5.7), (5.10), (5.13) y de los resultados anteriores es posible obtener las siguientes expresiones recursivas:

$$c_{m-1}[n] = \sqrt{2}\sum_{i=-\infty}^{+\infty} h[i - 2n]\, c_m[i], \tag{5.17}$$

$$d_{m-1}[n] = \sqrt{2}\sum_{i=-\infty}^{+\infty} g[i - 2n]\, c_m[i], \tag{5.18}$$

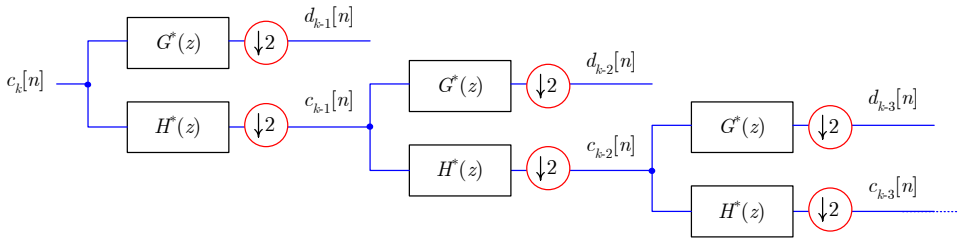

Figura 5.11: *Esquema del algoritmo de la FWT.*

donde $k > l$ y $m = l, \ldots, k-1$. Estas ecuaciones constituyen la base para el algoritmo de la FWT. Se puede observar como las onditas y las funciones escala aparecen aquí reemplazadas por las respectivas respuestas al impulso de los filtros pasa-altos y pasa-bajos. Así mismo los productos se han convertido en convoluciones.

Para la inversión (o reconstrucción de la señal) es posible partir de la descomposición, de manera de obtener nuevamente los coeficientes mediante otra expresión recursiva:

$$c_m[n] = \sum_{i=-\infty}^{+\infty} h[n-2i]c_{m-1}[i] + \sum_{i=-\infty}^{+\infty} g[n-2i]d_{m-1}[i], \tag{5.19}$$

donde $k > l$ y $m = l, \ldots, k-1$.

En ambos casos se puede lograr el análisis y la síntesis por medio de bancos de filtros. En las aplicaciones es importante contar con filtros de *respuesta finita al impulso* (FIR). En este caso h y g tendrán un número finito de coeficientes distintos de cero. Ésto corresponde al caso en el que la función de escala, y por ende la ondita, tienen soporte compacto. En la Figura 5.11 se muestra el esquema de la FWT que corresponde a la descomposición o análisis.

Hasta aquí se ha considerado el caso discreto pero todavía infinito. Supóngase ahora que $x(t)$ tiene su soporte en $[0, 1]$ y que se aproxima mediante un muestreo uniforme a intervalos de N^{-1}. La aproximación resultante c_k tiene $N = 2^{-k}$ muestras. Al calcular las convoluciones con $h[n]$ (5.17) y con $g[n]$ (5.18) cerca de los "bordes" de la señal se requiere conocer los valores de c_m más allá de estos bordes. Una solución sencilla consiste en reemplazar estas convoluciones lineales por convoluciones circulares. Esta estrategia suele producir coeficientes onditas muy grandes cerca de los bordes. Una solución más eficiente consiste en la utilización de filtros de borde especiales [127].

En cuanto al costo computacional, suponiendo que h y g poseen K coeficientes diferentes de cero, y $x[n]$ es una señal discreta de tamaño $N = 2^{-k}$, se requieren un total de $\mathcal{O}(2KN)$ operaciones para completar el algoritmo de la FWT.

Con las precauciones anteriores, y habiendo realizado la descomposición hasta la escala l, es posible finalmente escribir nuevamente la ecuación de síntesis de la señal discreta y finita $\mathbf{x} \in \mathbb{R}^N$ mediante una multiplicación matricial de la forma:

$$\mathbf{x} = \mathbf{\Phi}'\mathbf{c}',$$

donde $\mathbf{c}' = [c_l[0, \ldots, (N/2^l) - 1]; d_{k-1}[0, \ldots, (N/2) - 1], \ldots, d_l[0, \ldots, (N/2^l) - 1]]$, para $k > 0$ y las columnas de $\mathbf{\Phi}'$ son las versiones discretas de las onditas (y la función de

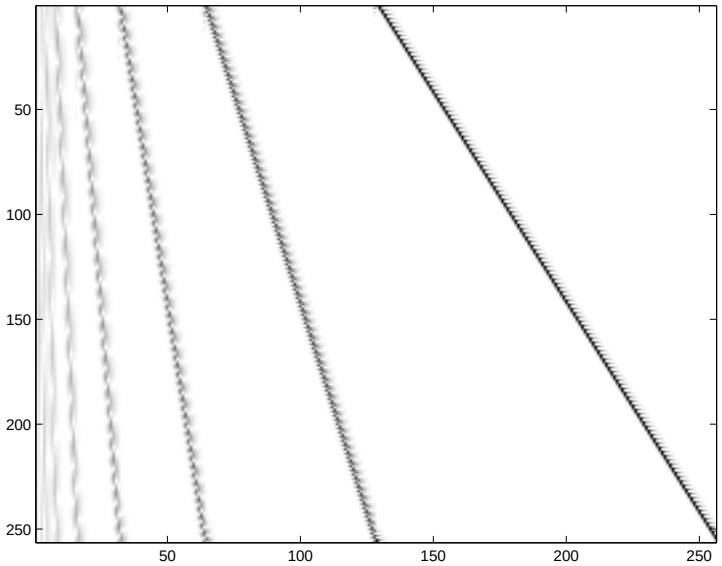

Figura 5.12: *Estructura del diccionario $\mathbf{\Phi}'$ de la DDWT para $N = 256$ y la ondita Daubechies con 8 momentos nulos (orden 8).*

escala), calculadas mediante (5.19) utilizando convoluciones circulares y como inicialización $c_0[n] = \delta[n]$ y $d_k[n] = 0$ ($c_0[n] = 0$, $d_0[n] = \delta[n]$ y $d_k[n] = 0$ para la función escala):

$$\mathbf{\Phi}' = \left[\left\{ \phi_l[i - 2^{l-k+1}n] \right\}_{n\in\mathbb{Z}} ; \left\{ \psi_m[i - 2^{m-k+1}n] \right\}_{k-1<m\leq l, n\in\mathbb{Z}} \right], \tag{5.20}$$

donde el subíndice $0 < i \leq N-1$ corresponde a las filas, y las columnas varían con m, n y l hasta completar N columnas en total. Una propiedad importante es que nuevamente $\mathbf{\Phi}'$ constituye una transformación unitaria. Se debe aclarar que este desarrollo corresponde al caso ortogonal, mientras que para el biortogonal es posible realizar un desarrollo similar pero utilizando diferentes diccionarios para análisis y para síntesis [127].

5.3.2. Familias de onditas

Una de las ventajas de la transformada onditas es que con condiciones bastante generales se tiene a disposición una gran cantidad de funciones con distintas características, que dan lugar a diferentes familias de onditas. Esta ventaja trae aparejado el problema de la elección de la familia óptima para la aplicación particular. Aunque no existe un único criterio para realizar esta elección, es posible relacionar las familias con características deseables para el análisis logrado:

- Reales o analíticas: para algunas aplicaciones puede ser importante trabajar con bases o diccionarios complejos. Por ejemplo si se quiere separar componentes frecuenciales en magnitud y fase. Por otra parte resulta más sencillo el tratamiento para el caso real, que suele ser útil para detección de cambios en la señal analizada.

- Expresión analítica: aunque resulta poco común es interesante que la ondita posea una expresión analítica cerrada.

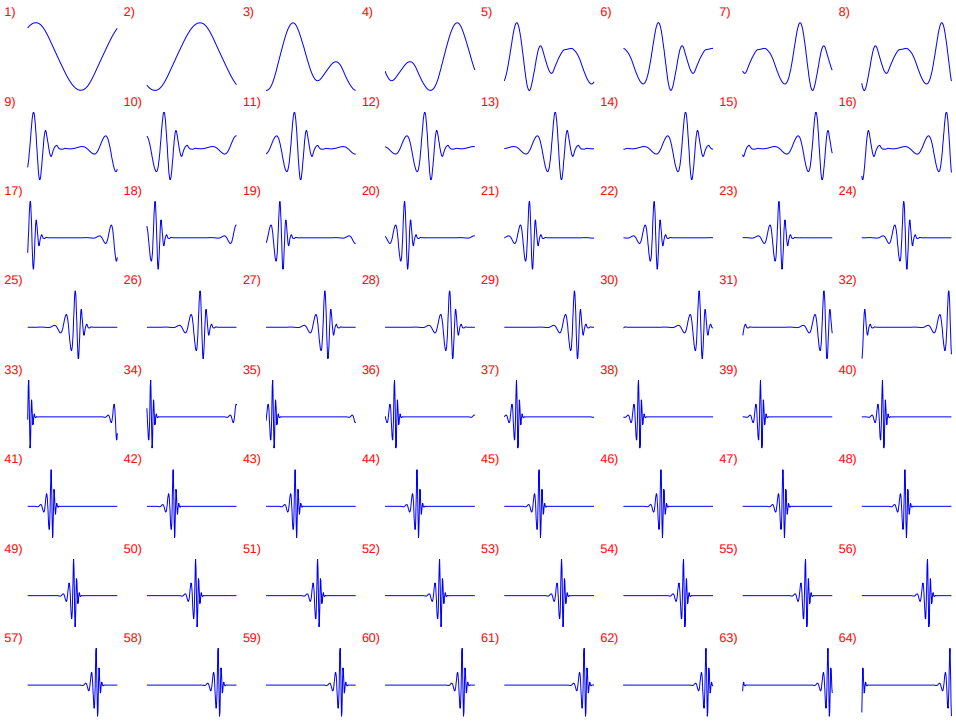

Figura 5.13: *Algunos átomos del diccionario* $\mathbf{\Phi}'$ *de la DDWT para* $N = 256$ *y la ondita Daubechies con 8 momentos nulos (orden 8).*

- Ortogonalidad: la ortogonalidad posee varias ventajas prácticas, según se discutió, entre las que se pueden citar el poder lograr una transformación unitaria.

- Soporte Compacto: si las onditas poseen soporte compacto, entonces los filtros correspondientes tienen respuesta al impulso finita. En el caso de no tener soporte compacto, se desea que tengan un decaimiento rápido.

- Simetría: si las onditas son simétricas ésto permite que los filtros sean de fase lineal. Este hecho resulta importante en algunas aplicaciones de procesamiento de señales. Para ello se requiere un esquema biortogonal.

- Coeficientes racionales: la posibilidad de trabajar con coeficientes raciones es importante desde el punto de vista computacional.

- Regularidad o suavidad: la regularidad es una propiedad deseable en algunas aplicaciones como la compresión de señales. Además, la suavidad de la ondita se corresponde con una mejor localización en frecuencia de los filtros. La suavidad está relacionada también con el número de *momentos nulos* de la ondita.

- Localización $t - f$: la localización resulta especialmente importante en la mayoría de las aplicaciones.

- Otras: existen otras características deseables como la posibilidad de utilización de

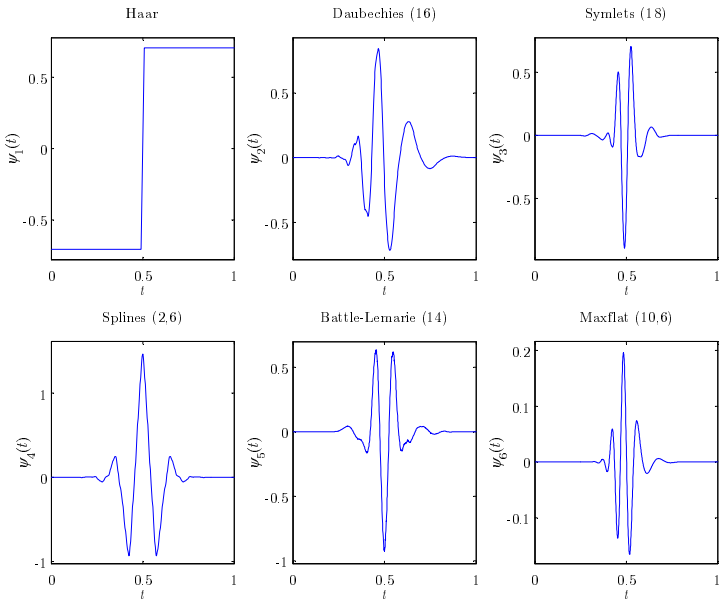

Figura 5.14: *Ejemplos de onditas madres correspondientes a diferentes familias (y parámetros).*

filtros especiales, el diseño de onditas basadas en modelos (por ejemplo auditivos), etc.

Esto permite que, para una aplicación particular, sea posible elegir aquella familia que mejor resalte las propiedades distintivas de la señal bajo estudio. En la Figura 5.14 se presentan algunos ejemplos de onditas madre y en la Figura 5.15 se muestra la variación de las características de una ondita de acuerdo con el valor de sus parámetros asociados. Para una revisión más amplia acerca de este aspecto ver [167].

5.4. Tranformada paquetes de onditas

La *transformada paquetes de onditas* (WPT) surge de la utilización de un razonamiento sugerido por Wickerhauser [213], que generaliza el análisis multiresolución. De acuerdo con este enfoque es posible descomponer también las componentes de alta frecuencia (detalles), en la misma forma que los componentes de baja frecuencia (aproximaciones) según se verá a continuación [127].

Sea V_k un espacio de aproximación multiresolución que se descompone en un espacio V_{k+1} de menor resolución y un espacio de detalle W_{k+1}. Ésto se realiza dividiendo la base ortogonal $\left\{\phi_k(t-2^k n)\right\}_{n\in\mathbb{Z}} \in V_k$ en dos nuevas bases ortogonales:

$$\left\{\phi_{k+1}(t-2^{k+1}n)\right\}_{n\in\mathbb{Z}} \in V_{k+1} \quad \text{y} \quad \left\{\psi_{k+1}(t-2^{k+1}n)\right\}_{n\in\mathbb{Z}} \in W_{k+1}. \tag{5.21}$$

Como se ha visto en la Sección 5.3, las descomposiciones de ϕ_{k+1} y ψ_{k+1} en la base $\left\{\phi_k(t-2^k n)\right\}_{n\in\mathbb{Z}}$ son especificadas por un par de filtros de cuadratura espejo $h[n]$ y $g[n] = (-1)^{1-n}h[1-n]$.

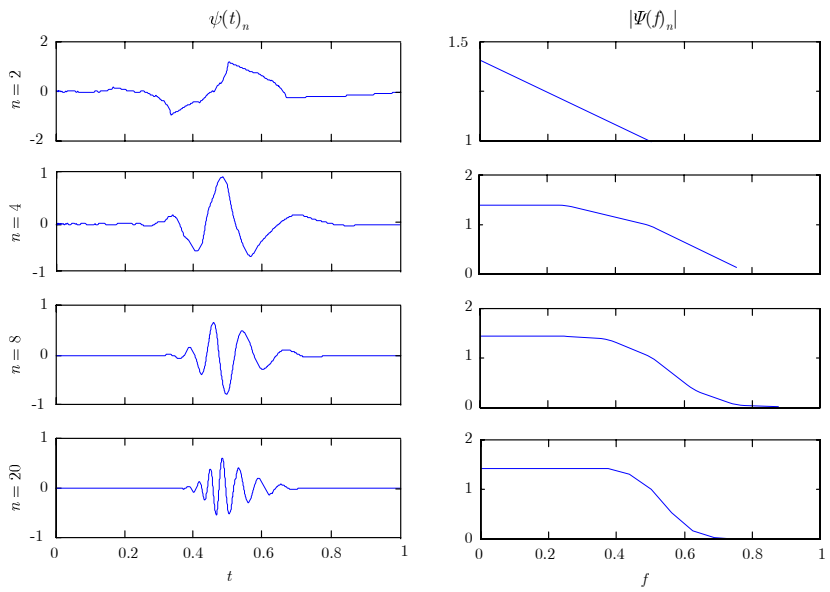

Figura 5.15: *Ejemplos de onditas de la familia de Daubechies correspondientes a diferentes cantidades de momentos nulos (izquierda) y sus respectivas magnitudes espectrales (derecha).*

El siguiente teorema generaliza este resultado a cualquier espacio U_k que admita una base ortogonal de funciones trasladadas por $n2^k$, para $n \in \mathbb{Z}$ [127].

Teorema 5.2 *Sea* $\left\{\theta_k(t-2^k n)\right\}_{n\in\mathbb{Z}}$ *una base ortonormal del espacio* U_k*. Sean* h *y* g *un par de filtros espejo conjugados. Se define:*

$$\theta^0_{k+1}(t) = \sum_{n=-\infty}^{n=+\infty} h[n]\ \theta_k(t-2^k n) \quad \text{y} \quad \theta^1_{k+1}(t) = \sum_{n=-\infty}^{n=+\infty} g[n]\ \theta_k(t-2^k n). \tag{5.22}$$

Entonces la familia:

$$\left\{\theta^0_{k+1}(t-2^{k+1}n), \theta^1_{k+1}(t-2^{k+1}n)\right\}_{n\in\mathbb{Z}},$$

es una base ortonormal de U_k*.*

El Teorema 5.2 muestra que los filtros espejo conjugados transforman una base ortogonal $\left\{\theta_k(t-2^k n)\right\}_{n\in\mathbb{Z}}$ en dos familias ortogonales $\left\{\theta^0_{k+1}(t-2^{k+1}n)\right\}_{n\in\mathbb{Z}}$ y $\left\{\theta^1_{k+1}(t-2^{k+1}n)\right\}_{n\in\mathbb{Z}}$. Sea U^0_{k+1} y U^1_{k+1} los espacios generados por cada una de estas familias. Claramente U^0_{k+1} y U^1_{k+1} son ortogonales y:

$$U^0_{k+1} \oplus U^1_{k+1} = U_k.$$

Si se calcula la FT de (5.22) es posible relacionar los espectros de θ^0_{k+1} y θ^1_{k+1}, con el de θ_k:

$$\Theta^0_{k+1}(f) = H(2^k f)\ \Theta_k(f), \quad \Theta^1_{k+1}(f) = G(2^k f)\ \Theta_k(f). \tag{5.23}$$

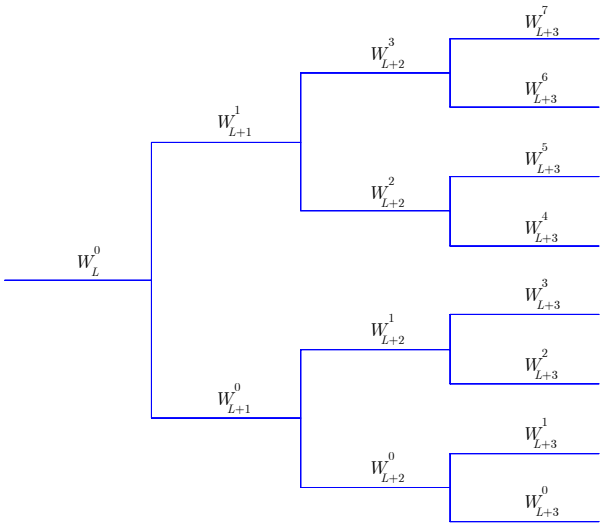

Figura 5.16: *Ejemplo de árbol binario de espacios de paquetes de onditas.*

Como las funciones de transferencia $H(2^k f)$ y $G(2^k f)$ poseen su energía concentrada en diferentes intervalos de frecuencia, esta transformación puede interpretarse como una división del soporte en frecuencia de $\Theta_k(f)$.

En lugar de dividir sólo los espacios de aproximación V_k para construir los espacios de detalle W_k y las bases de onditas tradicionales, se puede también dividir los espacios de detalle. El Teorema 5.2 afirma que es posible hacer $U_k = W_k$ y dividir estos espacios de detalle para obtener nuevas bases. La partición recursiva de los espacios vectoriales puede ser representada mediante un árbol binario. Si las señales son aproximadas a la escala 2^L, a la raíz del árbol se le asocia el espacio de aproximación V_L. Este espacio admite una base ortogonal de funciones de escala $\left\{\phi_L(t-2^L n)\right\}_{n\in\mathbb{Z}}$ con $\phi_L(t) = 2^{-L/2}\phi(2^{-L}t)$.

Cada nodo del árbol binario está etiquetado por (k,p), donde $k-L \geq 0$ corresponde a la profundidad del árbol, y p es el número de nodos que quedan debajo de él a la misma profundidad $k-L$. La Figura 5.16 muestra un ejemplo de la descomposición en subespacios de un árbol binario de WPT. Para cada nodo (k,p) se asocia un espacio W_k^p, que admite una base ortonormal $\left\{\psi_k^p(t-2^k n)\right\}_{n\in\mathbb{Z}}$, a medida que se desciende en el árbol. En la raíz se tiene $W_L^0 = V_L$ y $\psi_L^0 = \phi_L$. Supóngase que se ha construido W_k^p y su base ortonormal $\mathcal{B}_k^p = \left\{\psi_k^p(t-2^k n)\right\}_{n\in\mathbb{Z}}$ en el nodo (k,p). En el Teorema 5.2 se demuestra que $\mathcal{B}_{k+1}^{2p} = \left\{\psi_{k+1}^{2p}(t-2^{k+1}n)\right\}_{n\in\mathbb{Z}}$ y $\mathcal{B}_{k+1}^{2p+1} = \left\{\psi_{k+1}^{2p+1}(t-2^{k+1}n)\right\}_{n\in\mathbb{Z}}$ son bases ortonormales correspondientes a dos espacios ortogonales W_{2p}^{k+1} y W_{2p+1}^{k+1} de manera tal que:

$$W_{2p}^{k+1} \oplus W_{2p+1}^{k+1} = W_p^k. \tag{5.24}$$

Esta partición recursiva define un árbol binario de espacios de paquetes de onditas que depende de los filtros $h[n]$ y $g[n]$. En la Figura 5.17 se muestran los 8 paquetes de onditas ψ_k^p a la profundidad $k-L$ (con $k=3$ y $L=0$), calculados con el filtro de Daubechies de orden 5.

En un sentido un poco más general es posible ver a cada paquete de onditas $\psi(t)$

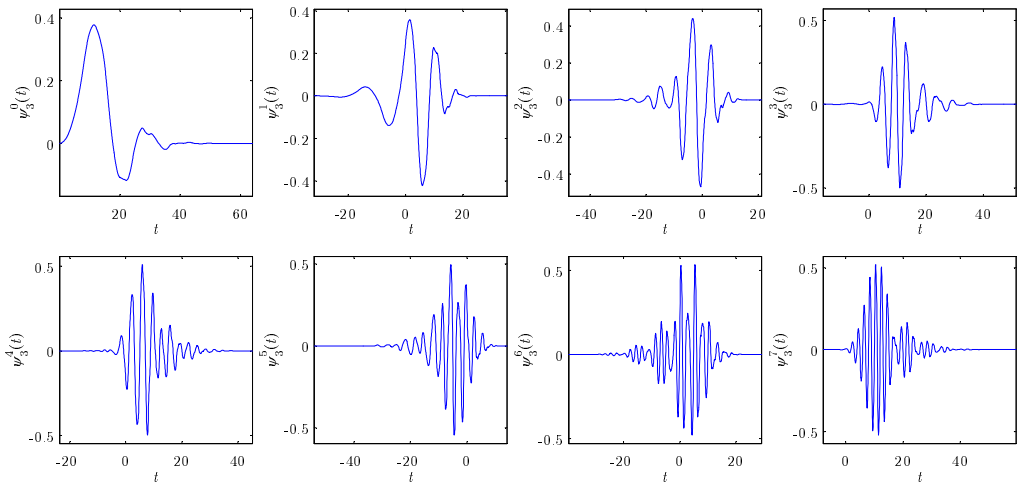

Figura 5.17: *Paquetes de onditas para la profundidad $k = 3$ del árbol correspondiente, calculados con el filtro de Daubechies (5). Están ordenados de izquierda a derecha y de arriba a abajo, desde las frecuencias bajas a las altas.*

como una función en $L^2(\mathbb{R})$ bien localizada tanto en el tiempo como en la frecuencia. Un ejemplo podría ser una nota musical. Es posible entonces describir a cada uno de estos átomos por medio de sus características temporales y frecuenciales. En la Figura 5.18 se muestra un ejemplo generado a partir de la ondita de Daubechies. Desde esta perspectiva es posible asignar 3 parámetros que permiten identificar a esta función a saber: localización temporal, escala y frecuencia. La primera y la tercera se pueden calcular a partir de los centros de masa de $|\psi(t)|^2$ y $|\Psi(f)|^2$. El segundo parámetro se puede obtener a partir del ancho característico de $|\psi(t)|^2$ o lo que es equivalente, la incertidumbre temporal. Por el principio de Heisenberg, éste resulta también recíproco de la incertidumbre en la frecuencia. Con estas ideas es sencillo construir ejemplos de estas funciones basados en ondas moduladas, lo que da lugar a la siguiente definición [213].

Definición 5.5 *Sea una señal $\psi(t)$ para la cual se definen los operadores de modulación, dilatación y traslación como sigue:*

$$\mu_f \psi(t) = e^{jft}\psi(t), \tag{5.25}$$

$$\delta_a \psi(t) = a^{1/2}\psi(t/a), \tag{5.26}$$

$$\tau_b \psi(t) = \psi(t-b). \tag{5.27}$$

La colección de funciones $\psi(t)$ moduladas, dilatadas (escaladas) y trasladadas forma una familia de paquetes de onditas *con parámetros f, a, b.*

Los operadores de la Definición 5.5 conservan la energía, por lo que las ondas pueden ser normalizadas a un vector unitario en $L^2(\mathbb{R})$. La componente de una función $x(t)$ en f, a, b es, como antes, el producto interno de $x(t)$ con la onda modulada cuyos parámetros son f, a y b. Si este valor es grande se podría concluir que $x(t)$ tiene energía considerable a la escala a cerca de la frecuencia f y la posición temporal b.

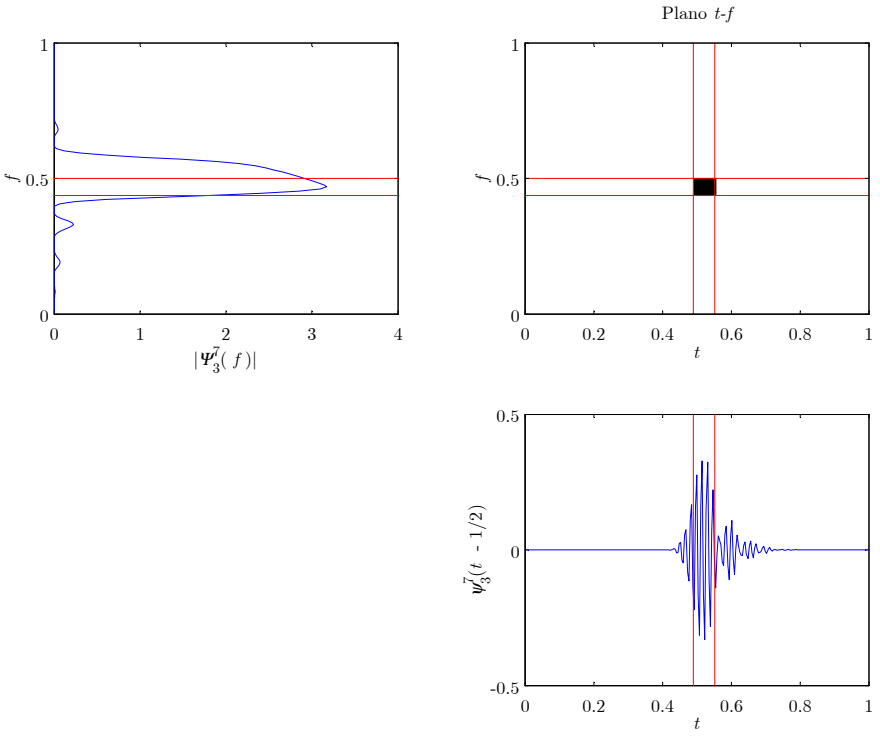

Figura 5.18: *Representación de un átomo WPT en el dominio de la frecuencia $|\Psi(f)|$ (arriba izquierda), en el dominio del tiempo $|\psi(t)|$ (abajo derecha), y el correspondiente rectángulo de Heisenberg en el plano $t - f$, calculado mediante el filtro de Daubechies (5).*

$x[0]$	$x[1]$	$x[2]$	$x[3]$	$x[4]$	$x[5]$	$x[6]$	$x[7]$
$c[0]$	$c[1]$	$c[2]$	$c[3]$	$d[0]$	$d[1]$	$d[2]$	$d[3]$
$cc[0]$	$cc[1]$	$dc[0]$	$dc[1]$	$cd[0]$	$cd[1]$	$dd[0]$	$dd[1]$
$ccc[0]$	$dcc[0]$	$cdc[0]$	$ddc[0]$	$ccd[0]$	$dcd[0]$	$cdd[0]$	$ddd[0]$

Figura 5.19: *Rectángulo diádico de coeficientes paquetes de onditas para $k = 3$. Para las muestras de la señal $x[n]$ asociadas a la raíz del árbol deben tenerse en cuenta similares consideraciones a las discutidas en la Sección 5.3.1. Para resaltar el efecto de las particiones en subespacios de detalle y aproximación se ha modificado ligeramente la notación utilizando secuencias c...d para cada coeficiente (adaptado de [213]).*

El conjunto de funciones $\phi(t)$ definidas en la forma anterior forma una biblioteca paramétrica de funciones. Por supuesto que para el caso general con $f, a, b \in \mathbb{R}$, esta familia no constituye una base ortogonal. Sin embargo, es posible también parametrizar un conjunto ortogonal obtenido a partir de un árbol binario en estos términos. Este esquema se puede generalizar para árboles binarios arbitrarios con propiedades similares de ortogonalidad, obtenidos utilizando diferentes filtros.

Una forma útil de representar las relaciones entre los coeficientes de los paquetes de onditas, dadas por un árbol binario de subespacios como el de la Figura 5.16, es a través de un rectángulo de bloques diádicos como el de la Figura 5.19. Aquí el número de fila dentro del rectángulo indica la escala del paquete de onditas contenido dentro. El número de columna indica los parámetros de la frecuencia y la posición. Es posible seleccionar un grupo de paquetes de onditas tanto por la posición como por la frecuencia. Agrupando por la posición se llenará cada fila del rectángulo con los espectros adyacentes en frecuencia, análogo a lo que se obtendría con la STDFT, con una ventana de tamaño determinado por el número de fila y la posición de la ventana correspondiente a la ubicación del grupo. El parámetro frecuencia se incrementa dentro del grupo.

5.4.1. Cantidad de bases de paquetes de ondita

Como se ha visto cada una de las bases ortogonales obtenidas a partir de la WPT puede verse como un árbol binario. Por lo tanto es necesario definir cuales de estos árboles pueden considerarse como *admisibles*, en el sentido de que generan una base ortogonal [127].

Definición 5.6 *Se denomina* árbol admisible *a cualquier árbol binario cuyos nodos poseen 0 o 2 hijos. Sean $\{k_i, p_i\}_{1 \leq i \leq I}$ las hojas de un árbol binario admisible. Aplicando la partición recursiva (5.24) a lo largo de las ramas del árbol, se verifica que los espacios $\left\{W_{k_i}^{p_i}\right\}_{1 \leq i \leq I}$ son mutuamente ortogonales y se suman hasta W_L^0:*

$$W_L^0 = \bigoplus_{i=1}^{I} W_{k_i}^{p_i}, \tag{5.28}$$

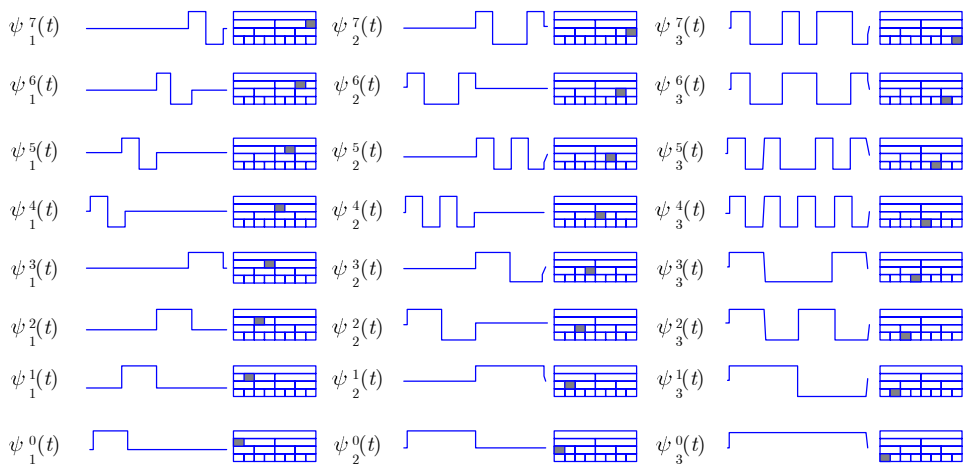

Figura 5.20: *Todos los paquetes de onditas de Haar para un árbol de profundidad $k = 3$ y las posiciones correspondientes en el rectángulo diádico de coeficientes (adaptado de [213]).*

La unión de las correspondientes bases de paquetes de onditas define entonces una base ortogonal para $W_L^0 = V_L$:

$$\left\{\psi_{k_i}^{p_i}(t - 2^{k_i}n)\right\}_{n\in\mathbb{Z},1\leq i\leq I}. \tag{5.29}$$

El número de bases ortogonales de paquetes de onditas para V_L es entonces igual al número de árboles binarios admisibles. La siguiente Proposición da cuenta de la cantidad de árboles diferentes que pueden formarse con una determinada profundidad [127].

Proposición 5.1 *El número B_K de bases de paquetes de onditas en un árbol binario completo de profundidad K satisface la siguiente relación:*

$$2^{2^{K-1}} \leq B_K \leq 2^{\frac{5}{4}2^{K-1}}. \tag{5.30}$$

5.4.2. Transformada paquetes de ondita rápida

Como se ha visto la DDWT es realmente un subconjunto de la WPT. La WPT generaliza el análisis tiempo-frecuencia realizado por la DDWT, dando como resultado una familia de bases ortonormales, una de las cuales es la DDWT. De la misma manera que la FWT, existe un algoritmo rápido (obtenido a partir de ésta) para el cálculo de la WPT denominado *transformada rápida de paquetes de onditas* (FWPT). La Figura 5.21 se muestra el esquema de la descomposición correspondiente, en forma de un banco de filtros, para la implementación del algoritmo de cálculo rápido, ejemplificado para un árbol completo de profundidad $k = 3$ o cuasi-Fourier. La partición $t - f$ asociada para este tipo de árbol puede observarse en la Figura 5.22.

Para el caso de señales discretas y finitas deben tenerse en cuenta similares consideraciones a las discutidas en la Sección 5.3.1. De manera análoga al caso de la DDWT es posible ahora escribir nuevamente la ecuación de síntesis de $\mathbf{x} \in \mathbb{R}^N$ en forma matricial:

$$\mathbf{x} = \mathbf{\Phi c}, \tag{5.31}$$

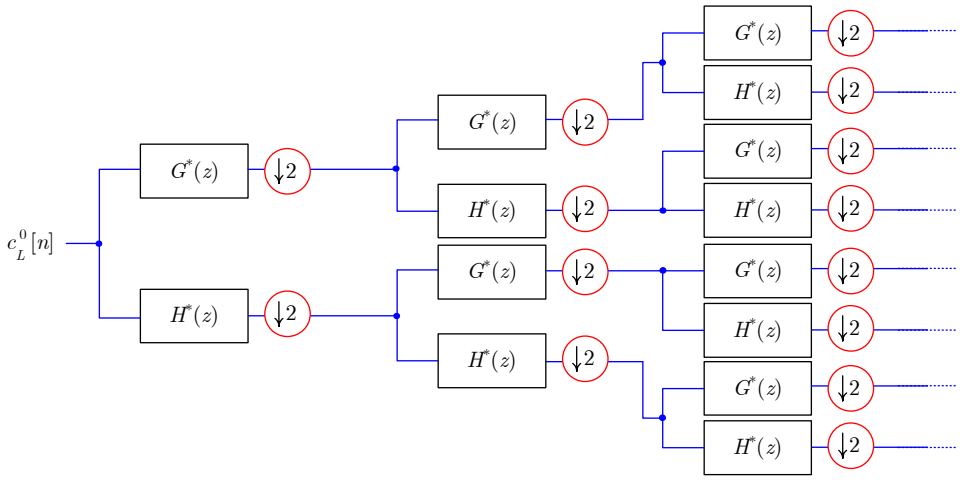

Figura 5.21: *Algoritmo rápido para el cálculo la WPT obtenido a partir del de la FWT (ejemplificado para el caso cuasi-Fourier).*

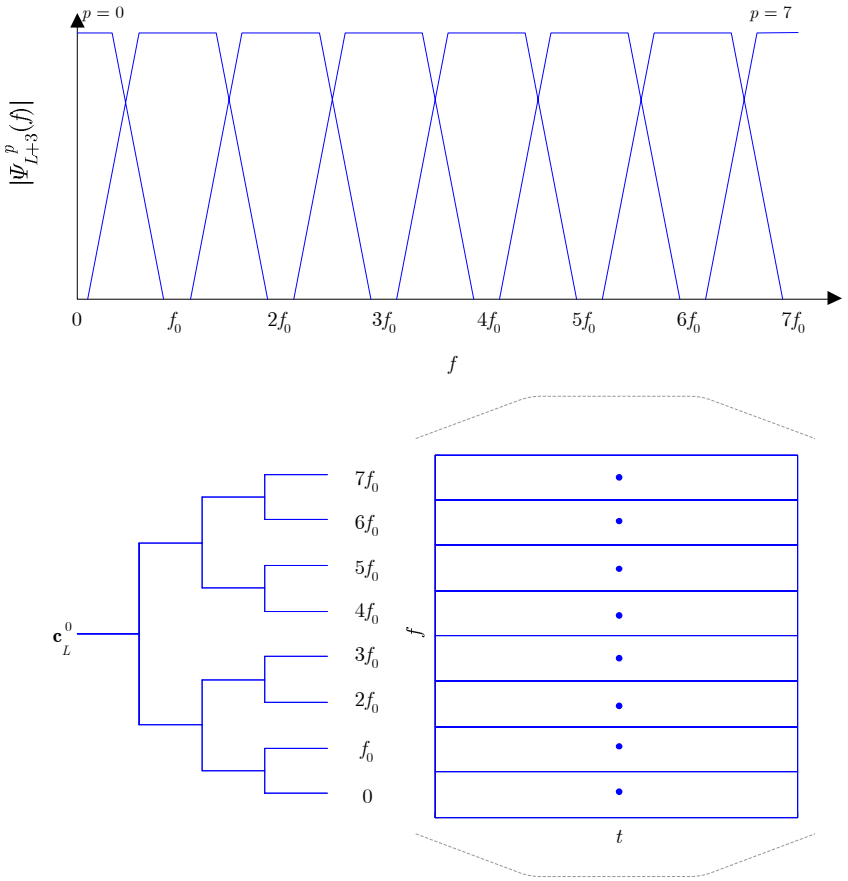

Figura 5.22: *Tipo de banco de filtros generado por la WPT ejemplificado para el caso cuasi-Fourier (arriba), representación esquemática del mismo (abajo izquierda) y partición $t - f$ asociada (abajo derecha). Compárese con el correspondiente a la STDFT (Fig. 5.6).*

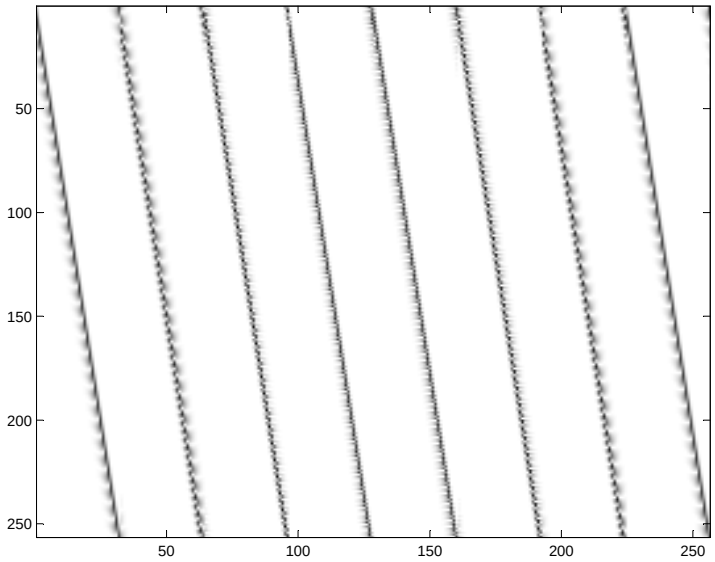

Figura 5.23: *Estructura del diccionario* $\mathbf{\Phi}$ *de la WPT (cuasi-Fourier,* $k = 3$*) para* $N = 256$ *y la ondita Daubechies con 8 momentos nulos (orden 8).*

donde $\mathbf{c}$ es el vector de coeficientes y $\mathbf{\Phi}$ es la matriz cuyas columnas están formadas por los paquetes de onditas discretos, obtenidas a partir del árbol binario correspondiente, en forma recursiva mediante el algoritmo rápido con las inicializaciones adecuadas. En las Figuras 5.23 y 5.24 se puede apreciar la estructura de la matriz $\mathbf{\Phi}$ y algunos átomos para el caso cuasi-Fourier. En cuanto al costo computacional, suponiendo que h y g poseen K coeficientes diferentes de cero, y $x[n]$ es una señal discreta de tamaño $N = 2^{-L}$, para el árbol completo de profundidad $\log_2 N$ se requieren un total de $\mathcal{O}(KN \log_2 N)$ operaciones para completar el algoritmo de la FWPT [127].

5.5. Transformada paquetes de cosenos

Hasta aquí se han presentado las diferentes variantes dentro de la familia de transformaciones basadas en onditas. La idea principal consiste en dividir el eje de las frecuencias en forma diádica mediante bases ortogonales adecuadas. Ahora bien, mediante funciones del tipo cosenos ventaneados es posible obtener una colección de bases ortogonales denominadas *bases de cosenos locales*(LCB). Estas bases permiten segmentar el eje temporal en intervalos solapados $[a_p, a_{p+1}]$ de longitud arbitraria l_p [127]:

$$\left\{ g_{p,k}(t) = g_p(t) \sqrt{\frac{2}{l_p}} \cos\left[\pi \left(k + \frac{1}{2} \right) \frac{t - a_p}{l_p} \right] \right\}_{k \in \mathbb{N}, p \in \mathbb{Z}},$$

donde $g_p(t)$ es una función ventana que posee propiedades de simetría y cuadratura en los intervalos de solapamiento. Este resultado es más general que la construcción de las bases WPT, que sólo pueden dividir el eje de frecuencia en intervalos diádicos cuya longitud es proporcional a potencias de 2. Sin embargo Coifman y Meyer [25] mostraron

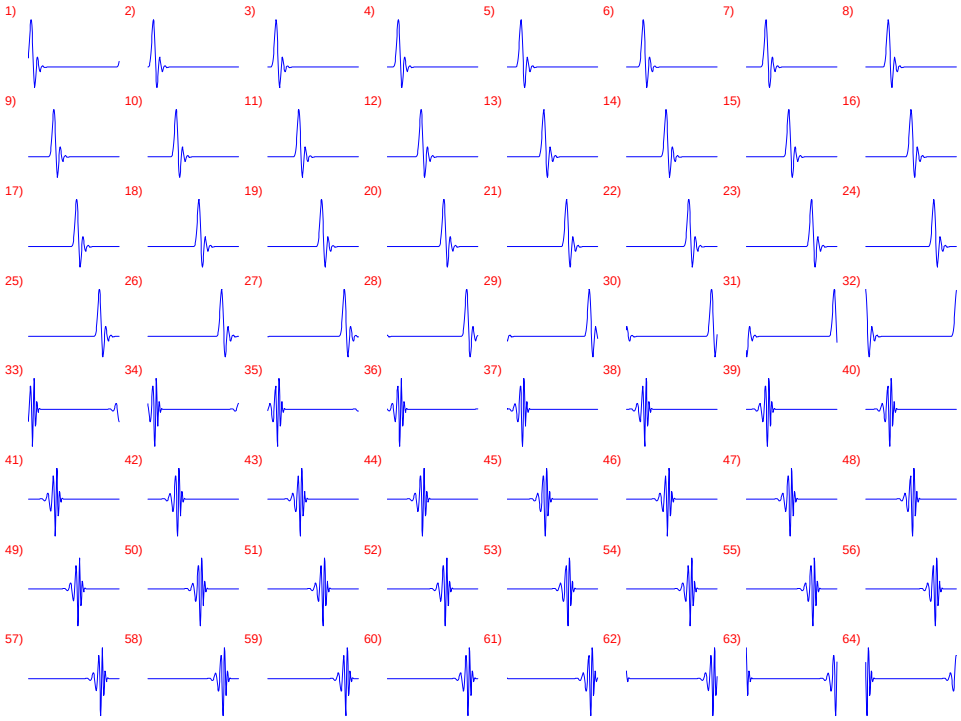

Figura 5.24: *Algunos átomos del diccionario* $\mathbf{\Phi}$ *de la WPT (cuasi-Fourier,* $k = 3$*) para* $N = 256$ *y la ondita Daubechies con 8 momentos nulos (orden 8).*

que si se restringen estos intervalos a tamaños diádicos, entonces se puede crear una estructura de árboles similar a la de los WPT. A las funciones obtenidas a partir de estos árboles se las denomina *paquetes de cosenos* (CPT) [127]. De esta manera los CPT se construyen dividiendo recursivamente espacios construidos con funciones LCB, lo que resulta también en un algoritmo de cálculo rápido. En las Figuras 5.25 y 5.26 se muestra la estructura y algunos átomos del diccionario $\mathbf{\Phi}$ que permite la síntesis de una señal discreta y finita $\mathbf{x}$. Debido a que las consideraciones a realizar son similares al caso de la WPT no se darán aquí más detalles acerca de la CPT a fin de no engrosar innecesariamente el presente volumen [127].

5.6. Comentarios de cierre del capítulo

En este capítulo se han descripto una serie de técnicas de análisis de señales discretas, basadas principalmente en la teoría de onditas, que permiten generar una gran variedad de bases ortogonales, marcos o diccionarios (Ver Figura 5.27). Dentro de este espectro adicional de posibilidades de representación, que se suman a las presentadas en los capítulos anteriores, el problema consiste nuevamente en elegir la base o diccionario más adecuado para una aplicación particular.

Para el caso de la utilización de diccionarios que no formen una base ortogonal (por ejemplo sobrecompletos) se requiere seleccionar algún subconjunto de elementos que sí constituyan una base, o bien realizar consideraciones adicionales. Estos aspectos serán

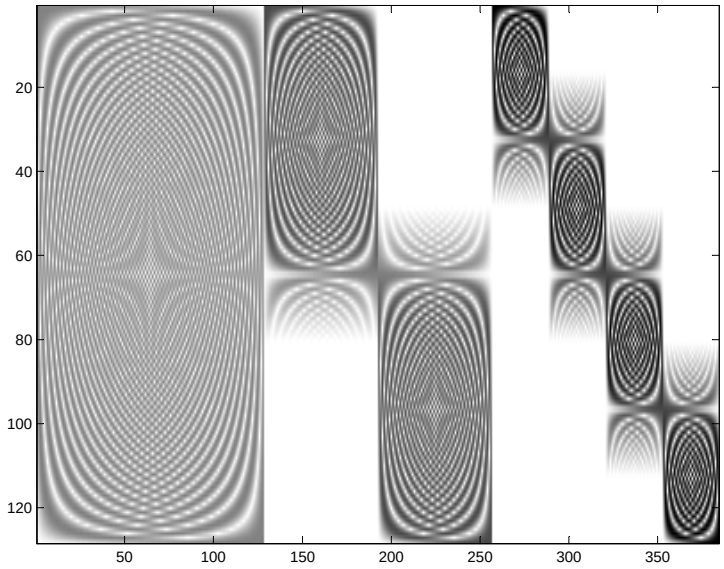

Figura 5.25: *Estructura del diccionario* $\mathbf{\Phi}$ *de la CPT para* $N = 128$ *y* $k = 3$.

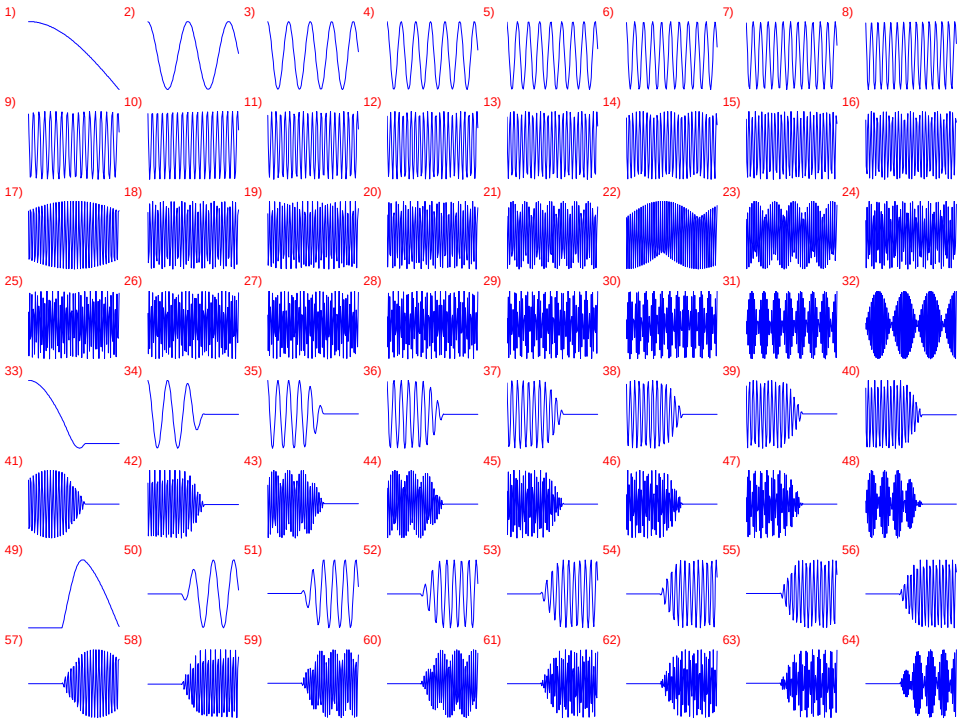

Figura 5.26: *Algunos átomos del diccionario* $\mathbf{\Phi}$ *de la CPT para* $N = 128$ *y* $k = 3$.

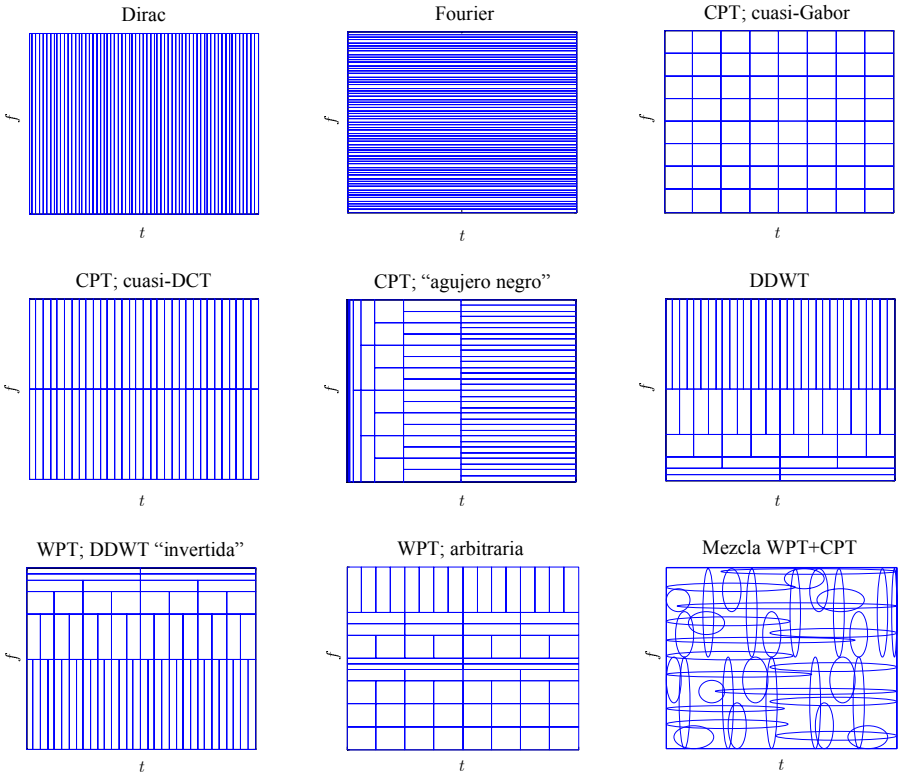

Figura 5.27: *Comparación entre las particiones $t-f$ generadas por diferentes bases y diccionarios $\mathbf{\Phi}$. Obsérvese el caso del diccionario formado por la mezcla de algunos átomos de WPT y CPT (abajo derecha). El diccionario así formado no resulta ortogonal debido a la superposición de las características $t-f$ de varios átomos. Se han reemplazado los rectángulos por elipses para hacer más evidentes las superposiciones. Además tampoco resulta completo porque no cubre todo el plano $t-f$. Para encontrar una representación en términos de este tipo de diccionarios se requieren consideraciones adicionales que serán presentadas en el capitulo siguiente.*

discutidos en el capítulo siguiente. Para algunos diccionarios especiales que pueden descomponerse en bases ortogonales, como WPT o CPT, se ha desarrollado el algoritmo de denominado *mejor base ortogonal* (BOB) [213]. Este algoritmo resulta útil en aplicaciones relacionadas con compresión ya que minimiza una función de "entropía" de la señal analizada. También se desarrolló la técnica denominada *base discriminante local* (LDB) para encontrar una base adecuada en problemas de clasificación de señales [177]. En relación con la técnica de ICA, discutida anteriormente, se ha implementado el algoritmo de la *base menos dependiente estadísticamente* (LSDB) [178]. Inclusive existe una versión más reciente que permite encontrar una base para lograr la representación más rala posible denominado *base de mejor dispersión* (BSB) [180].

Capítulo 6

Representaciones ralas y/o independientes

"¿Por qué no entendéis mi lenguaje? Porque no podéis oír mi palabra."

(Juan 8,43)

Contenido

6.1. Introducción

En el capítulo anterior se revisaron diferentes posibilidades para representar una señal mediante diccionarios discretos. Para el caso ortogonal existen una variedad de técnicas disponibles que permiten encontrar la representación de una señal discreta en términos de estos diccionarios. Estas técnicas resultan particularmente sencillas para el caso de las transformaciones unitarias debido, entre otros aspectos, a que la representación es única. Sin embargo para el caso no ortogonal existen muchas representaciones posibles de una señal mediante un mismo diccionario. En estos casos es posible seleccionar una representación adecuada en base a pautas o criterios adicionales. En el Capítulo 2 se discutieron algunos criterios posibles en el contexto de la modelización de señales. De esta discusión se desprende que dos criterios útiles para lograr una buena representación de la señal consisten en lograr que la misma resulte rala e independiente.

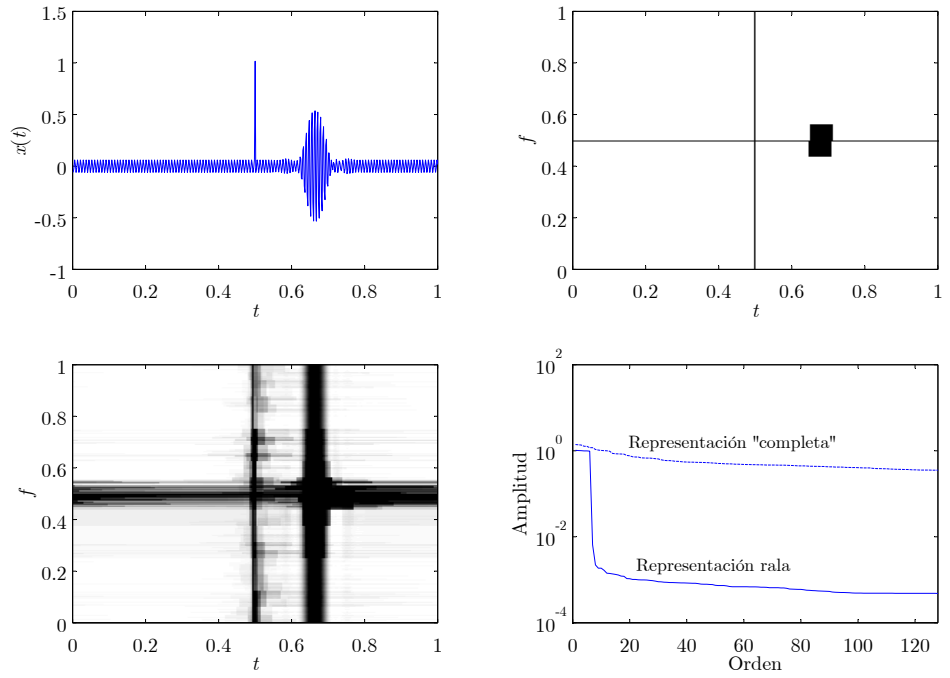

Figura 6.1: *Señal artificial $x(t)$ (denominada* carbon *en [21], arriba izquierda) y dos de sus posibles representaciones en función un diccionario prefijado, graficadas en el plano $t - f$: representación rala obtenida utilizando sólo los átomos más significativos (arriba derecha) y representación "completa" obtenida mediante las proyecciones de todos los átomos (abajo izquierda). El diccionario $\mathbf{\Phi}$ utilizado está formado por el árbol completo de funciones paquetes de onditas tipo Symlets con 8 momentos nulos (10x sobrecompleto). Obsérvese como la representación rala evidencia más claramente las tres componentes de la señal sintetizada. Se muestra también el gráfico de la evolución de los coeficientes de ambas representaciones ordenados en forma descendente por su amplitud (abajo derecha, adaptado de [21]).*

La Figura 6.1 ilustra la utilidad de lograr la representación rala de una señal. En esta figura aparecen dos posibles representaciones de una señal artificial en el plano $t - f$ mediante un diccionario sobrecompleto fijado de antemano. La señal artificial está formada por una componente senoidal, un delta de dirac y una "ondita" localizada adecuadamente en el plano $t - f$. Es posible observar como la representación rala es la que más se acerca a lo que podría considerarse el análisis ideal[1] para la señal original [21]. En este capítulo se retomará la temática de las representaciones ralas e independientes junto con los fundamentos de las técnicas disponibles para poder obtenerlas.

Como se mencionó, un código ralo es aquel que representa la información en términos de un número pequeño de descriptores tomados de un conjunto grande [145]. Ésto quiere decir que sólo una pequeña fracción de los elementos del código son usados activamente para representar un patrón típico. En términos numéricos, ésto significa que la mayoría de los elementos son cero, o "casi" nulos, la mayor parte del tiempo [71, 84]. La dificultad

[1] A veces se suele denominar a este análisis ideal directamente como "plano de síntesis" [21], o sea el que constituye la solución ideal del problema inverso al de la representación o análisis de la señal.

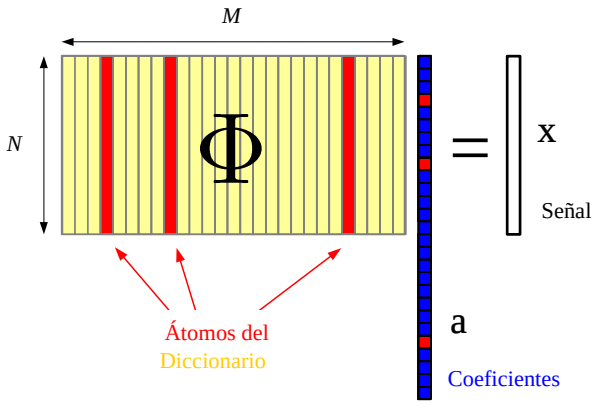

Figura 6.2: *Esquema que muestra la representación de una señal* $\mathbf{x}$ *discreta descripta en términos de un diccionario* $\mathbf{\Phi}$ *y un conjunto de coeficientes o pesos* $\mathbf{a}$ *(a partir de* $\mathbf{\Phi a} = \mathbf{x}$*). Aquí interesa el caso donde sólo unos pocos coeficientes son diferentes de cero (señalados con color rojo en el esquema).*

reside en definir de manera más precisa que significa "casi" nulos. La suposición implícita es que esos valores "cercanos a cero" pueden tratarse como si fueran exactamente cero, con una pequeña o ninguna pérdida de información útil.

Es posible definir varias medidas o normas que permitan cuantificar cuan rala es una representación (Ver Sección 2.4.2). Una forma alternativa de evaluar este tipo de representaciones es a través de su distribución de probabilidad. En general se trata de distribuciones con un valor de curtosis positivo grande. Ésto se traduce en que poseen un pico muy agudo en cero y colas largas a ambos lados. Un ejemplo es el caso de la distribución laplaciana, pero también pueden utilizarse otras distribuciones únicas o mixtas.

Para comprender mejor la importancia de obtener una representación rala de una señal se utilizará una analogía con el idioma inglés debida a Stefan Mallat, que proporciona la intuición correcta acerca de este tipo de representaciones [153]. Supóngase que se desean describir diferentes ideas utilizando un diccionario pequeño de sólo tres mil palabras inglesas. Entonces la descripción de la mayoría de los conceptos requeriría frases largas que usen todas o la mayoría de las tres mil palabras. Sin embargo, si se describieran estas mismas ideas utilizando un diccionario grande de cien mil palabras, sólo sería necesario ocupar un número pequeño de palabras para la mayoría de los conceptos.

Se ilustrará esta última analogía volviendo al caso de una señal $\mathbf{x}$ descripta en términos de un diccionario $\mathbf{\Phi}$ y un conjunto de coeficientes o pesos $\mathbf{a}$ como en la Figura 6.2. Para ello se muestra, en la Figura 6.3, la representación de una señal artificial sencilla mediante la base de Fourier y un diccionario sobrecompleto de paquetes de onditas. Obsérvese como, en el primer caso, no existe ningún elemento nulo en la representación, mientras que en el segundo la representación lograda posee muy pocos elementos diferentes de cero. Ésto tiene que ver también con el método utilizado para encontrar los coeficientes, el que será tratado con mayor detalle en la secciones siguientes.

Los métodos que permiten obtener una representación rala de una señal, dado un diccionario adecuado, suelen denominarse métodos de *selección de subconjuntos* [198].

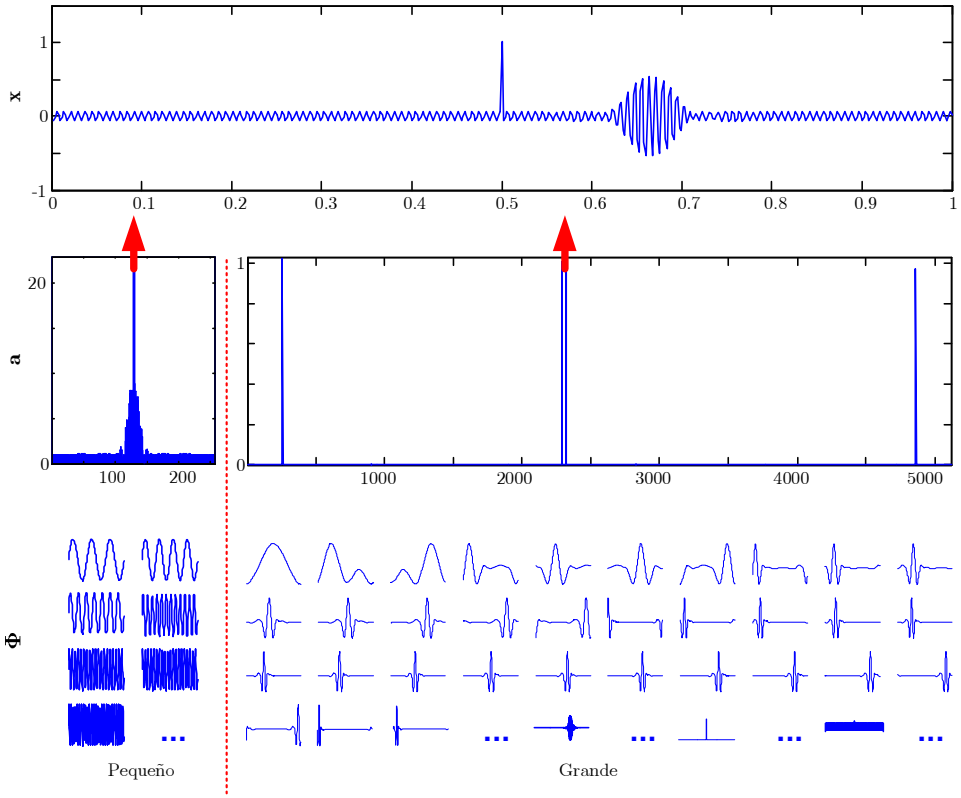

Figura 6.3: *Señal artificial* **x** *(como la de la Figura 6.1, arriba) y su representación en términos de los coeficientes* **a** *(centro) en función de dos diccionarios* **Φ***: uno "pequeño" formado por una base ortogonal de Fourier (se muestra la magnitud de los coeficientes complejos, izquierda) y uno "grande" formado por el árbol completo de funciones paquetes de onditas tipo Symlets con 8 momentos nulos (10 veces sobrecompleto, derecha). Se muestran también, a título ilustrativo, algunos elementos de cada uno de los diccionarios (abajo).*

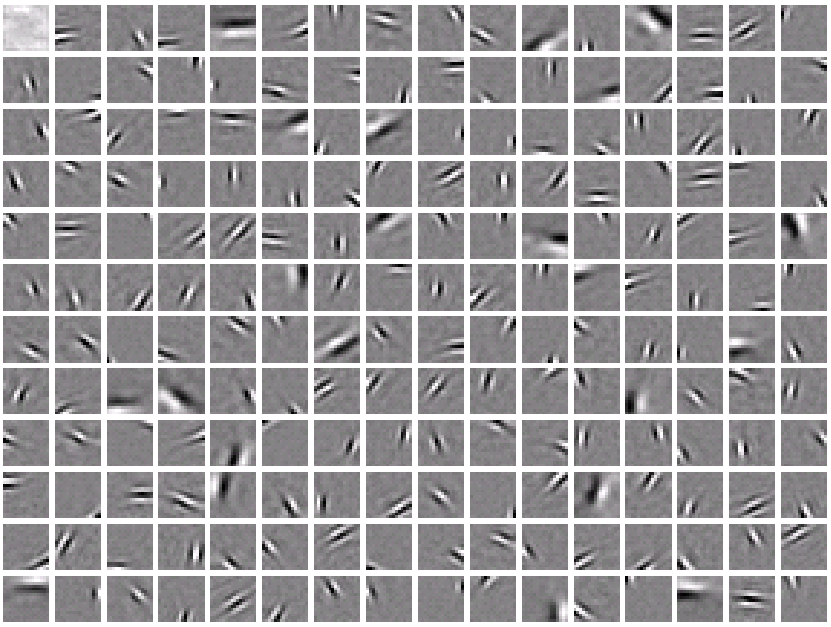

Figura 6.4: *Diccionario óptimo en términos de dispersión e independencia obtenido a partir de datos de imágenes tomados de la naturaleza. El método se basa en un modelo sensorial del análisis realizado a nivel de la corteza visual (tomado de [144]).*

Entre éstos pueden citarse, para el caso determinístico: el método de *búsqueda de bases* (BP) [21], el de *búsqueda por coincidencia* (MP) [128], o el de la *mejor base ortogonal* (BOB) [26]. El *método de marcos* (MOF) [33] se ha empleado a veces para comparación aunque estrictamente no resulta en una representación rala. También existen métodos derivados de un enfoque más estadístico para hallar los coeficientes, cuyos resultados pueden asimilarse como equivalentes a los del caso determinístico. Además estos métodos permiten encontrar también el diccionario óptimo [145]. En esta dirección existen trabajos orientados al análisis de imágenes "naturales" basados en modelos sensoriales [144, 147]. En la Figura 6.4 es posible apreciar el diccionario óptimo obtenido mediante estas técnicas a partir de imágenes de la naturaleza como las mostradas en la Figura 6.5. Más recientemente han aparecido trabajos similares pero dirigidos a señales sonoras de audio, música y sonidos naturales [1, 114]. El diccionario óptimo también puede ser encontrado por métodos determinísticos, aunque este enfoque no se ha explotado aún suficientemente.

En el enfoque estadístico la representación o codificación rala mediante diccionarios sobrecompletos también posee importantes relaciones con la técnica de *análisis de componentes independientes* (ICA) [86, 90, 112, 111]. Esta familia de procedimientos, cuyos fundamentos fueron descriptos en el Capítulo 2, maximiza la independencia estadística entre los coeficientes de la representación. Estas técnicas se han aplicado con bastante éxito al campo de las señales biomédicas en general [123, 124, 93], a problemas como el de la *deconvolución ciega*. En la Figura 6.6 se muestra un ejemplo de los resultados de un algoritmo ICA utilizado para separar tres señales de voz que han sido mezcladas en forma artificial [111]. Últimamente se han aplicado también al campo de la clasificación

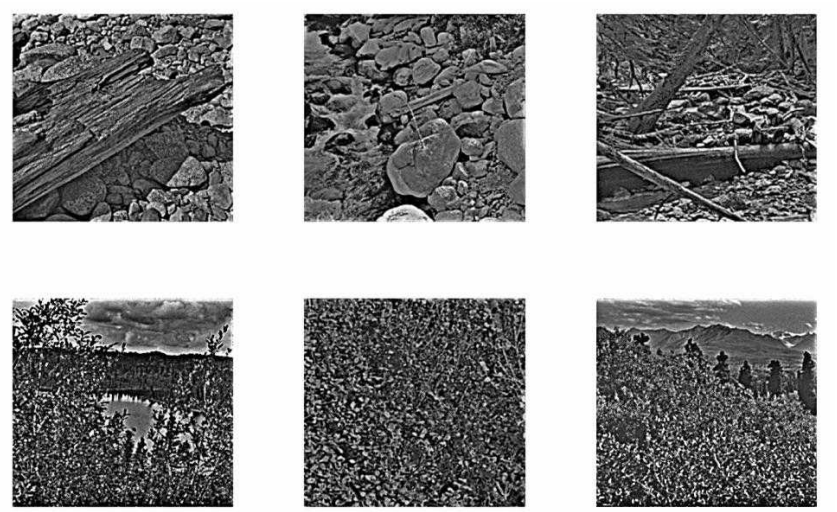

Figura 6.5: *Imágenes de escenas naturales utilizadas para generar el diccionario óptimo de la Figura 6.4. Generalmente a estas imágenes se las "blanquea" previamente para eliminar la estadística de segundo orden, de manera de facilitar la tarea del método de búsqueda del diccionario.*

de imágenes [113] y al ASR [110].

Éste capítulo se organiza de la siguiente forma. En la Sección 6.2 se describen con más detalle las ventajas de una representación rala, entre las que se pueden citar: robustez al ruido, mayor separabilidad, óptima generalización, eficiencia en la codificación de la información de la señal y mejor resolución de eventos. En la Sección 6.3 se presenta el planteo general del problema de la representación de una señal en este contexto. Este problema puede ser divido en dos partes: selección de los coeficientes (Sección 6.4) y búsqueda del diccionario óptimo (Sección 6.5). En cada caso se diferencian los enfoques basados en criterios determinísticos y estadísticos, tratando de rescatar las equivalencias entre ambos.

6.2. Ventajas y desventajas

Entre las ventajas de lograr una representación rala se pueden mencionar sus propiedades de superresolución [21], tanto en el tiempo como en la frecuencia. Ésto le permite comportarse como una transformación adaptable al tipo de señales a analizar, de manera de encontrar la representación óptima siempre que el diccionario esté diseñado adecuadamente (Ver nuevamente Figura 6.1).

Existen varias maneras para aprovechar la robustez intrínseca de este tipo de representaciones al ruido poco correlacionado con los elementos del diccionario. Se pueden establecer analogías con las técnicas de *limpieza de ruido* (en inglés *denoising*) apoyadas en la búsqueda de umbrales óptimos para la reconstrucción, como el caso de limpieza mediante onditas o paquetes de onditas [127, 83]. Es posible incluir el tratamiento del ruido en el cálculo de los coeficientes mediante esquemas determinísticos basados en

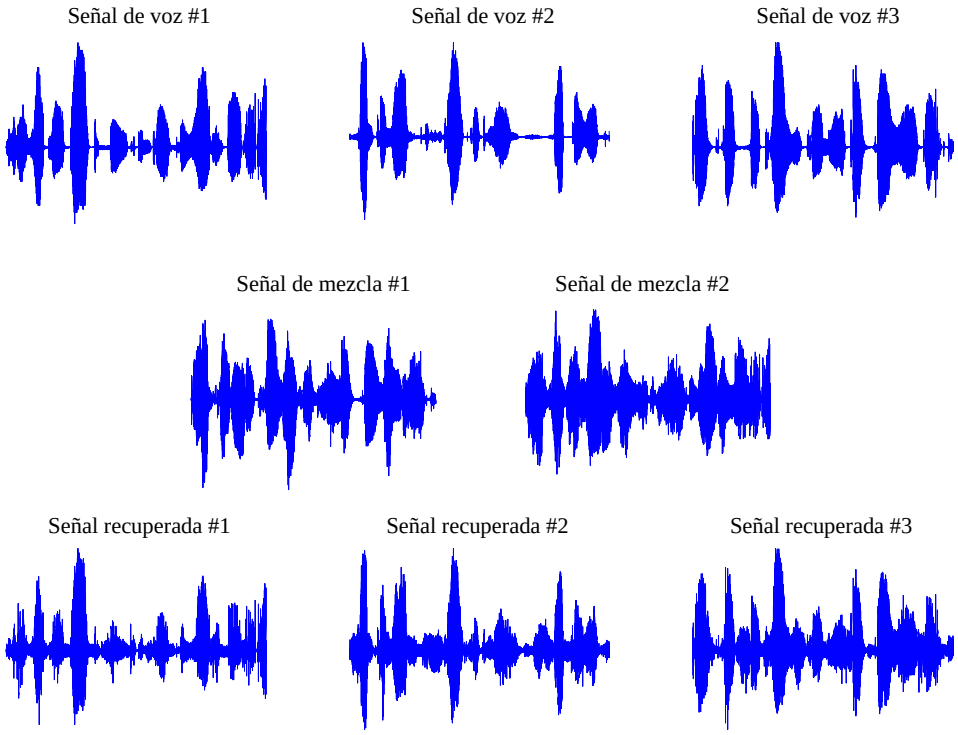

Figura 6.6: *Demostración del resultado de la separación ciega de tres señales de voz a partir de dos mezclas: señales originales (arriba), mezclas (centro) y señales restauradas (abajo, tomado de [111]). Las señales originales fueron obtenidas de la base de datos TIMIT.*

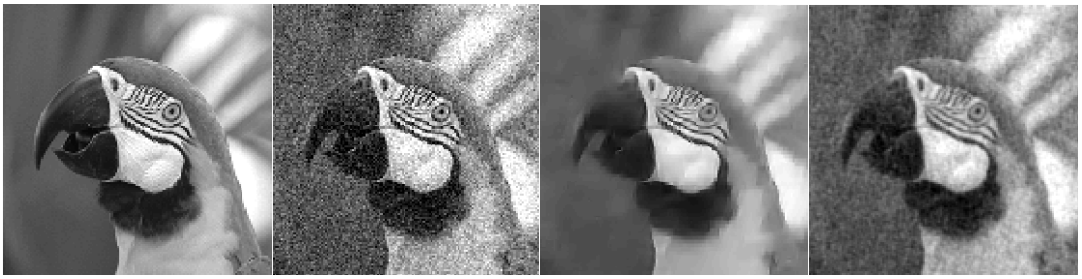

Figura 6.7: *Experimento de limpieza de ruido en imágenes (de izquierda a derecha): imagen original, con ruido de varianza unitaria, limpiada mediante reducción del código ralo y limpiada mediante el clásico filtro de Wiener para comparación (adaptado de [84]).*

teoría de la regularización (RT) [57]. También se pueden encontrar formas de combatir el ruido en conexión con el enfoque probabilista y el denominado *ICA con ruido* [113]. Este enfoque provee una visión alternativa que permite frasear los problemas en una forma equivalente a la encontrada mediante RT. Además, el enfoque basado en ICA aporta un esquema teórico más general, que incluye el caso particular de algunas representaciones ralas. En la Figura 6.7 es posible apreciar los resultados de un experimento de limpieza de ruido mediante la técnica denominada *contracción del código ralo* (en inglés *sparse code shrinking*) para el caso de una imagen [84].

Se ha demostrado también que cuanto más rala es una representación, también posee mejores características para la predicción y generalización a partir de ella [57]. El resultado sugiere que la representación rala de una señal en términos de un diccionario grande de rasgos es óptima para la generalización[2]. En este sentido se han establecido relaciones con las *máquinas de soporte vectorial* (SVMs). Éstas constituyen un método de aprendizaje artificial orientado a la clasificación o la regresión desarrollado por Vapnik [208], donde muchos de los parámetros son iguales a cero. También se puede esperar que si el diccionario se selecciona adecuadamente para que realice un análisis en términos de una serie de características bien discriminativas de las clases a analizar entonces, a pesar de que el espacio en el que se realiza la clasificación posea muchas dimensiones, los datos pueden ser más fácilmente separables en clases. Ésto último estaría indicando que se requeriría un clasificador más sencillo para realizar la tarea en este nuevo espacio de mayores dimensiones, lo cual es compatible de nuevo con las ideas detrás de SVM. En el caso de la señal de voz, dado que ésta posee características dinámicas importantes, es también posible capturar parte de esta dinámica si se utilizan ventanas de tiempo grandes para generar los átomos del diccionario. Ésto permite utilizar posteriormente un clasificador estático o alguno con una dinámica sencilla.

Existen interesantes relaciones entre la dispersión y la independencia estadística de los coeficientes de la representación. Si bien éstos son criterios diferentes, ambos resultan útiles en el contexto de modelización de señales y pueden dar lugar a representaciones similares. Inclusive algunos métodos particulares los aplican simultáneamente para obtener representaciones ralas y factoriales [115]. En estos casos existen evidencias de que los

[2]Para que ésto pueda cumplirse realmente es necesario contar con suficientes ejemplos de los patrones de activación característicos para las diferentes señales a analizar.

códigos generados poseen baja entropía, lo que los convierte en óptimos desde el punto de vista de la teoría de información [71].

Una representación rala puede también obtenerse a partir de una red neuronal, lo que permite algunas interpretaciones más biológicas de este mecanismo [145, 71]. Además, como ya se ha mencionado, se ha demostrado su utilización como esquema de codificación eficiente a nivel de los sistemas sensoriales biológicos. Se puede decir que el propio código neuronal, según se describió en el Capítulo 3 también resulta sumamente ralo. Ésto permite también la codificación más directa de las representaciones ralas obtenidas en términos de trenes de pulsos para su incorporación en sistemas de cómputo biológicamente inspirados[3].

Como posibles desventajas se pueden citar la gran cantidad de dimensiones de las representaciones logradas, en comparación con las representaciones tradicionales. Sin embargo ésta es una desventaja relativa también a su tratamiento mediante los paradigmas tradicionales. Otra es el mayor tiempo de cálculo necesario para obtener la representación. En este sentido se debe recalcar nuevamente que el objetivo de la presentación de estas técnicas no es el de reemplazar al enfoque más convencional, si no más bien proveer herramientas alternativas para cuando este enfoque no permite resolver adecuadamente los problemas que se presentan en las aplicaciones.

6.3. Planteo del problema

A continuación se formaliza el planteo del problema acerca de cómo encontrar una representación rala de una señal, comenzando con el caso más general de representación de una señal mediante un diccionario.

Sea $\mathbf{x} \in \mathbb{R}^N$ una señal a la cual se la quiere representar en términos de un diccionario $\mathbf{\Phi}$, de tamaño $N \times M$, y un conjunto de coeficientes $\mathbf{a} \in \mathbb{R}^M$. De este modo la expresión que describe a la señal es la siguiente:

$$\mathbf{x} = \sum_{\gamma \in \Gamma} \boldsymbol{\phi}_\gamma a_\gamma + \boldsymbol{\varepsilon} = \mathbf{\Phi}\mathbf{a} + \boldsymbol{\varepsilon}, \tag{6.1}$$

donde $\boldsymbol{\varepsilon} \in \mathbb{R}^N$ constituye el término de ruido de aditivo y $M \geq N$. El diccionario $\mathbf{\Phi}$ resulta en una colección de formas de onda o funciones parametrizadas $(\boldsymbol{\phi}_\gamma)_{\gamma \in \Gamma}$ (Definición 2.10), donde cada forma de onda $\boldsymbol{\phi}_\gamma$ constituye un átomo.

Aunque la apariencia de la ecuación (6.1) resulta sencilla, el principal problema consiste en que para el caso más general $\mathbf{\Phi}$, $\mathbf{a}$ y $\boldsymbol{\varepsilon}$ son desconocidos, existiendo infinitas soluciones. Aún en el caso sin ruido ($\boldsymbol{\varepsilon} = \mathbf{0}$) y conociendo $\mathbf{\Phi}$ de antemano, si los átomos son más que la cantidad de muestras de $\mathbf{x}$ o si no forman una base, ésto produce representaciones no únicas de la señal. Por lo tanto se debe encontrar un criterio que permita seleccionar alguna de ellas. En este caso y, a pesar de que la ecuación es lineal, los coeficientes que se eligen para formar parte de la solución resultan en general de una función no lineal de los datos $\mathbf{x}$. Se han propuesto diferentes métodos para obtener una

[3]Un ejemplo de ello es la clasificación de los patrones utilizando *redes neuronales pulsadas* o con sinápsis dinámicas [92].

descomposición de este tipo como BP, MP, BOB o incluso MOF. Para el caso completo y sin ruido la relación entre los datos y los coeficientes resulta lineal y está dada por $\mathbf{\Phi}^{-1}$. Para las transformaciones tradicionales como la DFT esta inversión se simplifica debido a que $\mathbf{\Phi}^{-1} = \mathbf{\Phi}^T$ (Ver Sección 5.2).

Por lo discutido hasta aquí un criterio de interés para seleccionar una representación, de entre todas las factibles, consiste en que ésta sea lo más rala posible (y muchas veces también la más "independiente"). Nuevamente, ésto significa que se espera que sólo unos pocos coeficientes, a_γ en (6.1), sean diferentes de cero. Existen varios criterios posibles para medir cuán rala resulta una representación, según se vió en la Sección 2.4.2, aunque generalmente el que más se utiliza es el de la norma ℓ_0.

Definición 6.1 *Sea $\mathcal{C}(\mathbf{a}, \mathbf{\Phi}|\mathbf{x})$ una función criterio con valores escalares que da cuenta de la dispersión de los coeficientes* $\mathbf{a}$*. Entonces, a partir de (6.1), es posible definir el problema de la representación rala de* $\mathbf{x}$ *con respecto a $\mathcal{C}(\mathbf{a}, \mathbf{\Phi}|\mathbf{x})$ como*[4]*:*

$$\left[\hat{\mathbf{a}}, \hat{\mathbf{\Phi}} \right] = \arg\min_{\mathbf{a}, \mathbf{\Phi}} \mathcal{C}(\mathbf{a}, \mathbf{\Phi}|\mathbf{x}) \quad \text{sujeto a} \quad \mathbf{\Phi a} + \boldsymbol{\varepsilon} = \mathbf{x}. \tag{6.2}$$

El problema planteado en la Definición 6.1 se podría dividir en dos partes o subproblemas, que en lenguaje cotidiano podrían plantearse de la siguiente forma:

1. ¿Cómo encontrar la menor cantidad de coeficientes que representen mejor a la señal original, eliminando también los efectos del ruido?

2. ¿Cómo construir el diccionario que mejor describa el tipo de señales a analizar?.

Utilizando la terminología utilizada en el modelado estadístico de los sistemas de percepción biológicos (Ver Sección 2.4), a estos problemas se los denomina el de *inferencia* y el de *aprendizaje* respectivamente [117]. Éste último suele ser el más complejo (y más demandante de recursos). Es posible también incluir restricciones adicionales a fin de disminuir la cantidad de soluciones posibles del problema, que resulten a su vez útiles para en la aplicación final. Ésto implica concebir un método que permita encontrar una solución adecuada en base a la formulación original y a las nuevas restricciones planteadas.

Formalmente el problema de la inferencia, siempre a partir de (6.1), se puede expresar como una minimización (o maximización según corresponda) de una función $\mathcal{F}$ que incluya el o los criterios considerados:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \left\{ \mathcal{F}\left(\mathbf{a}|\mathbf{x}, \mathbf{\Phi}\right) \right\}, \tag{6.3}$$

donde $\mathbf{x}$ y $\mathbf{\Phi}$ están dados de antemano.

[4]Dependiendo del criterio seleccionado es posible también definir este problema como uno de maximización.

En forma análoga es posible expresar el problema del aprendizaje como:

$$\hat{\mathbf{\Phi}} = \underset{\mathbf{\Phi}, a_1, \cdots, a_M}{\arg\min} \left\{ \mathcal{G}\left(\mathbf{\Phi}, \mathbf{a}|\mathbf{x}\right) \right\} \tag{6.4}$$

donde $\mathcal{G}$ también es una función que incluye a los criterios considerados y puede ser igual a $\mathcal{F}$.

Existen diferentes enfoques para plantear las funciones $\mathcal{F}$ y $\mathcal{G}$ de manera de obtener una solución de (6.3) y (6.4). Desde el punto de vista determinístico las restricciones aparecen en forma de la minimización de ciertas medidas, distancias o costos (Ver Sección 2.4.2) y la solución en forma de un problema de optimización o mediante teoría de regularización. Con el enfoque estadístico las restricciones aparecen sobre el tipo de funciones de densidad de probabilidad de los coeficientes y el ruido (Ver Sección 2.4.2), y la solución se obtiene nuevamente como un problema de optimización, maximizando alguna *verosimilitud* o *probabilidad posterior*. Para un análisis más detallado de las conexiones entre el enfoque probabilista y el determinista puede consultarse [106]. Las mismas medidas empleadas en la búsqueda de una solución permiten también valorar la eficacia de la representación lograda, es decir la "bondad" de la codificación de los datos $\mathbf{x}$ en términos de los coeficientes $\mathbf{a}$, mediante (6.1).

Para tener una idea gráfica del problema es posible utilizar la ecuación (6.1) como un modelo generativo $\mathcal{M}$ que permita sintetizar señales artificiales sencillas. Para ello se han elegido señales aleatorias $\mathbf{x} \in \mathbb{R}^2$, representadas mediante un diccionario de tres átomos ($\mathbf{\Phi} \in \mathbb{R}^{2\times 3}$). En la Figura 6.8 se muestran los resultados obtenidos para una corrida de esta simulación, junto con los vectores del diccionario original $\mathbf{\Phi}$ y del estimado $\hat{\mathbf{\Phi}}$ mediante el método planteado en [117]. Este método puede verse como una versión particular para solucionar el problema de ICA sobrecompleto y con ruido, para el cual se supone que los coeficientes son estadísticamente independientes y poseen una función de distribución de probabilidad a priori de tipo laplaciana. El método maximiza la verosimilitud de los datos dado el modelo para la base que posea más información en las direcciones predominantes de los datos y será descripto con mayor detalle en las secciones siguientes. Como puede observarse, y al contrario que PCA, éste puede encontrar direcciones no ortogonales con la única restricción de que los coeficientes deben tener distribuciones supergaussianas. La utilización de una laplaciana permite encontrar una solución que consiste en minimizar la norma ℓ_1 del vector de coeficientes. Sin embargo, si los elementos del diccionario son muy parecidos puede ser necesario utilizar distribuciones aún más ralas que la laplaciana. Ésto puede entenderse fácilmente si se imagina que los vectores del diccionario original en la Figura 6.8 se encuentran más cercanos. Entonces será más difícil extraer la dirección preponderante a partir de la estadística de los datos. Si por otra parte esta estadística fuera de segundo orden también es claro un fenómeno similar, y por eso la necesidad de utilizar información proveniente de ordenes superiores a dos. Además es posible observar como el método separa completamente los datos, esta operación es equivalente a una decorrelación pero de orden superior.

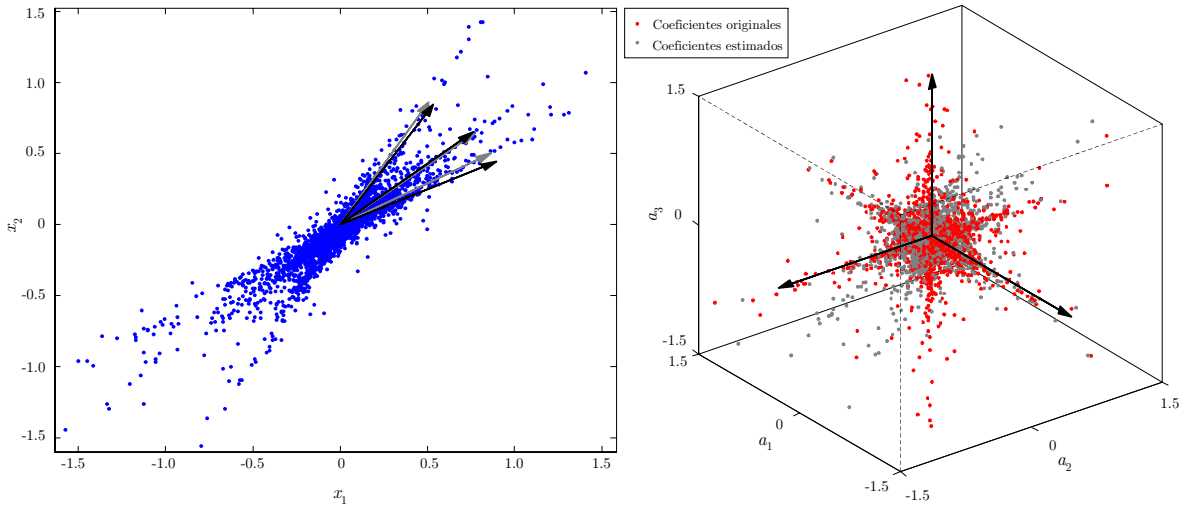

Figura 6.8: *Valores* $\mathbf{x}$ *obtenidos a partir del modelo generativo* $\mathcal{M}$ *(6.1) en dos dimensiones (izquierda), donde los coeficientes* $\mathbf{a}$ *tienen una distribución supergaussiana (más rala que la laplaciana) y el ruido* $\boldsymbol{\varepsilon}$ *posee una distribución gaussiana. Los vectores del diccionario o columnas de la matriz* $\boldsymbol{\Phi}$ *se muestran en gris y el diccionario estimado por el método descripto en [117] en color negro. Valores originales de los coeficientes* $\mathbf{a}$ *utilizados para generar los datos* $\mathbf{x}$ *en un espacio tridimensional y valores estimados por el mismo método (derecha).*

6.4. Selección de coeficientes o inferencia

En esta sección se plantean algunas soluciones posibles al problema de selección de coeficientes o inferencia, suponiendo el diccionario $\boldsymbol{\Phi}$ conocido de antemano. Primero se trata el caso limpio ($\boldsymbol{\varepsilon} = \mathbf{0}$) y posteriormente el caso ruidoso ($\boldsymbol{\varepsilon} \neq \mathbf{0}$). A estos métodos se los denominó anteriormente como métodos de aproximación (Sección 4.4.2).

6.4.1. Caso limpio, enfoque determinístico

Dado un diccionario $\boldsymbol{\Phi}$, el caso ideal consistiría en poder solucionar el problema de la representación rala de $\mathbf{x}$ en términos de (6.1) con respecto a la medida "real" de dispersión ℓ_0, es decir el número total de elementos iguales a cero. De esta forma el problema sería:

$$\min_{\mathbf{a}} \|\mathbf{a}\|_0 \quad \text{sujeto a} \quad \boldsymbol{\Phi}\mathbf{a} = \mathbf{x}. \tag{6.5}$$

Desafortunadamente este problema de optimización constituye uno de *programación entera* bastante difícil de resolver[5], y su costo computacional resulta actualmente prohibitivo para muchas aplicaciones prácticas [57]. Por ello se han planteado varias alternativas que se revisaran a continuación.

[5]De hecho se trata de un problema *NP-completo*.

Búsqueda de bases

En [21], Chen y colaboradores proponen un método, denominado BP el cual se diseñó para producir una representación rala. Ellos frasearon el problema de hallar una representación conveniente como uno de optimización con respecto a la norma ℓ_1. Más precisamente, si como en (6.1) la señal $\mathbf{x}$ tiene longitud N y existen M formas de onda en el diccionario, entonces el problema para resolver es:

$$\min_{\mathbf{a}} \|\mathbf{a}\|_1 \quad \text{sujeto a} \quad \mathbf{\Phi a} = \mathbf{x}, \tag{6.6}$$

donde $\mathbf{a}$ un es un vector en $\mathbb{R}^M$ que representa los coeficientes y $\mathbf{\Phi}$ es una matriz de $N \times M$ que da los valores de las M formas de onda en el diccionario.

Éste problema puede convertirse en uno de *programación lineal* tradicional (con coeficientes sólo positivos) haciendo la substitución $\mathbf{a} \leftarrow [\mathbf{u}, \mathbf{v}]$ y resolviendo (c.f. [21]):

$$\min_{\mathbf{u},\mathbf{v}} \mathbf{1}^{\mathrm{T}}[\mathbf{u}, \mathbf{v}] \quad \text{sujeto a} \quad [\mathbf{\Phi}, -\mathbf{\Phi}]\,[\mathbf{u}, \mathbf{v}] = \mathbf{x}, \text{ con } \mathbf{0} \leq \mathbf{u}, \mathbf{v} \tag{6.7}$$

Ésta formulación puede resolverse eficaz y exactamente con los métodos de punto interior de la programación lineal. El orden de la cantidad de operaciones necesarias para resolver este problema crece de forma cuasi-lineal con la cantidad de átomos[6] como $C\,\mathcal{O}(M \log_2(M))$, donde la constante C depende de la exactitud con la que se resuelve (6.6).

Una desventaja de BP es que en realidad ℓ_1 es sólo una aproximación a ℓ_0, pero resulta mucho más fácil de resolver que su contraparte. Además, si $\mathbf{x}$ puede sintetizarse a partir de muy pocos elementos, BP recobra perfectamente los átomos y los coeficientes específicos utilizados en la síntesis. En la Figura 6.9 se muestra una representación geométrica en $\mathbb{R}^2$ de las soluciones de:

$$\min_{a_1,a_2} \{|a_1|^q + |a_2|^q\} \quad \text{sujeto a} \quad x = \phi_1 a_1 + \phi_2 a_2, \tag{6.8}$$

para $0 \leq q < 1$ y $q = 1$ respectivamente. Resulta sencillo apreciar como ambas soluciones pueden coincidir cuando alguno de los coeficientes es igual a cero.

En la Figura 6.10 se puede apreciar el resultado de realizar el análisis de una señal artificial mediante BP con los diccionarios WPT y CPT (ambos 10 veces sobrecompletos o con profundidad 10).

Chen y colaboradores dan varios ejemplos con señales artificiales que muestran los beneficios de su método, en términos de dispersión y super-resolución, comparados a las representaciones correspondientes encontradas por MOF, MP y BOB. Sin embargo no parece haberse realizado un estudio sistemático de la técnica de BP aplicada a datos del mundo real. A continuación se presentarán brevemente otros métodos que permiten encontrar una representación de las señales en términos de (6.1) a los fines de su comparación.

[6] El costo computacional depende también del diccionario utilizado [20]. En lo que sigue se supone que se trata de diccionarios que proveen métodos rápidos implícitos para el cálculo de $\mathbf{\Phi a}$, como la mayoría de los analizados en el Capítulo 5.

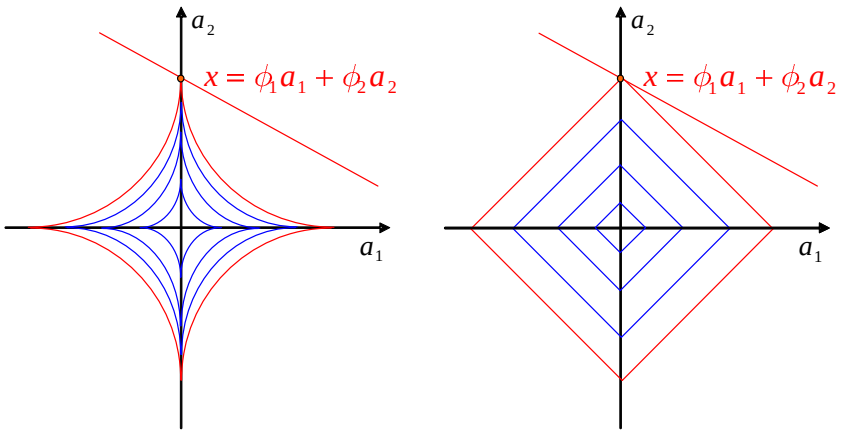

Figura 6.9: *Representación geométrica de la soluciones de (6.8) para* $0 \leq q < 1$ *(izquierda) y* $q = 1$ *(derecha). Obsérvese como ambas soluciones pueden coincidir cuando existen coeficientes iguales a cero. Ésto significa que si* $\mathbf{x}$ *puede sintetizarse a partir de muy pocos elementos, la minimización con respecto a* ℓ_1 *(BP) coincide con la de* ℓ_0 *(ideal).*

Búsqueda por coincidencia

En 1993 Mallat y Zhang [128] presentaron un método general para aproximar la descomposición (6.1) que encara el tema de la dispersión directamente. Comenzando a partir de una aproximación inicial $\mathbf{x}^{(0)} = \mathbf{0}$ y un residuo $\mathbf{R}^{(0)} = \mathbf{x}$, construye una secuencia de aproximaciones ralas paso a paso. En la etapa k se identifica el átomo $\boldsymbol{\phi}_\gamma^{(k)}$ que mejor se correlaciona con el residuo y luego se suma a la aproximación actual un múltiplo escalar de este átomo:

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + a_\gamma^{(k)} \boldsymbol{\phi}_\gamma^{(k)}, \tag{6.9}$$

donde $a_\gamma^{(k)} = \langle \mathbf{R}^{(k-1)}, \boldsymbol{\phi}_\gamma^{(k)} \rangle$, y $\mathbf{R}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$. Luego de m pasos, se obtiene una representación de la forma (6.1), con residuo $\mathbf{R} = \mathbf{R}^{(m)}$. Cabe mencionar que el valor de m marca una cota inferior a la dispersión de la representación obtenida.

Se puede decir que MP constituye una solución *voraz*[7] para encontrar una representación rala por pasos en términos de (6.1) [44]:

$$\min_{\mathbf{a}} \|\mathbf{x} - \boldsymbol{\Phi}\mathbf{a}\|_2 . \tag{6.10}$$

Es por ello que posee los inconvenientes y también las ventajas de este tipo de métodos de optimización. Entre las ventajas puede decirse que si $\mathbf{x}$ puede sintetizarse a partir de muy pocos elementos, entonces MP puede encontrar la solución. Sin embargo en general encuentra soluciones aproximadas y no exactas, aunque resulta para estos casos más rápido que BP. El orden de la cantidad de operaciones necesarias para este algoritmo

[7]Los algoritmos voraces (en inglés *greddy*) son aquellos que siguen el método de resolución metaheurístico que consiste en realizar elecciones localmente óptimas en cada paso, con la esperanza de encontrar el óptimo global. Raramente encuentran este óptimo, pero resultan rápidos y generalmente devuelven buenas aproximaciones.

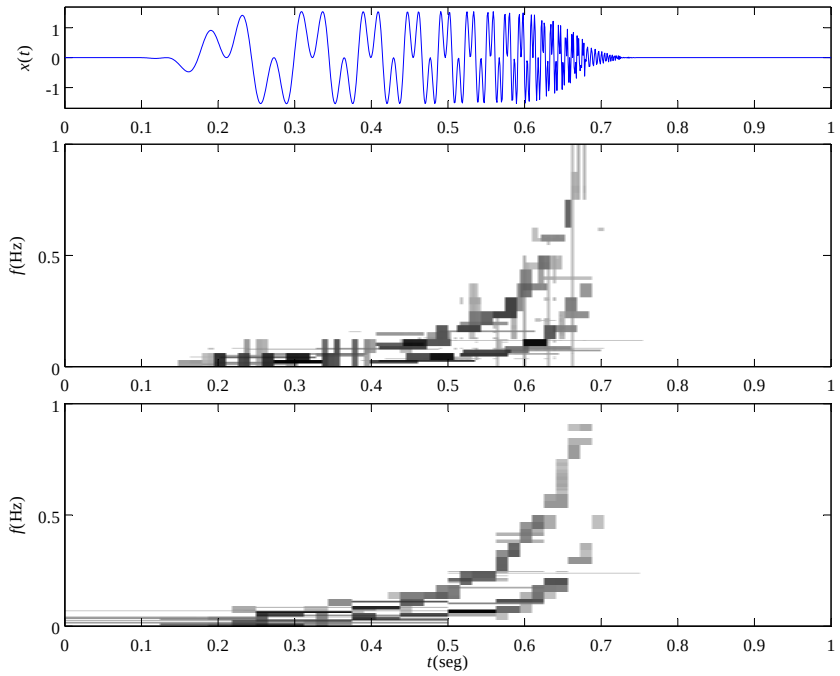

Figura 6.10: *Señal $x(t)$ compuesta de dos tonos modulados en frecuencia o "chirps" (arriba), representación en el plano $t-f$ de $x(t)$ calculada a partir de BP para el caso de un diccionario WPT (centro). Idem anterior pero para un diccionario CPT (abajo). Obsérvese la diferencia entre los átomos seleccionados para ambos diccionarios. Las frecuencias de ambos tonos pueden seguirse adecuadamente en esta representación, especialmente para el caso de CPT.*

crece también de forma cuasi-lineal con la cantidad de átomos como $C\,\mathcal{O}(M\log_2(M))$, pero la constante C depende ahora de la cantidad de pasos m utilizados [20].

En la Figura 6.11 se puede apreciar el resultado de realizar el análisis de una señal artificial mediante MP con los diccionarios WPT y CPT.

Mejor base ortogonal

Para algunos diccionarios, es posible desarrollar esquemas de descomposición específicos. Los diccionarios tipo WPT y CPT son ejemplos de ello, ya que poseen propiedades muy particulares (que se han descripto en el Capítulo 5). Algunas subcolecciones especiales de elementos en estos diccionarios son bases ortogonales. De esta forma se obtiene un amplio rango de posibles bases ortonormales. Coifman y Wickerhauser [26] propusieron un método para seleccionar una sola base ortogonal en forma adaptativa de entre todas estas bases ortogonales. Esta base resulta la "mejor base" en el sentido de una función de costo, generalmente dada como una función de "entropía"[8] de la energía normalizada de los coeficientes $\mathbf{a}$.

Para este caso se consideran diccionarios $\mathbf{\Phi}$ que resultan de la unión de bases orto-

[8]Ésta entropía es diferente a la relacionada con la función de distribución de probabilidad de los coeficientes tratada en el Capítulo 2 [127].

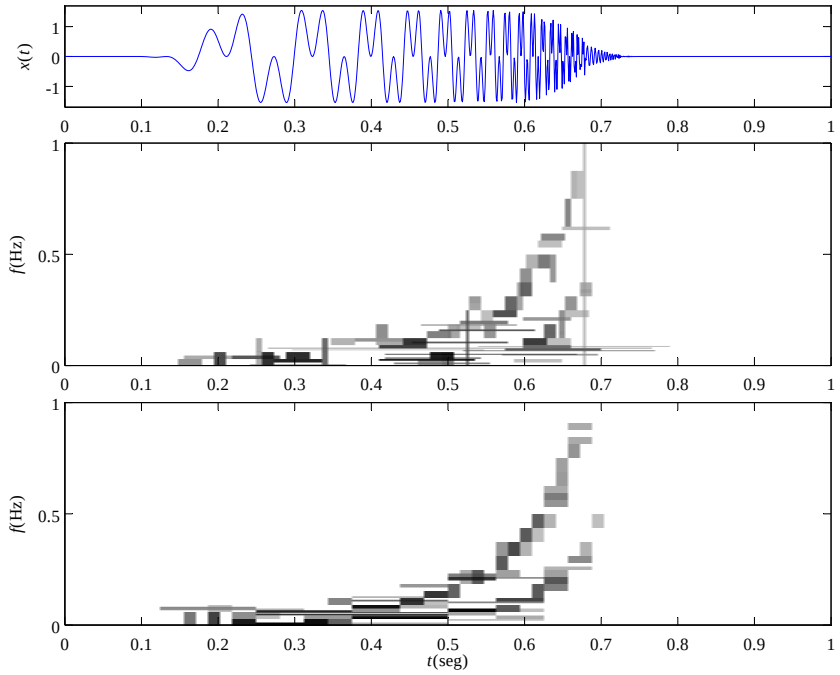

Figura 6.11: *Señal $x(t)$ compuesta como la de la Figura 6.10 (arriba), representación en el plano $t-f$ de $x(t)$ calculada a partir de MP para el caso de un diccionario WPT (centro). Idem anterior pero para un diccionario CPT (abajo). La cantidad máxima de átomos a utilizar se ha fijado de antemano para incrementar la dispersión. El resultado es una representación similar a la de la Figura 6.10.*

gonales en un espacio de señales de dimensión N [127]:

$$\mathbf{\Phi} = \bigcup_{\gamma \in \Gamma} \mathcal{B}_\gamma, \tag{6.11}$$

donde cada base ortogonal es una familia de N vectores:

$$\mathcal{B}_\gamma = \left\{ \boldsymbol{\phi}_\gamma^m \right\}_{1 \leq m \leq N}. \tag{6.12}$$

Si $\mathbf{a} = [\mathbf{x}\ \mathcal{B}_\gamma]_\gamma = a_\gamma$ denota el vector de coeficientes $\mathbf{a}$ de la señal $\mathbf{x}$ en términos de la base ortogonal $\mathcal{B}_\gamma$, y se define la función de costo como [21]:

$$\mathcal{E}\left(\mathbf{x}, \mathcal{B}_\gamma\right) = \sum_\gamma e\left(a_\gamma\right), \tag{6.13}$$

donde $e(\cdot)$ es una función escalar con argumento escalar[9]. Entonces es posible encontrar una representación de $\mathbf{x}$ en términos de (6.1) a partir de un algoritmo rápido para resolver:

$$\min_\gamma \left\{ \mathcal{E}(\mathbf{x}, \mathcal{B}_\gamma) \right\}. \tag{6.14}$$

[9]Por ejemplo, si $e(a) = |a|$ entonces $\mathcal{E} = \|\mathbf{a}\|_1$ lo que cual resulta en cierto modo similar a BP. Sin embargo en BOB los elementos de $\mathbf{a}$ que pueden tomar valores diferentes de cero están restringidos a los a_γ en alguna de las bases ortogonales $\mathcal{B}_\gamma$.

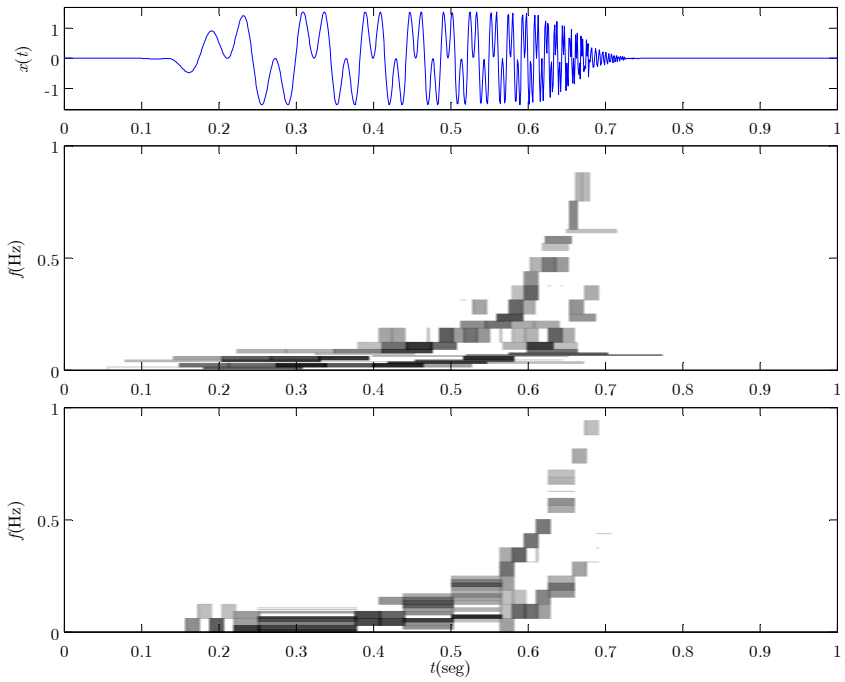

Figura 6.12: *Señal $x(t)$ compuesta como la de la Figura 6.10 (arriba), representación en el plano $t-f$ de $x(t)$ calculada a partir de BOB para el caso de un diccionario WPT (centro). Idem anterior pero para un diccionario CPT (abajo). La representación lograda preserva las características importantes de la señal, aunque la restricción de ortogonalidad impide una mejor resolución de algunos eventos.*

El orden de la cantidad de operaciones necesarias para este algoritmo es $\mathcal{O}(M \log_2(M))$ [20]. En algunos casos este algoritmo da representaciones ralas cercanas al óptimo, sin embargo ésto sólo es posible cuando la representación se puede realizar en términos de una base ortogonal [21]. Otra desventaja del método es que está atado a los diccionarios ya citados. Otros enfoques relacionados, aunque orientados a otras aplicaciones diferentes de la compresión (como por ejemplo LDB [177], LSDB [178] o incluso BSB [180]), sufren inconvenientes similares y no serán explorados en este trabajo por razones de espacio.

En la Figura 6.12 se puede apreciar el resultado de realizar el análisis de una señal artificial mediante BOB con los diccionarios WPT y CPT.

Método de los marcos

MOF [33] selecciona entre todas las posibles representaciones de $\mathbf{x}$ en términos de (6.1), una cuyos coeficientes posean norma ℓ^2 mínima:

$$\min_{\mathbf{a}} \|\mathbf{a}\|_2 \quad \text{sujeto a} \quad \mathbf{\Phi a} = \mathbf{x}. \tag{6.15}$$

La solución de este problema es única, y se llamará $\tilde{\mathbf{a}}$. Geométricamente, la colección de todas las soluciones de (6.1), con $\boldsymbol{\varepsilon} = \mathbf{0}$, es un *subespacio afín* en $\mathbb{R}^M$. MOF selecciona el elemento de este subespacio más cercano al origen. Ésto es llamado a veces una *solución*

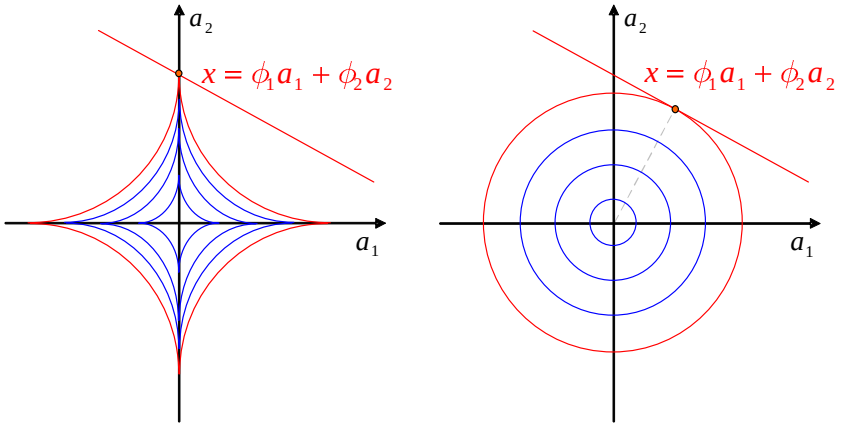

Figura 6.13: *Representación geométrica de la soluciones de (6.8) para* $0 \leq q < 1$ *(izquierda) y* $q = 2$ *(derecha). Obsérvese como ambas soluciones son diferentes. Ésto significa que aunque* $\mathbf{x}$ *pueda sintetizarse a partir de muy pocos elementos, la minimización con respecto a* ℓ_2 *(MOF) no encuentra esta representación, sino una diferente a la de* ℓ_0 *(ideal) y por consiguiente menos rala.*

de longitud mínima. Existe una matriz $\tilde{\mathbf{\Phi}}$, la inversa generalizada de $\mathbf{\Phi}$, la cual calcula la solución de longitud mínima del sistema de ecuaciones lineales:

$$\tilde{\mathbf{a}} = \tilde{\mathbf{\Phi}}\mathbf{x} = \mathbf{\Phi}^T \left(\mathbf{\Phi}\mathbf{\Phi}^T\right)^{-1} \mathbf{x}. \tag{6.16}$$

Para los diccionarios del tipo denominado *marco ajustado* (Ver Definición 2.9), MOF se halla disponible en una forma cerrada. Un buen ejemplo es el caso del diccionario WPT usual. Se puede calcular para todos los vectores $\mathbf{v}$:

$$\left\|\mathbf{\Phi}^T\mathbf{v}\right\|^2 = L_m \left\|\mathbf{v}\right\|^2, \tag{6.17}$$

donde $L_m = \log_2(M)$. O sea que $\tilde{\mathbf{\Phi}} = L_m^{-1}\mathbf{\Phi}^{\mathrm{T}}$. Note que $\mathbf{\Phi}^T$ es simplemente el operador de análisis. El orden de la cantidad de operaciones necesarias para este algoritmo es $\mathcal{O}(M \log_2(M))$ [20].

Existen dos dificultades claves con MOF. La primera es que no preserva la dispersión, es decir que aunque exista una representación muy rala de una señal, los coeficientes encontrados por MOF serán seguramente mucho menos ralos. En la Figura 6.13 se muestra una representación geométrica en $\mathbb{R}^2$ de las soluciones de (6.8) para $0 \leq q < 1$ y $q = 2$ respectivamente. Es posible apreciar como ambas soluciones difieren siempre. Otra dificultad es su limitación intrínseca de resolución. Para mayor detalle acerca de estos inconvenientes consultar [21].

En la Figura 6.14 se puede apreciar el resultado de realizar el análisis de una señal artificial mediante MOF con los diccionarios WPT y CPT.

6.4.2. Caso ruidoso

Un aspecto importante en la solución del problema considerado es cuando se incluye explícitamente el término $\boldsymbol{\varepsilon}$ referente al ruido. Ésto permite encontrar formas de realizar

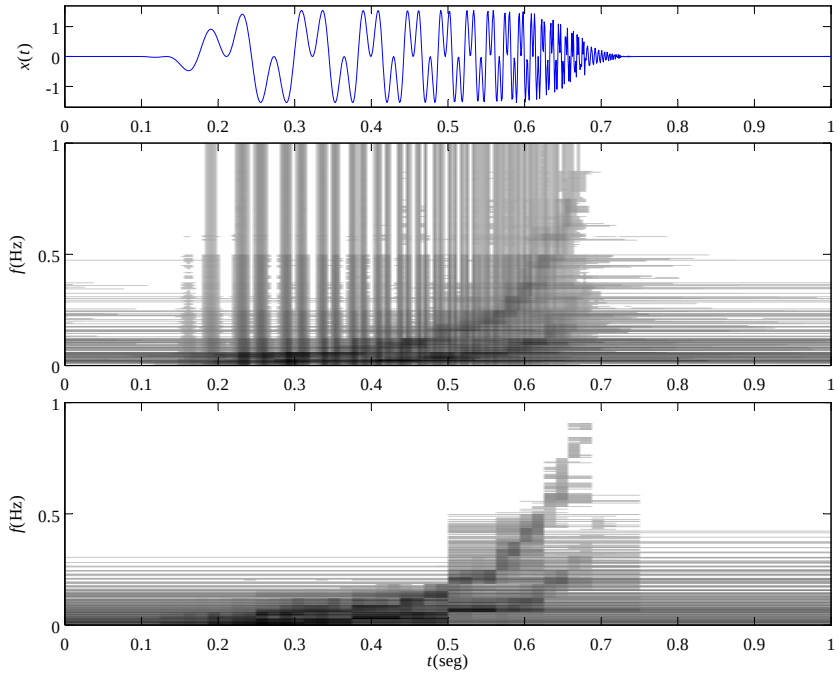

Figura 6.14: *Señal $x(t)$ compuesta como la de la Figura 6.10 (arriba), representación en el plano $t - f$ de $x(t)$ calculada a partir de MOF para el caso de un diccionario WPT (centro). Idem anterior pero para un diccionario CPT (abajo). Se puede observar cómo la representación obtenida no resalta las características importantes de la señal sino que, por el contrario, tiende a ocultarlas.*

una *limpieza de ruido* (en inglés *denoising*) al mismo tiempo que se encuentran los coeficientes $\mathbf{a}$[10]. Cabe recalcar el hecho de que, según se discutió en la Sección 4.6, la mayoría de los enfoques para solucionar problemas de análisis de habla ruidosa primero procesan la señal para extraer la información relevante y luego sobre esta transformación utilizan alguna técnica de limpieza de ruido. Es decir que no se busca necesariamente que la representación posea algún tipo de robustez intrínseca como la planteada en esta sección[11].

Como se mencionó en la Sección 6.2 algunas de las ideas aquí discutidas están relacionadas con las técnicas de limpieza de ruido por umbralamiento de los coeficientes de la transformada ondita, que han sido propuestas por varios investigadores [127]. Sin entrar en detalles se puede decir que este método consiste en:

1. Aplicar la DDWT[12] a la señal original $\mathbf{x}$,

2. Umbralar los coeficientes de detalle $\mathbf{a}$ mediante una función adecuada $\mathbf{a}_\theta = f(\mathbf{a}, \theta)$,

[10] Es posible obtener una versión limpia $\mathbf{x}_{dn}$ de $\mathbf{x}$, a partir de $\mathbf{\Phi a}$.

[11] Salvo algunos casos aislados ya revisados, donde se ha incluido la robustez de otra forma, como por ejemplo ΔC o RASTA-PLP.

[12] También puede aplicarse la WPT con un base ortogonal seleccionada mediante BOB, o a partir de algún otro criterio.

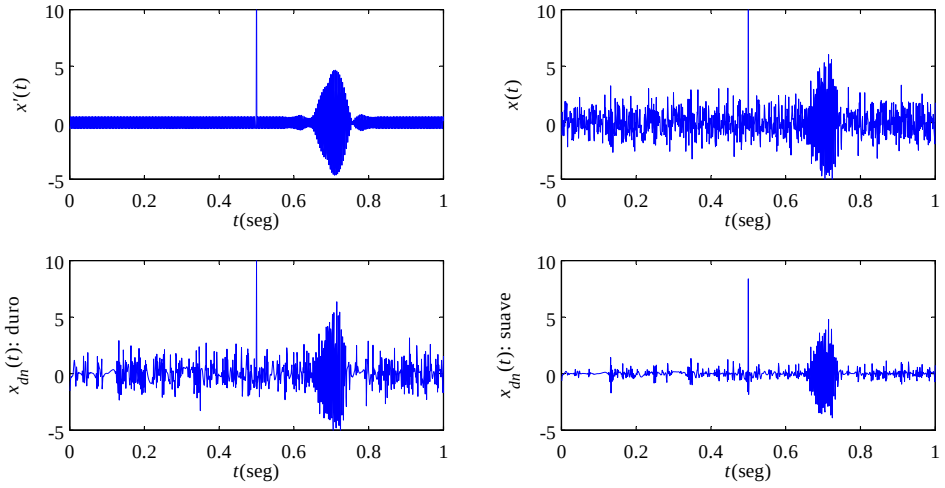

Figura 6.15: *Señal $x(t)$ artificial limpia (como la de la Figura 6.1, arriba izquierda) y ensuciada con ruido blanco (SNR=0 dB, arriba derecha). Resultado de la limpieza mediante: umbralamiento duro (abajo izquierda) y blando (abajo derecha). En ambos casos se utilizó la DDWT con ondita madre Symmlet 8. Se puede apreciar que ambas aproximaciones poseen importantes artefactos, aunque el umbralamiento blando preserva mejor algunas componentes de la señal original.*

3. Aplicar la transformada inversa correspondiente a los coeficientes umbralados $\mathbf{a}_\theta$.

De esta forma se obtiene una versión "limpiada" $\mathbf{x}_{dn}$ de la señal original $\mathbf{x}$. Existen varias maneras para calcular y aplicar los umbrales. La función de umbralamiento duro $f_H(\mathbf{a}, \theta)$ involucra igualar a cero a todos los coeficientes cuyos valores absolutos están debajo de un umbral positivo, mientras que el resto permanecen inalterados. La función de umbralamiento suave $f_S(\mathbf{a}, \theta)$ es similar, sólo que los coeficientes por debajo del umbral son modificados "encogiéndolos" hacia el cero.

En la Figura 6.15 se puede apreciar el resultado de realizar la limpieza de una señal artificial ruidosa mediante ambos métodos de umbralamiento de la DDWT.

A continuación se discutirá la forma de incluir el tratamiento del ruido en la etapa de inferencia del problema de la representación de $\mathbf{x}$, mediante los enfoques determinístico y estadístico respectivamente.

Enfoque determinístico

Dado un diccionario $\mathbf{\Phi}$, el caso ideal consiste nuevamente en encontrar la solución para el problema de representación rala de $\mathbf{x}$ en términos de (6.1) con respecto a ℓ_0 , reescribiendo a (6.5) para incluir el ruido $\boldsymbol{\varepsilon}$ como:

$$\min_{\mathbf{a}} \|\mathbf{a}\|_0 \quad \text{sujeto a} \quad \|\mathbf{x} - \mathbf{\Phi}\mathbf{a}\|_2 \leqslant \sigma_{\boldsymbol{\varepsilon}}, \tag{6.18}$$

donde $\sigma_{\boldsymbol{\varepsilon}}$ es una constante que depende de $\boldsymbol{\varepsilon}$.

Este problema resulta incluso más difícil de resolver que (6.5). Debido a ello se han planteado varias alternativas que se revisarán brevemente a continuación.

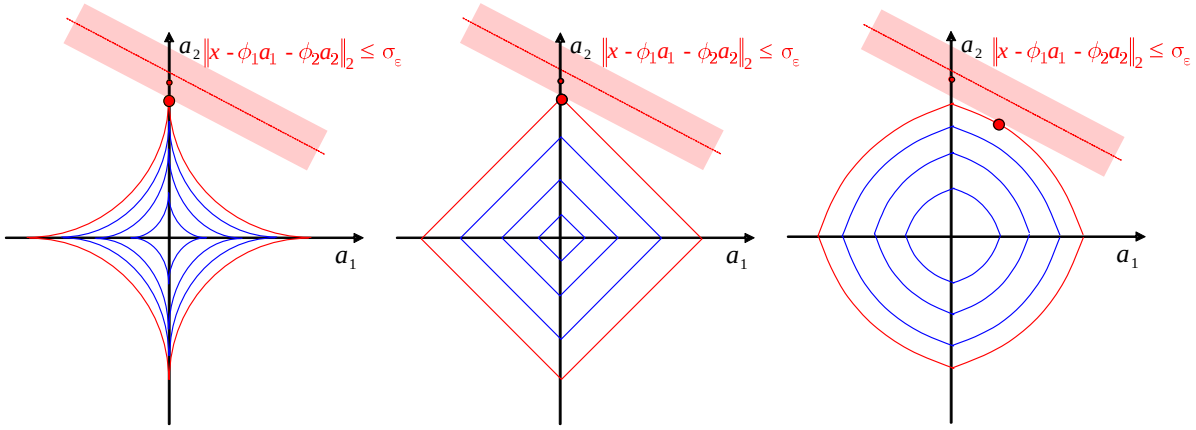

Figura 6.16: *Representación geométrica de la soluciones de (6.20) para* $0 \leq q < 1$ *(izquierda),* $q = 1$ *(centro) y* $q > 1$ *(derecha). Obsérvese que, aunque* $\mathbf{x}$ *pueda sintetizarse a partir de muy pocos elementos, la recuperación exacta de* $\mathbf{a}$ *es poco probable inclusive para el caso de la solución exhaustiva o ideal (*$q = 0$*). Se puede encontrar una representación rala cercana a la ideal para* $q = 1$ *(BP). Para* $q > 1$ *(relacionado con MOF) ambas soluciones son diferentes.*

Limpieza mediante búsqueda de bases y regularización En conexión con BP es posible plantear ahora a (6.18) como una minimización con respecto a la norma ℓ_1 de $\mathbf{a}$ como:

$$\min_{\mathbf{a}} \|\mathbf{a}\|_1 \quad \text{sujeto a} \quad \|\mathbf{x} - \mathbf{\Phi a}\|_2 \leqslant \sigma_{\boldsymbol{\varepsilon}}. \tag{6.19}$$

Este caso fue también planteado por Chen en su tesis [20] y constituye un problema más difícil de resolver que el original sin ruido. Para obtener algo de intuición acerca de este nuevo planteo se muestra, en la Figura 6.16, una representación geométrica en $\mathbb{R}^2$ de las soluciones de:

$$\min_{a_1,a_2} \{|a_1|^q + |a_2|^q\} \quad \text{sujeto a} \quad \|x - \phi_1 a_1 - \phi_2 a_2\|_2 \leqslant \sigma_\varepsilon, \tag{6.20}$$

para $0 \leq q < 1$, $q = 1$ y $q > 1$ respectivamente.

En esta figura puede apreciarse como, debido a los efectos del ruido, la recuperación exacta de $\mathbf{a}$ es poco probable inclusive para el caso de la solución exhaustiva respecto a ℓ_0. Sin embargo, nuevamente se puede encontrar una representación rala relativamente cercana a la ideal para el caso de $q = 1$. Para $q > 1$ las soluciones son muy diferentes.

Se puede frasear a los problemas del tipo de (6.19) en términos de la teoría de regularización [57]. La idea principal consiste en agregar a la minimización original respecto de alguna función de $\mathbf{a}$ un término adicional de penalización. Este término involucra generalmente alguna medida de ajuste de los datos al modelo, suponiendo la existencia de ruido aditivo, como ocurre en el caso que aquí se considera.

En general el problema en términos de la expresión regularizada tiene la siguiente forma:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \{d(\mathbf{x}, \mathbf{\Phi a}) + \lambda f(\mathbf{a})\}, \tag{6.21}$$

donde $f(\mathbf{a})$ es alguna función de los coeficientes, $d(\mathbf{x}, \mathbf{\Phi a})$ es alguna función que mide la distancia entre la señal y el modelo, y $\lambda \in \mathbb{R}$ es un factor de peso.

En el caso propuesto por Chen[13] se tiene:

$$f(\mathbf{a}) = \|\mathbf{a}\|_1 \text{ y } d(\mathbf{x}, \mathbf{\Phi a}) = \|\mathbf{\Phi a} - \mathbf{x}\|_2^2 . \tag{6.22}$$

Ésto complica un poco el problema de minimización original de (6.6) volviéndolo uno de *programación cuadrática*. Se podría decir que en realidad se están solucionando dos problemas en forma simultánea. El factor de proporcionalidad λ permite ajustar el peso relativo de ambos requerimientos. Chen introdujo estas ideas en su tesis en lo que denominó *limpieza por búsqueda de bases* (BPD), y posteriormente fueron extendidas a casos con ruidos más generales en el trabajo de Sardy [182] en la técnica que denominó de *búsqueda de bases generalizada* (GBP).

Limpieza mediante búsqueda por coincidencia A pesar de que se ha presentado a MP dentro de los casos "limpios" es posible también interpretar al residuo $\mathbf{R}$ como equivalente a $\boldsymbol{\varepsilon}$ en (6.1). Ésto permite aplicar MP también al caso de limpieza de ruido tomando nuevos átomos mientras que:

$$\|\mathbf{x} - \mathbf{\Phi a}\|_2 > \sigma_\varepsilon .$$

El valor de la cantidad de pasos m óptima para separar la señal del ruido está relacionado con la SNR. En la Figura 6.17 se puede apreciar el resultado de realizar la limpieza de una señal artificial ruidosa mediante los diferentes métodos discutidos en esta sección.

Equivalencia entre enfoques Como se ha mencionado anteriormente el enfoque determinístico resulta paralelo al estadístico, que se revisará a continuación, y arroja una serie de soluciones equivalentes a las de las ecuaciones (6.29, 6.31 y 6.33) de la sección siguiente. Obsérvese que para el caso determinista los problemas se plantean en términos de minimizaciones (del error o distancia) y para el probabilista de maximizaciones (de la verosimilitud o probabilidad). Existen trabajos donde se exploran otras posibilidades tratando a las $f(\mathbf{a})$ en (6.21) como funciones de activación de una red neuronal. Ésto permite una interpretación biológica más directa, aunque con un enfoque también probabilístico [71].

Enfoque estadístico

Como se ha visto la inclusión del ruido complica el problema de encontrar una representación rala adecuada. Sin embargo, permite también su tratamiento más directo desde el punto de vista probabilista o estadístico. Se puede decir entonces que existen señales "más probables" que otras para un diccionario y un conjunto de coeficientes dados. Ésto

[13] Aunque el enfoque es determinístico, en este caso se está suponiendo implícitamente que el ruido es gaussiano.

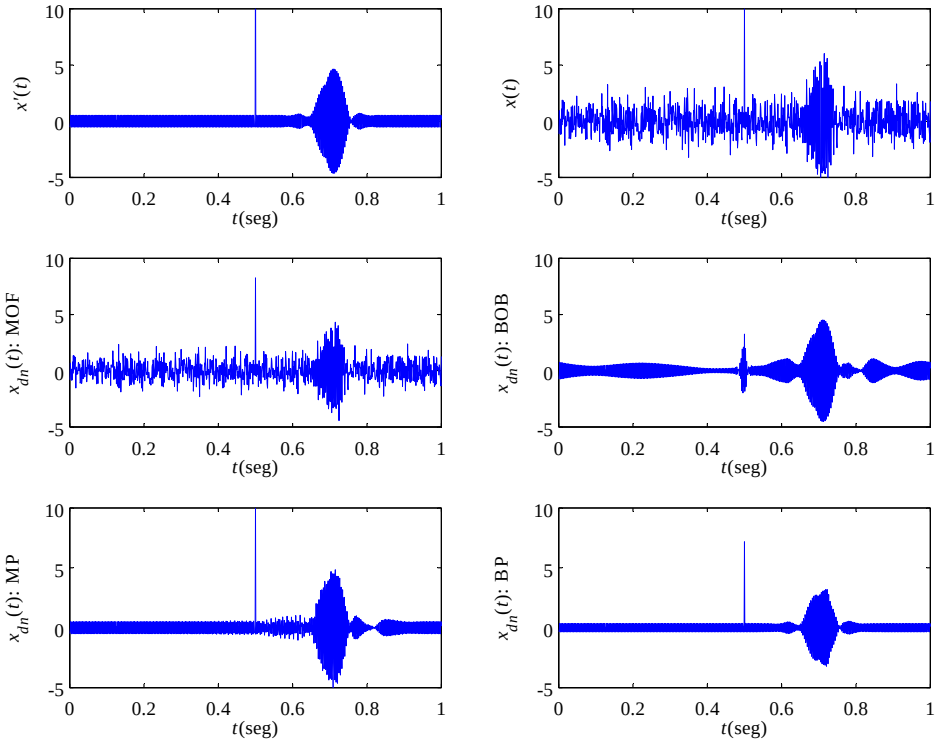

Figura 6.17: *Señal $x(t)$ artificial limpia (como la de la Figura 6.1, arriba izquierda) y ensuciada con ruido blanco (SNR=0 dB, arriba derecha). Resultado de la limpieza mediante: MOF (centro izquierda), BOB (centro derecha), MP (abajo izquierda) y BP (abajo derecha). En todos los casos se utilizó un diccionario WPT Symmlet 8 de profundidad 10. Se puede apreciar a simple vista que la aproximación realizada por BP presenta menos artefactos, aunque resulta ligeramente atenuada en amplitud. Adaptado de [20].*

permite tener un modelo para tomar decisiones si se conoce algo de la estadística del proceso, como por ejemplo la distribución de probabilidad $P(\mathbf{x}|\mathbf{\Phi}, \mathbf{a})$[14].

Para obtener una representación rala puede suponerse una distribución con curtosis positivo para cada coeficiente a_i. Según se mostró anteriormente otra característica deseable es suponer que los a_i sean estadísticamente independientes con una distribución a priori conjunta[15]:

$$P(\mathbf{a}) = \prod_i P(a_i). \tag{6.23}$$

Ésto conecta los resultados de esta sección con la técnicas de ICA descriptas en la Sección 2.5. Es posible ver nuevamente a (6.1) como un modelo generativo. Siguiendo la terminología utilizada en el campo de ICA, ésto significa que la señal $\mathbf{x} \in \mathbb{R}^N$ se genera a partir de un conjunto de fuentes a_j (arregladas en la forma de un vector de estado $\mathbf{a} \in \mathbb{R}^M$) utilizando una matriz de mezcla $\mathbf{\Phi}$ (de tamaño $N \times M$, con $M \geq N$), e incluyendo un término de ruido aditivo $\boldsymbol{\varepsilon}$ (generalmante gaussiano).

Si se conoce $\mathbf{\Phi}$ y $\mathbf{x}$, es posible estimar $\mathbf{a}$ considerando la distribución a posteriori:

$$P(\mathbf{a}|\mathbf{\Phi}, \mathbf{x}) = \frac{P(\mathbf{x}|\mathbf{\Phi}, \mathbf{a})P(\mathbf{a})}{P(\mathbf{x}|\mathbf{\Phi})}. \tag{6.24}$$

Una estimación de $\mathbf{a}$ de *probabilidad a posteriori máxima* (MAP) sería:

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} \left[\log P(\mathbf{x}|\mathbf{\Phi}, \mathbf{a}) + \log P(\mathbf{a})\right]. \tag{6.25}$$

Si la posterior $P(\mathbf{a}|\mathbf{\Phi}, \mathbf{x})$ es suficientemente suave, puede encontrarse el máximo por gradiente ascendente. La solución depende de la forma de la distribución supuesta para el ruido (que está relacionada con $P(\mathbf{x}|\mathbf{\Phi}, \mathbf{a})$) y para los coeficientes (distribución a priori $P(\mathbf{a})$), dando lugar a diferentes métodos para el cálculo de los coeficientes que se mencionarán a continuación.

Distribución a priori exponencial-logarítmica y ruido gaussiano En [143] se utiliza para la a priori una función exponencial de la forma:

$$P(a_i) = \alpha\ e^{-f_i(a_i)}, \tag{6.26}$$

donde $\alpha = 1 \,/ \int e^{-f_i(a_i)} da_i$ es una constante de normalización y $f_i(\cdot)$ es una función de costo no convexa con valores en $\mathbb{R}^M$ y parámetros β y $\sigma_{a_i} \in \mathbb{R}$, como por ejemplo:

$$f_i(a_i) = \beta\ \log\left(1 + (a_i/\sigma_{a_i})^2\right), \tag{6.27}$$

Éste caso corresponde a una función de distribución de probabilidad para a_i tipo Cauchy, siendo β el parámetro que controla cuan "picuda" resulta la distribución, y σ_{a_i} el parámetro de escala relacionado con la desviación standard de los coeficientes (Ver Figura 6.18).

[14]Aunque estrictamente corresponde utilizar $P(\cdot)$ para las distribuciones de probabilidad y $p(\cdot)$ para las densidades correspondientes, en esta sección se utilizará genéricamente el término distribución con su correspondiente notación, debiendo quedar claro el significado exacto de acuerdo al contexto. Para simplificar el desarrollo no se empleará una notación especial para designar a las variables aleatorias.

[15]Por esta última propiedad a los códigos generados de esta forma se los suele llamar también *códigos factoriales*.

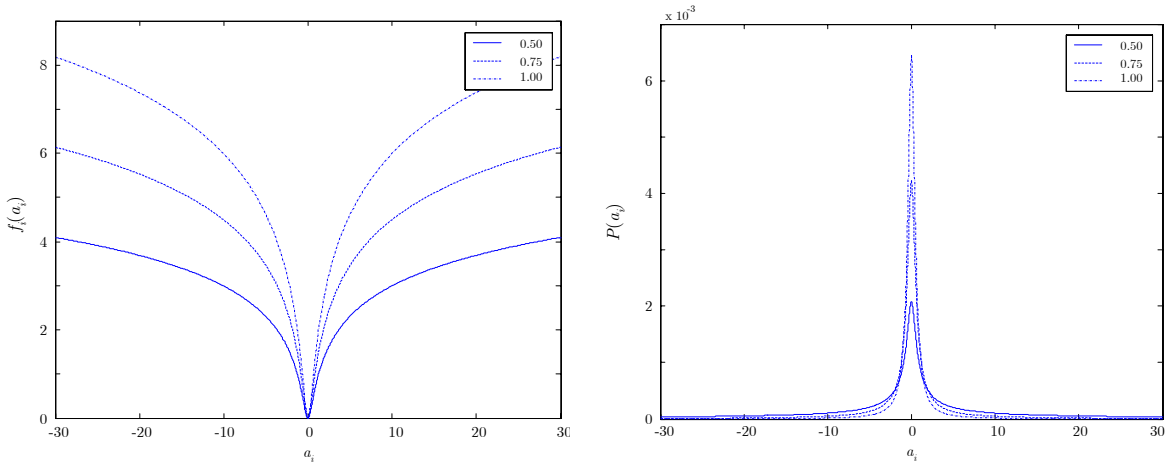

Figura 6.18: *Función de costo $f(a_i)$ (izquierda) y su distribución $P(a_i)$ correspondiente (derecha) para $\sigma_{a_i} = 0{,}5$ y diferentes valores de β en (6.27). Los parámetros de la función de costo permiten modificar la curtosis de la ditribución para ajustar la dispersión de los coeficientes a_i.*

Si se supone ruido aditivo gaussiano $\boldsymbol{\varepsilon}$ con matriz de covarianza $\mathcal{E}\left[\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}\right] = \boldsymbol{\Lambda}_{\boldsymbol{\varepsilon}}^{-1}$, entonces la probabilidad de observar una $\mathbf{x}$ particular, dado un diccionario $\boldsymbol{\Phi}$ conocido y coeficientes $\mathbf{a}$, es:

$$P(\mathbf{x}|\boldsymbol{\Phi},\mathbf{a}) = \rho\, e^{-\frac{1}{2}\boldsymbol{\varepsilon}^T\boldsymbol{\Lambda}_{\boldsymbol{\varepsilon}}\boldsymbol{\varepsilon}}, \tag{6.28}$$

donde ρ es una constante de normalización.

A partir de aquí la solución MAP mediante gradiente ascendente se obtiene la siguiente regla de actualización para $\mathbf{a}$:

$$\Delta\mathbf{a} = \boldsymbol{\Phi}^T\boldsymbol{\Lambda}_{\boldsymbol{\varepsilon}}\boldsymbol{\varepsilon} - \nabla_{\mathbf{a}} f(\mathbf{a}). \tag{6.29}$$

Distribución a priori laplaciana y ruido gaussiano En [115] Lewicki y Olshausen proponen utilizar una distribución a priori de tipo laplaciana con paramétro β_i :

$$P(a_i) = \alpha\, e^{-\beta_i|a_i|}, \tag{6.30}$$

donde α es una constante de normalización.

En conjunción con la suposición de ruido gaussiano nuevamente, ésto lleva a la siguiente regla de actualización para $\mathbf{a}$:

$$\Delta\mathbf{a} = \boldsymbol{\Phi}^T\boldsymbol{\Lambda}_{\boldsymbol{\varepsilon}}\boldsymbol{\varepsilon} - \boldsymbol{\beta}^T\left|\mathbf{a}\right|. \tag{6.31}$$

Ésta regla resulta equivalente a BPD propuesto por Chen [21] y presentado en la sección anterior [115].

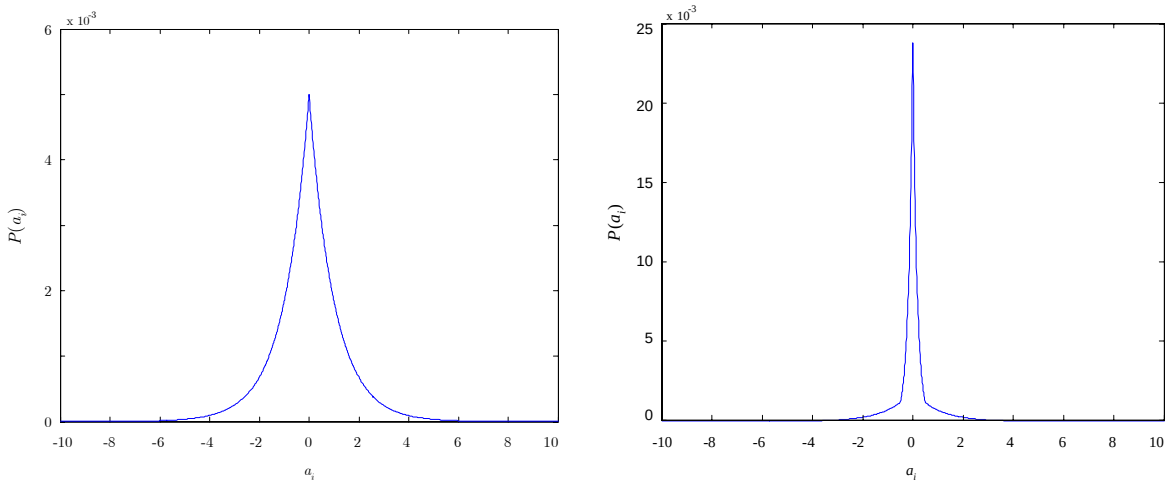

Figura 6.19: *Comparación entre una distribución laplaciana simple (izquierda) y una mixta (derecha), a partir de (6.32) con $K = 6$ y $\mu = 0{,}5$. Mediante estos parámetros se tiene más control sobre la relación entre el pico y las colas de la distribución, lo que permite ajustar la dispersión de los coeficientes a_i.*

Distribución a priori mixta y ruido gaussiano La aproximación seguida por Abdallah y Plumbley [1, 2] para lograr una distribución rala, consiste en una versión mixta formada por dos laplacianas de la siguiente forma:

$$P(a_i) = \begin{cases} \alpha\, e^{-|a_i|} & \text{si } |a_i| \geq \mu\ , \\ \alpha\, C\, e^{-K|a_i|} & \text{si } |a_i| < \mu\ , \end{cases} \tag{6.32}$$

donde α es una constante de normalización, y $C = e^{\mu(K-1)}$ para asegurar continuidad. Los parámetros μ y K controlan el ancho y la masa relativa del pico central. En la Figura 6.19 se muestra un ejemplo de esta distribución mixta comparada con una laplaciana simple.

Éste tipo de distribución lleva a un comportamiento tipo umbralamiento al calcular $\hat{\mathbf{a}}$ [84] y la solución MAP mediante gradiente ascendente se convierte en la siguiente regla de actualización para $\mathbf{a}$:

$$\Delta\mathbf{a} = \mathbf{\Phi}^T \mathbf{\Lambda}_{\boldsymbol{\epsilon}} \boldsymbol{\epsilon} - \gamma(\mathbf{a}), \tag{6.33}$$

donde:

$$\gamma_i(a_i) = \begin{cases} \text{signo}(a_i) & \text{si } |a_i| \geq \mu\ , \\ \text{signo}(a_i)\ K & \text{si } |a_i| < \mu\ . \end{cases} \tag{6.34}$$

Distribución a priori laplaciana y ruido no gaussiano En [182] Sardy, desarrolla un método general para encontrar los coeficientes en el caso de que sea a priori laplaciana para distintas distribuciones convexas no necesariamente gaussianas mediante GBP. En dicho trabajo se muestran los casos para las distribuciones exponencial, Poisson y Bernoulli. La solución se plantea en términos de un problema de programación cuadrática similar al resuelto por Chen para BPD.

En [57] Girosi plantea otra alternativa que tendría que ver con otros tipos de ruido pero estableciendo una conexión entre RT y SVM a través de las *funciones núcleos*.

6.5. Búsqueda del diccionario o aprendizaje

En la sección anterior se presentaron varios métodos que permiten encontrar los coeficientes de una representación suponiendo que se cuenta con un diccionario predeterminado. Surge naturalmente el interrogante acerca de cómo elegir o generar un diccionario $\mathbf{\Phi}$ adecuado para una aplicación particular. En este sentido existen, en principio, dos enfoques posibles. El primero consiste en "armar a mano" el diccionario mediante la utilización del conocimiento a priori sobre las características que se quieren encontrar en la señal. Este enfoque requiere del criterio a priori del "diseñador" el análisis, por lo que se denominará *diseño a medida*. La otra posibilidad corresponde a la búsqueda no supervisada de un diccionario óptimo, en forma similar a como se encaró el problema de encontrar los coeficientes en la sección anterior. Ésto es, imponiendo algunas restricciones a la solución de la ecuación (6.1) y utilizando datos de las señales reales que se van a analizar[16]. A este último enfoque se lo denominará *ajuste automático*.

6.5.1. Diseño a medida

Para el primer enfoque se puede hacer uso de los diccionarios basados en funciones paramétricas tradicionales discutidos en el Capítulo 5. Algunos de los diccionarios comúnmente utilizados son los frecuenciales (DFT), temporales (basados en impulsos desplazados), tiempo-frecuencia (Gabor, STFT o CPT), tiempo-escala (DDWT, WPT), entre otros (Heaviside, polinomios, etc.). También pueden utilizarse combinaciones de estos diccionarios individuales para formar otros más grandes que incluyan una variedad importante de características[17].

Cuando existen diferentes clases de señales y resulta difícil elegir algún diccionario que represente en forma adecuada el comportamiento de todas las clases simultáneamente, puede ser útil buscar conjuntos de funciones apropiadas para cada tipo de señales y posteriormente mezclarlas en un único super-diccionario.

[16]Existe también la posibilidad de emplear enfoques mixtos, por ejemplo se puede inicializar con una base o diccionario conocido y luego adaptarlo a los datos.

[17]Aquí resulta necesario recalcar que, al igual que en el caso de un único diccionario simple, el procesamiento se realiza simultáneamente con todos los diccionarios seleccionados juntos y es el criterio empleado el que determina cuáles átomos van a entrar en juego para realizar el análisis de una señal determinada, lo que asegura finalmente la unicidad.

6.5.2. Ajuste automático

Enfoque deterministico

Es posible aprender el diccionario de manera que se adapte al ambiente planteando el siguiente problema de regularización [106]:

$$\hat{\mathbf{\Phi}} = \underset{\mathbf{\Phi}, a_1, \cdots, a_M}{\arg\min} \{d(\mathbf{x}, \mathbf{\Phi}\mathbf{a}) + \lambda f(\mathbf{a})\}. \tag{6.35}$$

Como se ha visto anteriormente la solución encontrada mediante este enfoque suele ser equivalente a la obtenida con el enfoque estadístico y restricciones análogas o "duales".

Enfoque estadístico:

Para estimar el valor de $\mathbf{\Phi}$, es posible maximizar la siguiente función objetivo:

$$\mathcal{L} = \mathcal{E}[\log P(\mathbf{x}|\mathbf{\Phi})]_{P(\mathbf{x})}\,, \tag{6.36}$$

donde $\mathcal{E}[\cdot]$ indica el valor esperado tomado sobre la distribución de vectores observados $\mathbf{x}$. A $\mathcal{L}$ se la denomina *verosimilitud* de los datos en relación al modelo y se estima en base a la evidencia de los datos. Esta evidencia puede estimarse marginalizando el producto de la distribución de los datos, dados el diccionario y los coeficientes, con la distribución a priori de los coeficientes de la siguiente forma:

$$P(\mathbf{x}|\mathbf{\Phi}) = \int_{\mathbb{R}^M} P(\mathbf{x}|\mathbf{\Phi}, \mathbf{a})\, P(\mathbf{a})\, d\mathbf{a}, \tag{6.37}$$

donde se trata con una integral en el espacio M-dimensional disponible para los estados $\mathbf{a}$. Si ahora se maximiza la función objetivo mediante gradiente ascendente igualando su derivada a cero:

$$\frac{\partial \mathcal{L}}{\partial \phi_{ij}} = 0, \tag{6.38}$$

se obtiene una regla de actualización para la matriz $\mathbf{\Phi}$:

$$\Delta\mathbf{\Phi} = \eta \mathbf{\Lambda}_{\boldsymbol{\varepsilon}}\ \mathcal{E}\left[\boldsymbol{\varepsilon}\mathbf{a}^T\right]_{P(\mathbf{a}|\mathbf{\Phi},\mathbf{x})}\,, \tag{6.39}$$

donde $\eta \in \mathbb{R}$ es un coeficiente de aprendizaje (que varía entre 0 y 1).

El problema principal consiste en el cálculo de esta regla de actualización ya que implica resolver la siguiente integral:

$$\mathcal{E}\left[\boldsymbol{\varepsilon}\mathbf{a}^T\right]_{P(\mathbf{a}|\mathbf{\Phi},\mathbf{x})} = \int (\mathbf{x} - \mathbf{\Phi}\mathbf{a})\, \mathbf{a}^T P(\mathbf{a}|\mathbf{\Phi}, \mathbf{x}) d\mathbf{a}, \tag{6.40}$$

la que crece exponencialmente a medida que la dimensión de $\mathbf{a}$ crece. Para hacer tratable esta integral distintos autores han realizado diferentes aproximaciones que se mencionarán a continuación.

Aproximación Delta Olshausen y Field [145, 143] propusieron colapsar la posterior a una distribución delta multivariada en su valor máximo $\hat{\mathbf{a}}$:

$$P(\mathbf{a}|\mathbf{\Phi}, \mathbf{x}) \approx \delta(\mathbf{a} - \hat{\mathbf{a}}), \tag{6.41}$$

lo que equivale a tomar una sola muestra y nos lleva a la siguiente solución de la ecuación de actualización:

$$\Delta\mathbf{\Phi} = \eta\mathbf{\Lambda}_{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}\hat{\mathbf{a}}^T, \tag{6.42}$$

donde se define $\hat{\boldsymbol{\varepsilon}} = \mathbf{x} - \mathbf{\Phi}\hat{\mathbf{a}}$.

Como esta aproximación pierde la información de volumen de la integral, si no se toman precauciones, la matriz puede crecer sin límites y los valores estimados de $\hat{\mathbf{a}}$ tienden a cero. Por lo tanto en este caso se requiere un paso explícito de normalización. Luego de cada paso de aprendizaje se re-escalan los átomos del diccionario de manera que su norma ℓ_2 mantenga un nivel adecuado de varianza en cada coeficiente a_i correspondiente:

$$\|\boldsymbol{\phi}_i^{nuevo}\|_2 = \left\|\boldsymbol{\phi}_i^{anterior}\right\|_2 \left[\frac{\mathcal{E}[a_i^2]}{\sigma_{a_i}^2}\right]^{\alpha}, \tag{6.43}$$

donde σ_{a_i} es el parámetro de escala de la función de costo (6.27) y $\alpha \in \mathbb{R}$ es un coeficiente de ajuste.

Aproximación Gaussiana Lewicki y Sejnowski [116] usaron como aproximación a la posterior alrededor de su máximo $\hat{\mathbf{a}}$, una gaussiana multivariada:

$$P(\mathbf{a}|\mathbf{\Phi}, \mathbf{x}) \approx \sqrt{\frac{|\mathbf{H}|}{(2\pi)^M}}\, e^{-\frac{1}{2}(\mathbf{a}-\hat{\mathbf{a}})^T\mathbf{H}(\mathbf{a}-\hat{\mathbf{a}})}. \tag{6.44}$$

De esta forma, por construcción la media de la gaussiana es $\hat{\mathbf{a}}$, y su matriz de covarianza es $\mathbf{H}^{-1}$, dónde $\mathbf{H}$ es el Hessiano de la log-posterior evaluado en $\hat{\mathbf{a}}$:

$$\mathbf{H} = -\nabla\nabla^T \log P(\mathbf{a}|\mathbf{\Phi}, \mathbf{x}), \tag{6.45}$$

lo que asegura una buena aproximación cerca de $\hat{\mathbf{a}}$.

Ésto resulta en una solución de la forma:

$$\Delta\mathbf{\Phi} = \eta\mathbf{\Lambda}_{\boldsymbol{\varepsilon}}(\hat{\boldsymbol{\varepsilon}}\hat{\mathbf{a}}^T - \mathbf{\Phi}\mathbf{H}^{-1}). \tag{6.46}$$

Teniendo en cuenta el ancho de la posterior cerca de su pico, la regla de actualización agrega el término de decaimiento $\mathbf{\Phi}\mathbf{H}^{-1}$ que soluciona el problema de crecimiento ilimitado de la aproximación anterior.

En la Figura 6.20 se puede apreciar el diccionario $\mathbf{\Phi}$ encontrado con un método como de ajuste automático a partir de señales de audio y música obtenidos de la radio BBC [1].

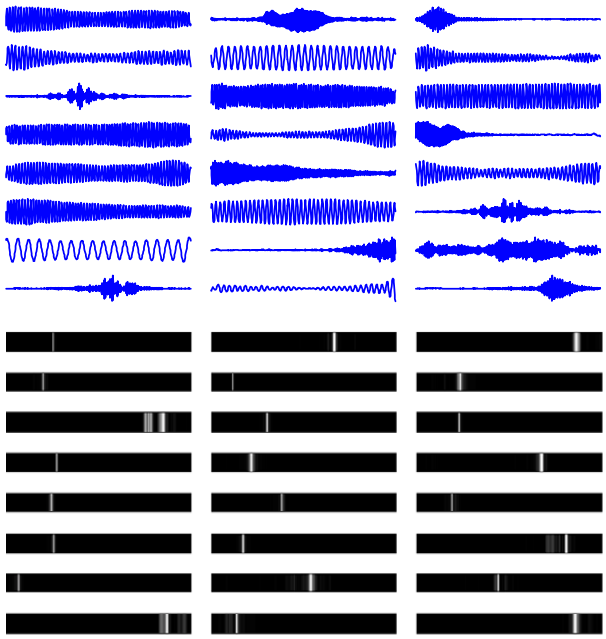

Figura 6.20: *Átomos del diccionario* $\mathbf{\Phi}$ *(arriba) y sus correspondientes espectros (abajo, representados como bandas horizontales) encontrados mediante un método de ajuste automático, a partir de señales de música obtenidas de la radio BBC [1]. Es posible observar que varios átomos se parecen a tonos puros, lo que se correlaciona con la naturaleza armónica de la música contenida en los datos.*

6.6. Comentarios de cierre del capítulo

En este capítulo se han presentado los fundamentos de las técnicas que permiten obtener una representación rala y/o independiente de una señal determinada. Se han discutido las ventajas y desventajas de este tipo de representaciones, junto con los métodos que permiten obtenerlas a partir de enfoques determinísticos o estadísticos. Entre las principales ventajas figura la posibilidad de obtener una representación "limpiada", es decir una representación en la cual se han "suprimido" los efectos del ruido. La cuestión acerca de cual es el análisis que resulta más adecuado para una aplicación particular continua implícita. En este capítulo se introducen herramientas adicionales para acercarse a una respuesta. Algunos de los métodos presentados permiten obtener no sólo los coeficientes, sino también el diccionario óptimo. Es posible imponer restricciones adicionales a las aquí planteadas para encontrar el diccionario, de manera que resulten útiles para una aplicación particular. Otra forma consiste en elegir uno o más de diccionarios de los presentados en el Capítulo 5, como por ejemplo los basados en onditas o paquetes de onditas.

En el siguiente Capítulo se presentará la aplicación de todas estas técnicas al caso del habla en comparación con los enfoques más convencionales.

Capítulo 7

Aplicaciones a la señal de voz

"Si alguno tiene oídos, que oiga."

(Marcos 4,23)

Contenido

7.1. Introducción

EN los capítulos previos se ha presentado un conjunto de técnicas que permiten obtener la representación de una señal basándose en diferentes enfoques, que van desde los convencionales o clásicos hasta los más recientes o no convencionales. En este capítulo se desarrolla una serie de alternativas para la aplicación de técnicas no convencionales a la representación de la señal de voz. Estas alternativas constituyen una parte de los aportes originales de este trabajo.

En el Capítulo 2 se han presentado varias medidas que permiten cuantificar algunos aspectos importantes para lograr una "buena" representación. Estas medidas "directas" forman parte del primer análisis cuantitativo de algunas de las representaciones obtenidas. Sin embargo, otra de las pautas discutidas para evaluar si la representación constituye un buen modelo de la señal, es su ajuste al propósito. El interrogante acerca de cuál resulta la representación óptima para la señal de voz, requiere entonces precisar

el contexto de aplicación. Por ello se presentan también algunos experimentos que tratan de orientar la búsqueda, principalmente en el contexto de la clasificación de fonemas.

Aunque en estos experimentos se utilizan diferentes técnicas de clasificación y modelado de unidades acústico-fonéticas del habla, este aspecto reviste aquí un carácter secundario. Ésto se debe a que el objetivo principal de la utilización de estas técnicas de clasificación en el presente contexto es el de cuantificar el desempeño de las diferentes representaciones de la señal de voz en una aplicación concreta. Por la misma razón tampoco se pretende dar una solución definitiva a alguno de los problemas planteados por los experimentos, sino más bien proponer un camino de búsqueda hacia esta solución, particularmente en lo que se refiere a la representación de la señal. No se entrará a mayores detalles acerca de los fundamentos teóricos de estas técnicas, los que pueden consultarse en la extensa bibliografía disponible al respecto (como por ejemplo [210, 82, 37, 132, 91]).

El capítulo está organizado de la siguiente manera. A continuación se describen los aspectos generales de las pruebas y experimentos realizados con el fin de evaluar las técnicas de representación desarrolladas en este capítulo. En la Sección 7.3 se describen los experimentos con las técnicas convencionales que serán utilizados como referencia. En la Sección 7.4 se describe un método sencillo para mejorar la algunas características de las representaciones convencionales basado en la inclusión de información relacionada con los cambios de dinámica del aparato fonador. En la Secciones 7.5 y 7.6 se presentan las técnicas basadas en onditas y representaciones ralas y/o independientes propuestas en el presente trabajo, y se las compara con los enfoques más tradicionales. Los comentarios finales del capítulo se realizan en la Sección 7.7.

7.2. Descripción de los experimentos

En esta sección se presentan los aspectos generales del camino seguido para la evaluación de las diferentes representaciones. La hipótesis de este trabajo consiste en suponer que pueden aprovecharse algunas propiedades de las representaciones basadas en técnicas no convencionales para mejorar el desempeño de los sistemas artificiales que emulan la comunicación humana. Ésto es, respecto al enfoque clásico que debido a numerosas simplificaciones no contempla ciertos aspectos explícitamente. Para probarla se llevaron a cabo diferentes comparaciones entre los resultados *de referencia* derivados de los experimentos con las representaciones convencionales, en contraste con los de las alternativas propuestas. El tipo de experimentos elegido para realizar esta comparación ha sido consecuencia de querer incluir aspectos o medidas directas de las características de las representaciones obtenidas, junto con otros que son producto del desempeño de los sistemas artificiales que las utilizan. Esto permite también establecer relaciones entre estas diferentes "miradas" acerca de las representaciones. Los resultados reportados en este capítulo corresponden principalmente a las siguientes instancias:

Experimentos de análisis cualitativo de los diccionarios: La utilidad de un determinado diccionario depende de que los elementos que lo constituyen permitan describir adecuadamente a la señal. En varios casos se han utilizado diccionarios de funciones con características bien conocidas, o bien se ha diseñado o adaptado

el diccionario para lograr una representación "a medida" de los datos a analizar. Ésto permite la realización de un análisis cualitativo de las bases o diccionarios utilizados y de su relación con las características significativas de la señal de voz.

Experimentos de evaluación cuantitativa de las representaciones: En la Sección 2.4.2 se presentaron diversas medidas que permiten evaluar la calidad de una representación. Entre estas medidas pueden contarse diferentes normas y otras derivadas de la estadística y de la teoría de información. Éstas permiten valorar cuan bien es representada una señal en términos del modelo subyacente en la representación. En relación con el ajuste al propósito se implementaron experimentos adicionales que se describen a continuación.

Experimentos de clasificación de fonemas: El objetivo de estos experimentos es el de clasificar un conjunto de fonemas del idioma inglés. Los experimentos se realizaron utilizando clasificadores basados en redes neuronales artificiales con retardos temporales, que han mostrado buenos resultados para este tipo de tareas [210]. El entrenamiento se realizó con habla limpia, proveniente de los fonemas */b/*, */d/*, */jh/*, */eh/*, y */ih/* de la base de datos TIMIT [56]. Dado que uno de los aspectos a ser evaluados es la robustez de la representación o sus posibilidades para implementar métodos de limpieza de ruido, se realizaron también pruebas con habla contaminada con ruido aditivo. Para mayores detalles referirse al Apéndice A.

Expermientos de reconocimiento de habla continua: Los HMMs poseen características útiles para el ASR, como por ejemplo el tratamiento integrado y uniforme de los distintos niveles dentro del reconocedor [160]. Por ello se realizaron también algunos experimentos utilizando esta técnica. Estos experimentos consistieron en el reconocimiento de habla continua en castellano a partir de un subconjunto de la base de datos Albayzin [17], según se describe en el Apéndice B. También se realizaron pruebas de robustez al ruido aditivo.

Además de los anteriores se realizaron algunos experimentos de limpieza de ruido sencillos y otros que utilizaron no sólo señales de voz reales, sino también datos generados artificialmente a fin de resaltar algunos aspectos específicos de las representaciones obtenidas.

7.3. Representaciones convencionales

Dentro de las representaciones convencionales se seleccionaron para tomar como referencia: la STDFT y los coeficientes cepstrales, en escala frecuencial lineal y psicoacústica de mel. Como se mencionó en la Sección 4.5 éstas constituyen las más utilizadas en el campo del habla. En la Figura 7.1 se puede apreciar una emisión típica correspondiente a un fonema de la base de datos TIMIT junto con su representación espectral tradicional y en escala de mel.

Para el caso de la STDFT, la forma del diccionario y los átomos correspondientes fueron descriptos en el Capítulo 5. En el caso de la transformada de Fourier en escala

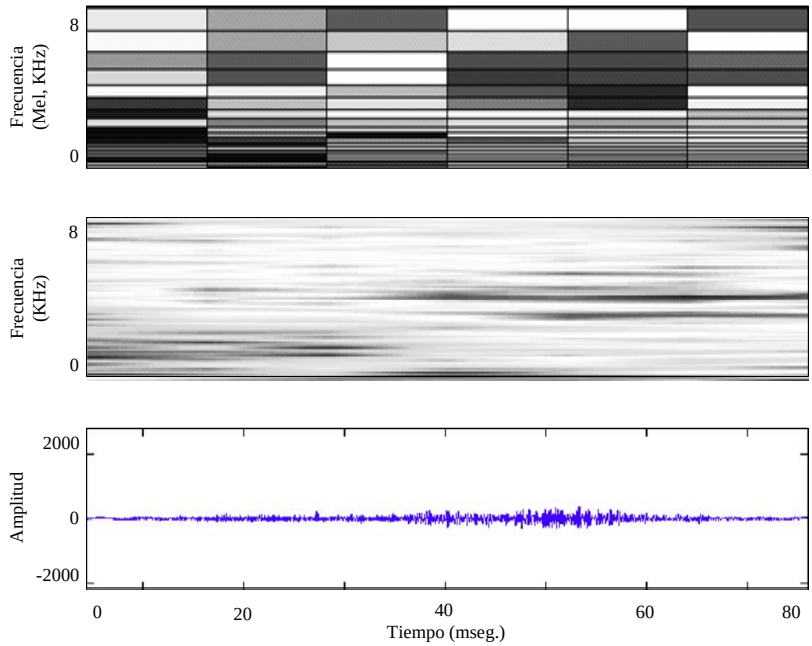

Figura 7.1: *Sonograma (abajo), espectrograma (centro) y transformada discreta de Fourier de corta duración en escala de mel (tramo 250 muestras, desplazamiento 10 mseg, arriba), correspondiente al fonema /jh/ de la frase "She had your dark suit in greasy wash water all year" (TIMIT). Para representar el espectrograma se ha utilizado una interpolación bidimensional que "oculta" su carácter discreto.*

de mel los átomos del diccionario toman la forma que se muestra en la Figura 7.2[1]. Con algunas aproximaciones es posible también "ver" al análisis cepstral real en escala de mel (MFCC) como realizado mediante un diccionario particular. Los átomos de este "diccionario" $\mathbf{\Phi'}$ pueden obtenerse a través de la inversión (aproximada) de la transformación, partiendo de cada uno de los coeficientes de la representación. No corresponde a un verdadero diccionario debido a que al multiplicar los coeficientes $\mathbf{a}$ por $\mathbf{\Phi'}$ no se obtiene la señal original exacta $\mathbf{x}$. El aspecto de los átomos de este "diccionario" se puede apreciar en la Figura 7.3, y se muestra aquí sólo con propósitos ilustrativos.

Con las representaciones obtenidas mediante las técnicas descriptas se realizaron los experimentos de clasificación de fonemas en inglés, cuyos resultados se muestran en la Tabla A.7. Como puede observarse los mejores resultados corresponden a los casos en escala de mel, en particular para el caso de Fourier (Tabla A.7, exp. Nº 3). Esto obedece, no sólo al hecho de la escala frecuencial adecuada, sino también a la reducción en la cantidad de dimensiones finales. Como referencia para poder comparar con otras representaciones basadas en diccionarios, el caso de Fourier en escala lineal (Tabla A.7, exp. Nº 1) puede resultar útil debido a que posee una mayor cantidad de dimensiones. Para el caso de MFCC se realizaron también pruebas de robustez del clasificador al ruido

[1]Si se toma una emisión completa esta representación corresponde a la STDFT en escala de mel. Sin embargo como el diccionario que aquí se utiliza corresponde estrictamente a un único tramo puede resultar más adecuado asociarlo a la DFT en escala de mel.

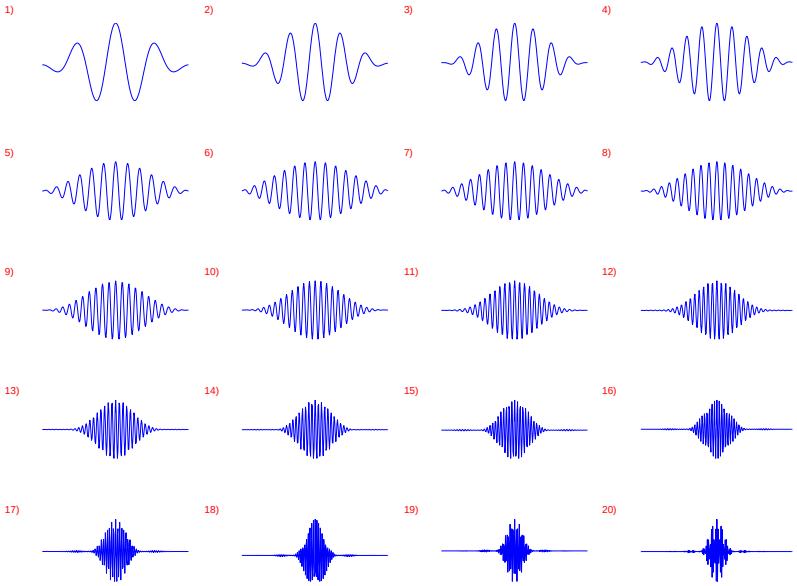

Figura 7.2: *Átomos o columnas del diccionario* $\mathbf{\Phi}$ *de la DFT en escala de mel para* $N = 256$ *(parte real). El valor indicado en la parte superior izquierda de cada átomo corresponde al índice* k *de cada columna. Este diccionario posee menos columnas que las dimensiones del espacio y los átomos están ubicados sólo alrededor de algunas frecuencias de interés y con un ancho de banda que aumenta con dicha frecuencia.*

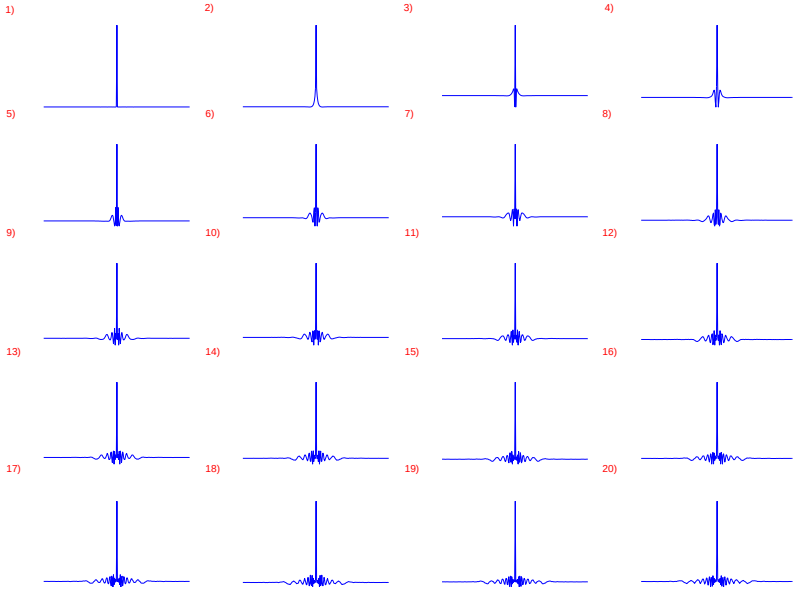

Figura 7.3: *Átomos o columnas del "diccionario" aproximado* $\mathbf{\Phi}'$ *para la transformación que devuelve los coeficientes cepstrales reales en escala de mel (MFCC), para* $N = 256$ *. El valor indicado en la parte superior izquierda de cada átomo corresponde al índice* k *de cada columna. Estos átomos resultan similares a respuestas al impulso de sistemas y poseen características especiales en frecuencia derivadas de la escala de frecuencial no lineal utilizada.*

aditivo cuyos resultados aparecen en la Tabla A.8.

En cuanto a los experimentos de ASR en castellano, estos se corrieron sólo con MFCC (incluyendo también el correspondiente coeficiente de energía y la derivada temporal) por ser la alternativa más utilizada en este tipo de sistemas. Los resultados con diversas variantes del sistema completo se encuentran reportados en el Apéndice B. En la Figura 1.1 del capítulo introductorio se mostraron los resultados de este sistema para ejemplificar la degradación sufrida por el mismo frente a diferentes situaciones de contaminación con ruido.

7.4. Inclusión de cambios de complejidad

Una alternativa bastante directa para mejorar algunas propiedades de las representaciones tradicionales consiste en agregar a éstas información adicional acerca de aspectos no contemplados originalmente. En esta sección se muestra como la adición a los coeficientes cepstrales 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 [175].

Como se ha visto los modelos lineales autoregresivos han sido ampliamente utilizados para modelar la señal de voz. Sin embargo algunos aspectos, como por ejemplo la radiación a nivel de los labios o las turbulencias producidas durante las constricciones, no pueden modelarse adecuadamente mediante un enfoque lineal [195, 193, 7]. El tracto vocal constituye entonces un sistema no lineal cuya dinámica varía en el tiempo de forma continua. Estos cambios de dinámica pueden ser detectados a partir de diferentes medidas de evolución de la complejidad del sistema –como las discutidas en la Sección 2.3.3– inclusive en presencia de ruido.

En la Figura 7.4 se muestra un trozo de una emisión de voz limpia y su versión contaminada con ruido de conversación. La correspondiente evolución de la q-entropía relativa $\mathcal{D}_q$ se muestra también en la figura para cada caso[2]. Es posible apreciar aquí la correspondencia entre las variaciones de $\mathcal{D}_q$ y los cambios fonéticos. Para el caso ruidoso los picos se mantienen en posiciones similares al caso limpio, lo que sugiere cierto grado de robustez de esta medida. Resultados similares se han reportado utilizando la entropía de Shannon en problemas de detección de voz [81]. En otras señales provenientes de sistemas biológicos la robustez se ha incrementado por medio de análisis multiresolución [4].

A partir de estas propiedades de las medidas de complejidad es posible proponer una alternativa sencilla para aumentar la robustez del enfoque clásico. Esta alternativa consiste en agregar información adicional acerca de los cambios de dinámica de la señal de voz a la parametrización tradicional basada en MFCC. Esta información agregada consiste en la adición de un coeficiente que mida la evolución temporal de la complejidad.

Para evaluar esta alternativa los resultados se contrastan con los del Apéndice B para el mismo experimento en presencia de ruido. Con el fin de que la dimensión de los patrones no fuera un elemento que pesara en los resultados se igualaron las dimensiones

[2]Esta evolución de la entropía relativa se calcula mediante un enfoque de análisis por tramos, entre el tramo actual y el anterior de la señal temporal de voz.

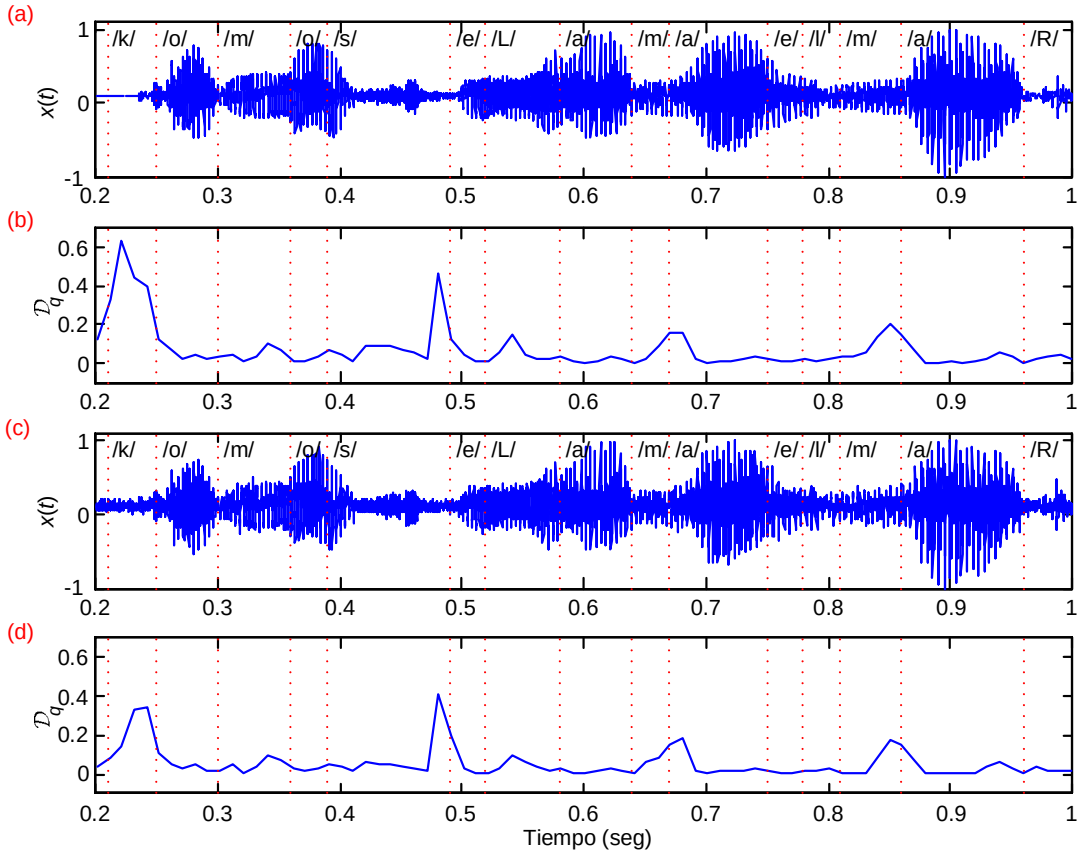

Figura 7.4: *Señal de voz segmentada y etiquetada junto con su correspondiente evolución por tramos para $\mathcal{D}_q$ en el caso limpio (arriba) y contaminado con ruido aditivo blanco a 20 dB SNR (abajo).*

Tabla 7.1: *Porcentaje relativo de mejora del error ($\Delta\epsilon_{\%}$) para diferentes medidas de complejidad comparadas con la referencia para habla contaminada con ruido blanco. Las cifras resaltadas indican el mejor desempeño para cada valor de SNR.*

SNR_{dB}	$\mathcal{H}$	$\mathcal{H}_{q=0,1}$	$\mathcal{H}_{q=0,5}$	$\mathcal{D}$	$\mathcal{D}_{q=0,1}$	$\mathcal{D}_{q=0,5}$
∞	0.89	3.55	2.55	0.96	**8.04**	3.89
50	-1.88	0.09	1.48	-2.01	**7.23**	1.42
25	1.20	-8.57	7.19	5.90	**12.83**	2.38
15	8.60	6.39	7.30	-7.43	**21.25**	2.51
10	13.02	15.79	12.29	8.99	**18.38**	-1.14
5	2.25	**3.20**	3.13	-5.45	-1.91	-8.53
0	-1.26	**0.31**	-0.75	-4.07	-2.44	-3.17

Tabla 7.2: *Porcentaje relativo de mejora del error ($\Delta\epsilon_{\%}$) para diferentes medidas de complejidad comparadas con la referencia para habla contaminada con ruido murmullo. Las cifras resaltadas indican el mejor desempeño para cada valor de SNR.*

SNR_{dB}	$\mathcal{H}$	$\mathcal{H}_{q=0,1}$	$\mathcal{H}_{q=0,5}$	$\mathcal{D}$	$\mathcal{D}_{q=0,1}$	$\mathcal{D}_{q=0,5}$
∞	0.89	3.55	2.55	0.96	**8.04**	3.89
50	1.48	1.68	3.79	1.66	**10.71**	3.92
25	6.46	-1.34	6.84	4.31	**14.11**	1.06
15	17.98	14.50	20.75	17.19	**24.31**	10.13
10	**14.83**	10.42	13.46	4.13	8.16	0.64
5	1.52	**2.67**	2.18	-6.73	-4.53	-11.36
0	-4.35	**-2.00**	-2.77	-8.56	-4.96	-7.25

de ambas experiencias. Para la alternativa propuesta se calcularon 12 MFCC, 1 coeficiente de energía, 1 coeficiente relacionado con una medida de complejidad del tramo correspondiente y sus derivadas temporales. Se consideraron los casos para la entropía de Shannon, las q–entropías y sus correspondientes informaciones relativas.

Para facilitar la comparaciones se calculó la mejora del error relativo para las diferentes medidas utilizadas, comparadas con los resultados de referencia:

$$\Delta\epsilon_{\%} = \frac{\epsilon_{ref} - \epsilon}{\epsilon_{ref}} \times 100,$$

donde ϵ es el porcentaje de errores por palabras.

Los resultados para las diferentes medidas de complejidad y señales contaminadas con ruido blanco y murmullo se muestran en las Tablas 7.1 y 7.2 respectivamente. En ambos casos se han remarcado los mejores resultados para cada valor de SNR. De estas tablas se puede concluir que $\mathcal{D}_{q=0,1}$ es la medida que provee mejores resultados, en particular para SNRs mayores que 10 dB.

Se evaluó también la significancia estadística de estos resultados calculando la probabilidad de que un reconocedor dado sea mejor que el de referencia ($P(\epsilon_{ref} > \epsilon)$). Para realizar esta prueba se supuso la independencia estadística de los errores de reconocimiento para cada palabra y se aproximó la distribución binomial de los errores por medio de una distribución gaussiana. Ésto es posible debido a que se cuenta con un número

suficientemente grande de palabras (11077 si se toman en cuenta todas las particiones). De esta forma, para $q = 0{,}1$ y SNR entre 10 y 25 dB para ambos tipos de ruido se tiene que $P(\epsilon_{ref} > \epsilon) > 99{,}999\,\%$.

7.5. Representaciones basadas en onditas

En lugar del análisis tradicional basado en la transformada de Fourier, que examina una señal a una resolución fija, se ha visto que la transformada onditas posee la característica de hacerlo a distintas escalas (o resoluciones). Esto implica un análisis más "similar" al realizado por el sistema auditivo. Este análisis posee mayor resolución frecuencial de los eventos lentos y mayor resolución temporal de los eventos rápidos. En una variedad de trabajos se han reportado beneficios empleando este tipo de transformación para tareas tales como la compresión y el filtrado de señales [161]. Sin embargo, se ha hecho relativamente poco en materia de clasificación de patrones dinámicos de longitud variable, como es el caso de la clasificación de los fonemas [167]. Ésto representa una tarea diferente debido a la necesidad de procesar un gran número de señales con diferentes características mediante una única familia de onditas. Para el caso de señales muestreadas existe la *transformada onditas discreta diádica* (DDWT), que posee además una implementación rápida de interés en las aplicaciones. A partir de la *transformada paquete de onditas* (WPT) aparece una gama adicional de posibilidades de representación.

A continuación se presentan los resultados de los diferentes experimentos relacionados con las representaciones basadas en onditas y su discusión. Para su realización se emplearon más de 2500 horas de tiempo de máquina[3]. Ésto se debe a que se ensayaron una gama bastante amplia de parámetros de las representaciones y configuración de los clasificadores. Aquí sólo se presentan los resultados más significativos.

7.5.1. Transformada discreta diádica

Existe una gran variedad de familias de onditas disponibles y parámetros ajustables para realizar el análisis mediante la DDWT. Ésto representa una ventaja debido a la flexibilidad que implica poder ajustar la representación a una aplicación particular. Sin embargo ésto también requiere algún criterio para seleccionar entre todas ellas. En la Tabla 7.3 se muestra un resumen de características principales de las familias de onditas utilizadas en las pruebas. Para ajustar los parámetros de cada familia se utilizó el criterio descripto en [167]. Este tiene que ver con la distancia promedio entre los centroides de las diferentes clases consideradas. Posteriormente se procedió a realizar los experimentos de clasificación con cada familia, cuyos resultados se muestran en la Tabla 7.4 [167, 170]. Los coeficientes de la representación se obtuvieron luego de calcular la magnitud en decibeles de la transformada, sin utilizar ningún agrupamiento particular (es decir que en cada tramo aparecen "mezclados" los coeficientes para las diferentes escalas y tiempos). No se utilizó ninguna ventana especial para cada uno de los tramos debido a las propiedades de la transformada para representar adecuadamente los transitorios.

[3]Con computadoras basadas en procesadores tipo Pentium III, 500 MHz de velocidad de reloj.

Tabla 7.3: *Resumen de características principales de las familias de onditas utilizadas en las pruebas.*

Familia	Soporte Compacto	Simetría	Regularidad	Localización	Comentario
Haar	SI	SI	NO	Mala	La más simple
Daubechies	SI	NO	Variable	Variable	Optimiza suavidad
Meyer	NO	SI	SI	Buena	Muy difundida
Vaidyanathan	SI	NO	SI	Buena	Codificación de voz
Splines	SI	SI	Variable	Variable	Biortogonal
Symmlets	SI	NO	Variable	Variable	La menos asimétrica

Tabla 7.4: *Resultados de experimentos de clasificación de fonemas con redes neuronales mediante las representaciones generadas con la DDWT y diferentes familias de onditas. El ancho de la ventana de análisis se mantuvo fijo.*

Nº	Experimento	Estructura Red	TRN	TST	*/b/*	*/d/*	*/jh/*	*/eh/*	*/ih/*
1	Haar (128)	128+128/150/5	53.15	50.00	30.80	8.00	96.50	97.50	3.20
2	Daubechies 16 (128)	128+128/150/5	63.76	63.11	44.90	63.20	67.10	50.60	75.60
3	Meyer (128)	128+128/150/5	65.46	67.58	65.80	61.30	85.20	43.30	84.70
4	Vaidyanathan (128)	128+128/150/5	69.25	68.11	33.30	63.00	81.90	68.90	68.00
5	Splines 9,37 (128)	128+128/150/5	**70.43**	**70.92**	**47.00**	**78.90**	**82.70**	**67.20**	**73.50**
6	Symmlets 10 (128)	128+128/154/5	66.43	64.30	24.32	73.33	93.18	40.45	85.61
1	Fourier (256, 128)	128+128/150/5	**79.67**	**77.53**	**52.60**	**63.60**	**97.20**	**83.60**	**71.70**

Para estas pruebas el mejor resultado es el correspondiente a la familia Splines (Tabla 7.4, exp. Nº 5). En la Figura 7.5 se puede apreciar el diccionario de síntesis correspondiente a esta familia (recuérdese que se trata del caso biortogonal, donde los diccionarios de análisis y síntesis resultan diferentes).

Sin embargo, como se puede apreciar, los resultados para la DDWT y las familias de onditas ensayadas no superan en nigún caso a los obtenidos por la STDFT con patrones de la misma dimensión (Tabla A.7, exp. Nº 1), que se ha reproducido nuevamente al final de la Tabla 7.4 para facilitar la comparación. Desde el punto de vista de las características tiempo-frecuencia de las representaciones logradas esto se puede explicar de la siguiente forma. La STDFT posee una mejor resolución frecuencial en el rango de las frecuencias medias y altas, contrastando con una baja resolución relativa de la DDWT a estas mismas frecuencias (y quizás una "excesiva" resolución temporal en el rango de altas frecuencias). Esto se evidencia por los valores relativamente bajos de la tasa de reconocimiento individual para el caso de las vocales (o en las tasas de confusión más altas reportadas en [167]). En particular la resolución en frecuencia de las bases de la DDWT no alcanza para distinguir las pequeñas diferencias entre las formantes de las vocales elegidas (que se eligieron precisamente con este objetivo, ver Figura 7.6). Algo similar ocurre para el caso de la */jh/*.

El tamaño de la ventana de análisis también puede influir en los resultados. Por ello se corrieron algunas pruebas adicionales con la ondita Symmlets (con 10 momentos nulos) para ver su efecto sobre los resultados. Éstos se reportan en la Tabla 7.5 [170]. Se puede observar que, si bien los resultados mejoran a medida que aumenta el ancho de la ventana, las mejoras más importantes aparecen en los fonemas con componentes transitorias (*/b/*, */d/*). Obsérvese que para el mejor caso (Tabla 7.5, exp. Nº 2) el ancho de la ventana es el mismo que el utilizado en los experimentos con Fourier, pero logrando una representación con el doble de dimensiones. Para este caso podría decirse que el

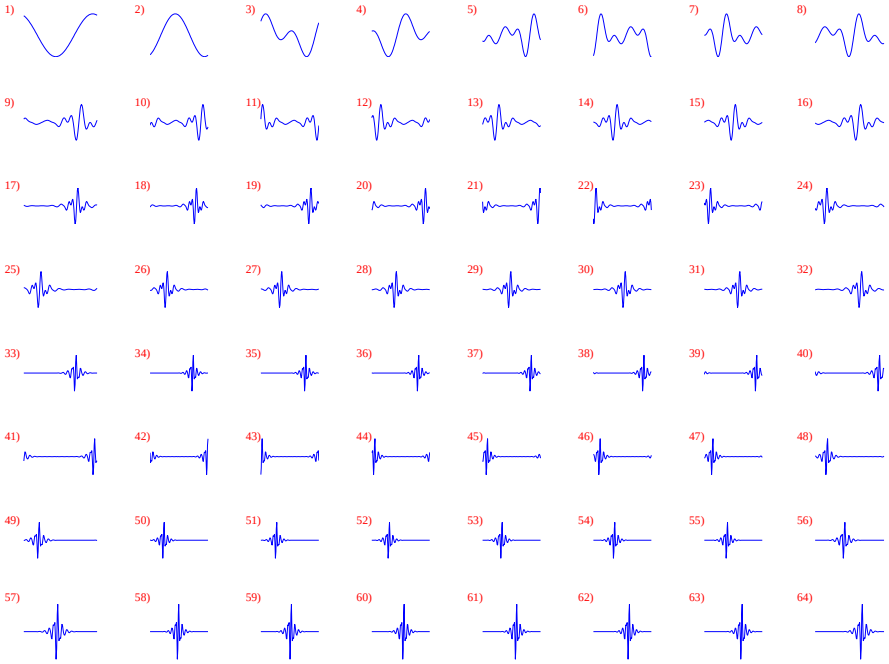

Figura 7.5: *Algunos átomos del diccionario de síntesis de la DDWT para $N = 128$ y la ondita Splines con parámetros 9, 37. Esta familia fue la que obtuvo mejor resultado en los experimentos de clasificación de fonemas.*

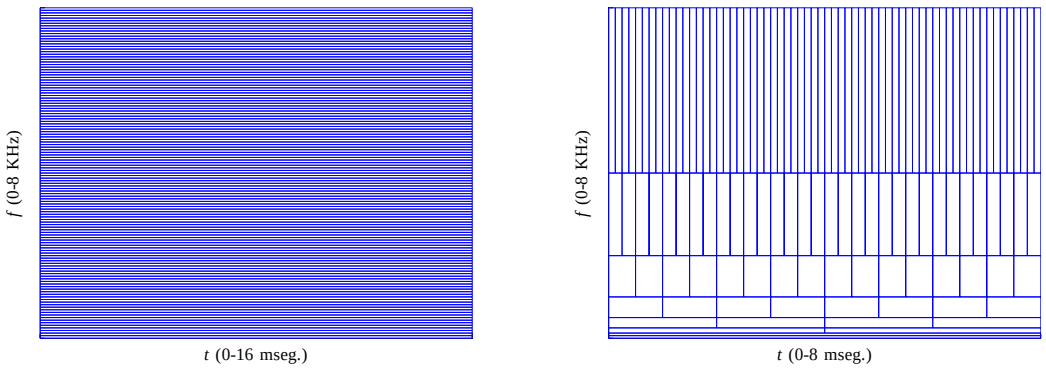

Figura 7.6: *Comparación entre la resolución $t - f$ para un tramo de la STDFT (izquierda) y de la DDWT (derecha), calculada para tramos de 256 y 128 muestras respectivamente y una frecuencia de muestreo de 16 KHz (valores como los de los experimentos reportados). Obsérvese la escasa resolución frecuencial relativa de la DDWT en la zona entre los 500 y 3500 Hz, que constituye un rango de frecuencias muy importante para la discriminación de las vocales del idioma inglés (Ver Figura A.2).*

Tabla 7.5: *Resultados de experimentos de clasificación de fonemas con redes neuronales mediante las representaciones generadas con la DWT, familia de onditas Symmlets con 10 momentos nulos y ancho ventana variable.*

Nº	Experimento	Estructura Red	TRN	TST	/b/	/d/	/jh/	/eh/	/ih/
1	Symmlets 10 (64)	64+64/125/5	61.41	56.14	16.49	46.76	61.74	82.46	36.61
2	Symmlets 10 (256)	256+256/188/5	**73.74**	**69.85**	**63.46**	**86.49**	**76.19**	**75.62**	**61.52**
6	Symmlets 10 (128)	128+128/154/5	66.43	64.30	24.32	73.33	93.18	40.45	85.61
1	Fourier (256, 128)	128+128/150/5	**79.67**	**77.53**	**52.60**	**63.60**	**97.20**	**83.60**	**71.70**

comportamiento es precisamente "inverso" al de Fourier con las vocales ya discutido. Es decir que los resultados sobre los fonemas */b/* y */d/* son mejores para la DDWT; mientras que los resultados sobre */jh/*, */eh/* y */ih/* son mejores para la STDFT. De esta forma se puede decir hasta aquí que la STDFT resultaría más adecuada para la clasificación de los fonemas estables y la DDWT para la de los que poseen comportamiento no estacionario. Esta hipótesis está sustentada también por los fundamentos teóricos de ambas técnicas.

7.5.2. Transformada paquetes de onditas

En los experimentos anteriores se discutieron algunas características de la DDWT que limitan su aplicación al problema considerado. Para intentar solucionarlo se ha desarrollado un método basado en la utilización de la WPT en lugar de la DDWT, que se presentará en esta sección. Aprovechando la mayor flexibilidad de la WPT, el banco de filtros utilizado en esta transformación se diseñó especialmente para tener una resolución en frecuencia más similar a la del oído. El objetivo se logra distribuyendo el ancho de banda de las señales de la base (filtros) de acuerdo a la escala de mel. Esto se apoya también en las mejoras relativas obtenidas en los experimentos anteriores con Fourier y cepstra, al cambiar de la escala frecuencial lineal a la de mel. Se ha denominado a este enfoque como *transformada paquetes de onditas orientadas perceptualmente* (POWPT) [202]. El árbol de filtros diseñado y la partición tiempo frecuencia lograda, para las señales consideradas, se muestran en la Figura 7.7. Los coeficientes de la representación se obtienen luego de calcular la magnitud de la energía en decibeles proveniente de la "integración" de un determinado número de coeficientes adyacentes correspondientes a las distintas escalas (Ver Figura 7.8). Esto obedece a que, según se hizo notar, la resolución temporal en algunas escalas podría resultar excesiva para la discriminación de los fonemas. Además esto disminuye la dimensión de los patrones de la representación, y permite incluir información de mayor duración que ha demostrado mejorar los resultados (Ver Tabla 7.5, exp. Nº 2). En la Figura 7.9 se puede observar la representación obtenida a partir de la POWPT para un fonema de TIMIT.

Los experimentos consistieron en utilizar la "mejor" ondita de los resultados anteriores, Splines (Tabla 7.4 exp. Nº 5), con ancho de ventana variable y diferentes esquemas de integración (directo o sin integración, todos los de una banda, agrupamiento de a 4 coeficientes y de a 8 coeficientes). Para tener otro punto de comparación se repitieron idénticos experimentos pero con la ondita Daubechies 16. Los resultados se muestran en las Tablas 7.6 y 7.7 respectivamente. Como se puede apreciar los mejores resultados son los obtenidos para Daubechies con una ventana de 512 y submuestreo por 4 (Tabla

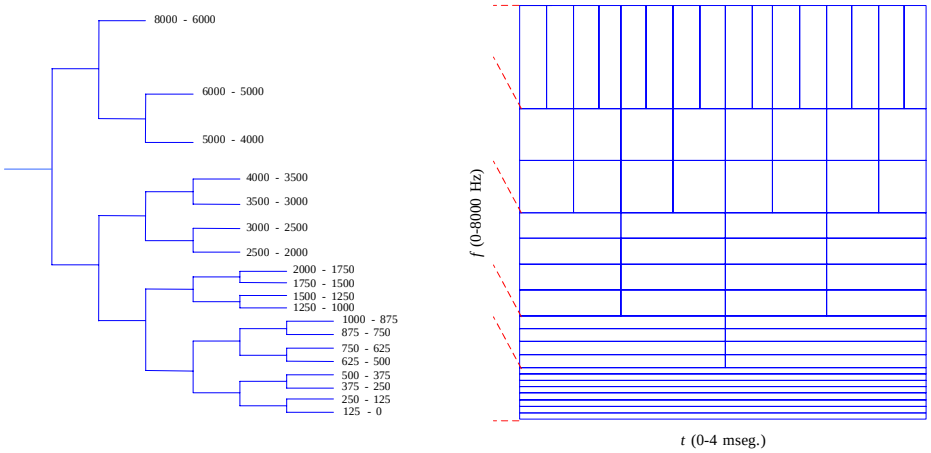

Figura 7.7: *Transformada paquetes de onditas orientadas perceptualmente: Árbol de filtros (izquierda) y partición tiempo frecuencia correspondiente (derecha) para una frecuencia de muestreo de 16 KHz y un ancho de ventana de 64 muestras.*

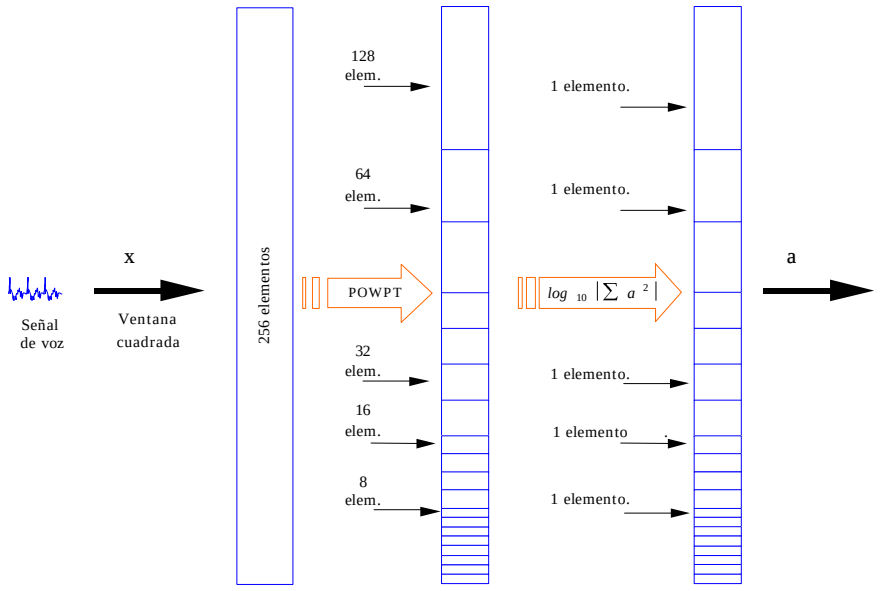

Figura 7.8: *Diagrama del cálculo de los coeficientes de la representación basada en la POWPT para el esquema de integración donde se suma la energía de todos los coeficientes de cada banda. Esto da lugar a 19 coeficientes en total, es decir un coeficiente por cada banda.*

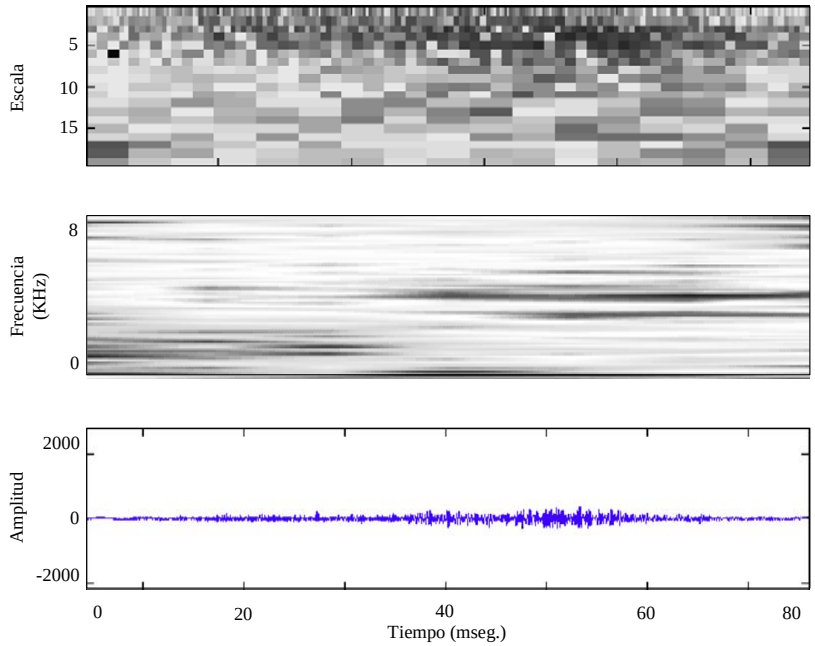

Figura 7.9: *Sonograma (abajo), espectrograma (centro) y transformada paquetes de onditas en escala de mel (Daubechies 16, tramo 256 muestras, esquema de integración directo, arriba), correspondiente al fonema /jh/ de la frase "She had your dark suit in greasy wash water all year" (TIMIT). Comparar con la Figura 7.1.*

7.7, exp. Nº 8). Éstos son sustancialmente mejores que los obtenidos para la DDWT con idénticas onditas madres (7.4, exp. Nº 2 y Nº 5), invirtiendo de hecho las posiciones relativas originales de ambas familias. El aumento en los porcentajes de clasificación se debe a la mejor discriminación de las frecuencias formantes al mejorar la resolución, respecto de la DDWT, para las frecuencias medias (anteriormente de 1 coeficiente por octava). Por ello se puede decir que el cambio en la partición $t-f$ resulta más importante para la discriminación de los fonemas considerados que las características propias de las diferentes familias de onditas. El comportamiento del clasificador es inclusive mejor que el obtenido procesando las señales con la STDFT para igual dimensión de los patrones (Tabla A.7, exp. Nº 1). Ésto ocurre debido a la mejor resolución temporal en las bandas de frecuencia media y alta que permite diferenciar mejor a los fonemas con componentes transitorias. El desempeño es ligeramente inferior a la STDFT en escala de mel (Tabla A.7, exp. Nº 3) que posee muchas menos dimensiones[4].

En las Figura 7.10 se muestran algunos átomos y sus respectivos espectrogramas, correspondientes al diccionario de la POWPT para $N = 256$ y la ondita Daubechies 16 (Tabla 7.7, exp. Nº 3). En la Figura 7.11 se presenta a modo de resumen un gráfico comparativo de los resultados de clasificación de fonemas para las distintas representaciones

[4]Para evaluar la influencia sobre los resultados del clasificador empleado se corrieron también algunos experimentos utilizando los mismos datos y representaciones, pero entrenando una única red neuronal con retardos por cada uno de los fonemas. Este enfoque resulta similar al empleado en los HMM, donde se entrena un "modelo" separado para cada fonema, y de hecho mejoró notablemente los resultados. Éstos resultados fueron reportados en [172] y no se presentan aquí por razones de espacio.

Tabla 7.6: *Resultados de experimentos de clasificación de fonemas con redes neuronales y las representaciones generadas mediante la POWPT, familia de onditas Splines 9, 37, ancho ventana variable y diferentes esquemas de integración.*

Nº	Experimento	Estructura Red	TRN	TST	/b/	/d/	/jh/	/eh/	/ih/
1	Splines 9,37 (64, directo)	64+64/160/5	47.99	47.42	8.10	50.00	62.79	68.13	29.72
2	Splines 9,37 (128, directo)	128+128/160/5	66.61	64.77	35.00	55.97	43.92	73.88	64.07
3	Splines 9,37 (256, directo)	256+256/300/5	73.00	63.80	63.49	54.55	50.00	75.00	57.14
4	Splines 9,37 (64, todos)	19+19/138/5	60.83	60.63	14.02	66.19	63.28	49.23	74.78
5	Splines 9,37 (128, todos)	19+19/100/5	66.50	64.05	31.63	65.61	65.10	70.74	60.75
6	Splines 9,37 (256, todos)	19+19/100/5	71.15	66.67	33.33	65.71	78.38	66.75	69.72
7	Splines 9,37 (256, de a 4)	64+64/95/5	74.66	70.89	45.65	75.38	87.80	72.75	69.56
8	Splines 9,37 (512, de a 4)	128+128/110/5	75.37	72.95	74.58	74.39	73.33	64.40	18.50
9	Splines 9,37 (512, de a 8)	64+64/78/5	**80.29**	**74.59**	**79.66**	**79.75**	**83.33**	**74.85**	**68.94**
5	Splines 9,37 (128)	128+128/150/5	**70.43**	**70.92**	**47.00**	**78.90**	**82.70**	**67.20**	**73.50**
1	Fourier (256, 128)	128+128/150/5	**79.67**	**77.53**	**52.60**	**63.60**	**97.20**	**83.60**	**71.70**
3	Mel Fourier (256, 20)	20+20+20/135/5	**82.56**	**81.83**	**72.95**	**79.28**	**90.19**	**84.74**	**78.33**

Tabla 7.7: *Resultados de experimentos de clasificación de fonemas con redes neuronales y las representaciones generadas mediante la POWPT, familia de onditas Daubechies 16, ancho ventana variable y diferentes esquemas de integración.*

Nº	Experimento	Estructura Red	TRN	TST	/b/	/d/	/jh/	/eh/	/ih/
1	Daubechies 16 (64, directo)	64+64/160/5	68.72	60.12	27.78	51.30	89.62	35.86	82.42
2	Daubechies 16 (128, directo)	128+128/200/5	66.99	63.91	32.63	70.44	66.44	50.56	77.88
3	Daubechies 16 (256, directo)	256+256/160/5	79.17	73.61	54.72	76.06	78.57	63.73	84.12
4	Daubechies 16 (64, todos)	19+19/138/5	63.02	60.61	38.54	61.67	77.84	25.83	93.26
5	Daubechies 16 (128, todos)	19+19/100/5	71.82	69.21	23.26	81.69	88.24	70.44	67.38
6	Daubechies 16 (256, todos)	19+19/100/5	68.76	67.79	38.18	34.85	92.31	66.99	75.17
7	Daubechies 16 (256, de a 4)	64+64/95/5	68.02	65.61	49.21	87.78	78.26	62.97	64.04
8	Daubechies 16 (512, de a 4)	128+128/110/5	**82.39**	**79.49**	**67.80**	**85.71**	**94.74**	**81.53**	**76.23**
9	Daubechies 16 (512, de a 8)	64+64/78/5	79.33	74.30	85.96	75.32	89.47	81.61	60.82
2	Daubechies 16 (128)	128+128/150/5	63.76	63.11	44.90	63.20	67.10	50.60	75.60
1	Fourier (256, 128)	128+128/150/5	**79.67**	**77.53**	**52.60**	**63.60**	**97.20**	**83.60**	**71.70**
3	Mel Fourier (256, 20)	20+20+20/135/5	**82.56**	**81.83**	**72.95**	**79.28**	**90.19**	**84.74**	**78.33**

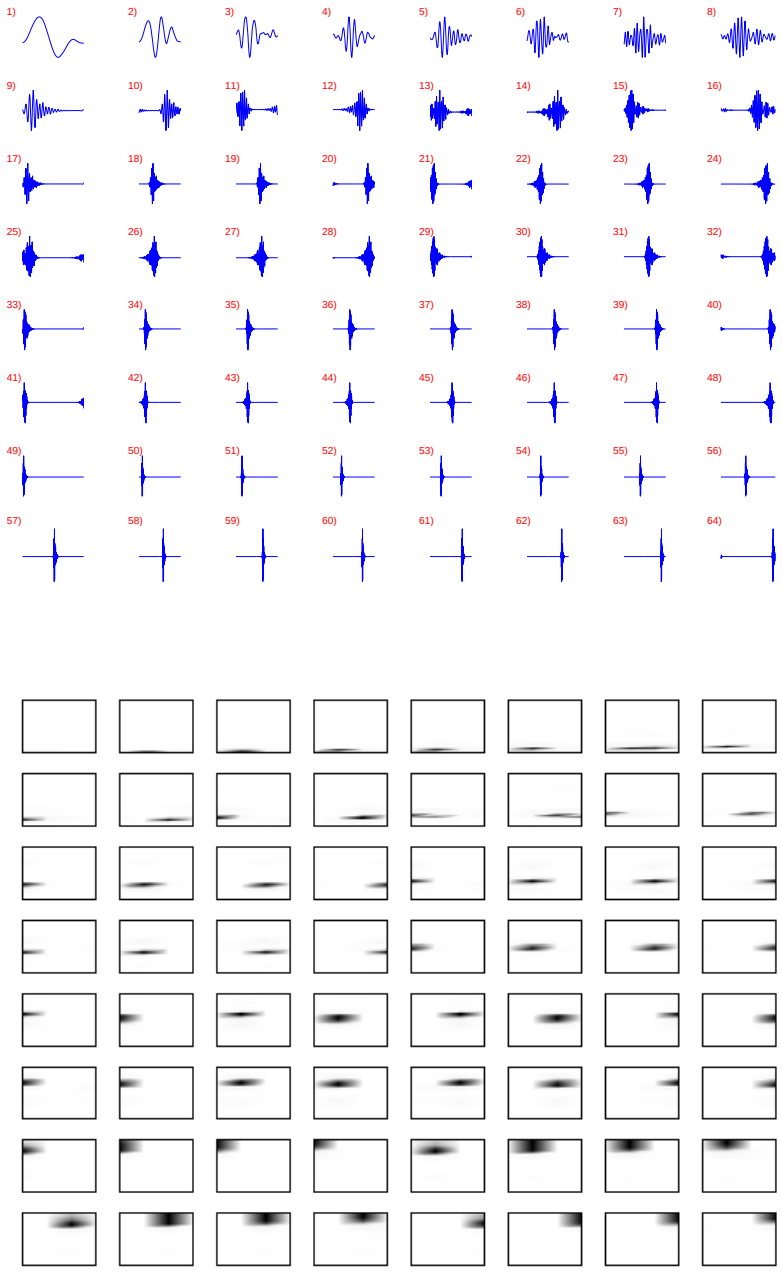

Figura 7.10: *Algunos átomos del diccionario de la POWPT para $N = 256$ y la ondita Daubechies 16 (arriba) y sus correspondientes espectrogramas (abajo). Esta familia fue la que obtuvo mejor resultado en los experimentos de clasificación de fonemas. Se observan varios átomos asimétricos similares a trozos de fonemas sonoros junto con otros más parecidos a impulsos. Esto se manifiesta mediante comportamientos que van desde muy localizados en la frecuencia hasta muy localizados en el tiempo.*

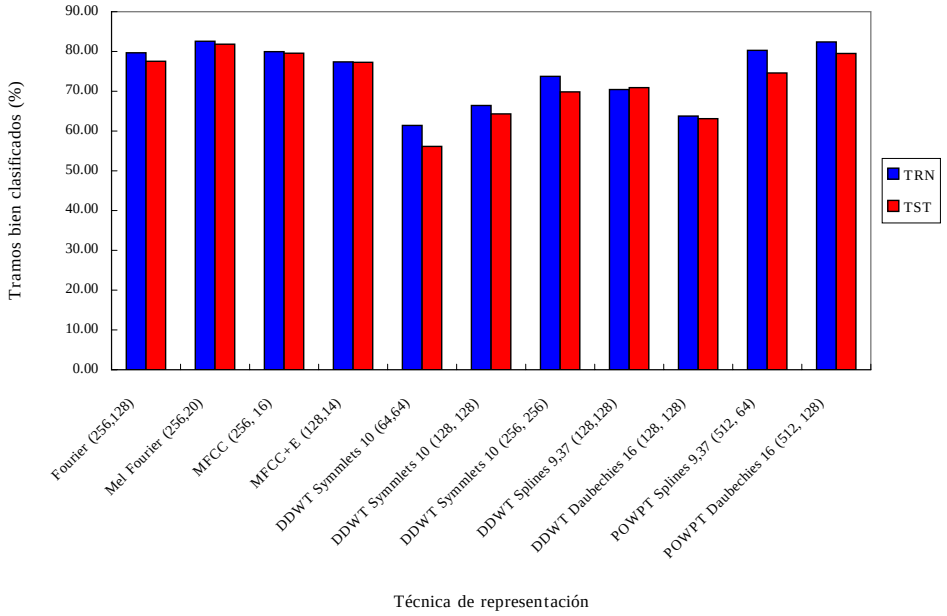

Figura 7.11: *Resumen comparativo de la tasa de tramos bien clasificados para las diferentes representaciones convencionales y basadas en onditas.*

basadas en DDWT y WPT, junto con los de las técnicas convencionales.

Por último se utilizó la representación obtenida mediante POWPT en experimentos con el sistema de ASR, con habla limpia y contaminada con ruido aditivo (Apéndice B). Sin embargo el desempeño relativo del sistema entrenado mediante POWPT respecto al de MFCC+E+Δ no resultó tan bueno para esta aplicación. Se obtuvo del orden de un 10 % menos de reconocimiento a nivel de palabras (absoluto) para los distintos niveles y tipos de ruido (blanco y murmullo).

Un dato importante a señalar, en el caso de POWPT, es que en ciertas particiones y para determinadas condiciones de ruido los HMM no generaban ninguna transcripción. Ésto podría implicar probabilidades demasiado bajas que hubieran desbordado la precisión numérica de los tipos de datos, o caído debajo del umbral mínimo. Se ensayaron distintos tipos de normalización de los datos para intentar solucionar este problema pero no se encontraron cambios significativos en los resultados. Debido a que en los experimentos iniciales se utilizaron matrices de covarianza diagonal, se rediseñaron los modelos para el caso más general de una matriz de covarianza completa. Al entrenar los modelos, estas matrices resultaron ser no inversibles, lo que indicaría que la cantidad de datos de entrenamiento resulta pequeña comparada con la complejidad del modelo o bien existe alguna dependencia lineal entre los coeficientes de la representación. Se ensayaron algunos métodos para "decorrelacionar" los coeficientes pero no parecieron solucionar el inconveniente. Se sabe también que las representaciones basadas en onditas son bastante más ralas que las otras representaciones empleadas [31]. El modelado por medio de mezclas de gaussianas de funciones de distribución de probabilidad con curtosis positivo requiere la utilización de un gran número de gaussianas para lograr una precisión adecuada. Esto lleva a repensar algunos de los fundamentos del modelo utilizado para el

reconocimiento en términos de ésta nueva representación, lo que queda fuera del alcance del presente trabajo.

Hasta aquí se han presentado y revisado distintas alternativas para obtener una representación óptima basada en onditas, principalmente aprovechando la flexibilidad de la WPT para incorporar aspectos psicoacústicos. Un inconveniente con algunas de las representaciones así generadas es que la transformación no resulta invariante al desplazamiento[5]. Otra posibilidad consiste en utilizar la *transformada ondita continua muestreada* (CSTW), muestreando los coeficientes en escala de mel. Esto mejora también la resolución en frecuencia en las bandas críticas para el habla con respecto al caso diádico y se han reportado mejores resultados que para la transformada de Fourier en experimentos de clasificación de fonemas [47]. El problema principal de este enfoque es que requiere bastante más tiempo de cálculo que el algoritmo rápido utilizado en los experimentos anteriores.

7.6. Representaciones ralas y/o independientes

En la sección anterior se analizaron varias representaciones basadas en onditas. Otra alternativa sería la utilización de una transformación "adaptable". La representación obtenida mediante este tipo de transformación se adaptaría (idealmente) para realizar el mejor análisis posible en términos de las características más significativas de la señal a analizar. Como se discutió anteriormente los métodos orientados a generar representaciones ralas poseen este tipo de comportamiento. Es posible plantear una discusión respecto a los resultados de la Sección 7.5 en términos de cuál resulta ser la base o diccionario más adecuado para representar a la señal de voz a los fines de su clasificación fonética o reconocimiento. Se podría realizar la pregunta acerca de si las representaciones basadas en Fourier u onditas resultaron mejores. Como se ha visto no hay una respuesta definitiva al respecto. Se verá a continuación como la utilización del enfoque basado en la utilización de representaciones ralas e independientes permite utilizar grandes diccionarios que incluyan simultáneamente elementos con características estables y transitorias. Es posible inclusive encontrar el diccionario óptimo a partir de las señales de voz.

Para comenzar se presentarán y estudiarán las representaciones de la señal de voz obtenidas de diccionarios diseñados a medida de antemano. Además se desarrollarán métodos específicos para aprovechar las características de las representaciones logradas en los casos considerados. Posteriormente se hará algo similar para las representaciones obtenidas a partir de diccionarios óptimos encontrados por ajuste automático.

[5]Ésto puede producir que se codifiquen en los coeficientes aspectos relacionados con la fase relativa de la señal respecto a la ventana de selección de cada tramo. La codificación de la fase de esta forma puede "confundir" al clasificador con aspectos que no resultan directamente relevantes para la tarea de clasificación o reconocimiento a nivel fonético. Se retomará la discusión acerca de este aspecto más adelante.

7.6.1. Diccionarios a medida

Criterios de diseño

Antes de seleccionar un diccionario particular resulta interesante plantear algunos criterios generales que orienten la búsqueda hacia algún subconjunto, dentro de todos los diccionarios posibles. Nuevamente el objetivo consiste en lograr una "buena" representación de la señal de voz mediante estos diccionarios. Un aspecto a considerar podría ser el de lograr cierta invariancia a la translación temporal de la representación obtenida. Como se mencionó anteriormente, ésto resulta importante a los efectos de evitar que cambios en la fase produzcan representaciones diferentes. Sin embargo la invariancia a la translación no depende sólo del dicionario empleado, sino también de otros factores que pueden tratarse por separado.

Los diccionarios seleccionados para las pruebas deberían contemplar las siguientes características:

- Adecuada cobertura del plano tiempo-frecuencia en diferentes formas[6].
- Representación de eventos a diferentes escalas temporales, desde pequeñas a grandes.
- Inclusión de cierta asimetría en los átomos que refleje la presente en el tipo de eventos a detectar.

El primer aspecto permite lograr una adecuada flexibilidad en los elementos utilizados para la representación. Esto se debe a que la señal de voz presenta comportamientos altamente transitorios junto con trozos relativamente estacionarios.

El segundo aspecto es deseable a fin de capturar la dinámica a diferentes escalas, que pueden ir desde el nivel subfonémico al suprasegmental. Sin embargo la utilización de un diccionario demasiado grande tiene inconvenientes de índole práctica relacionados con el tiempo necesario para encontrar una representación (Ver Capítulo 6).

El último aspecto mencionado se relaciona con algunos comentarios realizados en [61] donde se propone el uso de diccionarios de sinusoides amortiguadas (asimétricos) para reemplazar o complementar a los del tipo Gabor (simétricos). Ésto aporta algunas funciones no simétricas que representan mejor el tipo de señales que se encuentra por ejemplo en las vocales. De otra forma se requieren varios átomos simétricos para representar un trozo de señal asimétrico, produciendo representaciones menos ralas.

Teniendo en cuenta estos aspectos se han seleccionado para las pruebas los diccionarios basados en las familias de paquetes de onditas y paquetes de cosenos (con profundidad del árbol de filtros adecuada para incluir una variedad importante de comportamientos, o sea varias veces sobrecompletos.).

[6]Es decir que no sólo se requiere la cobertura del plano sino también superponer distintos tipos de estructuras, como por ejemplo mediante tonos, deltas y átomos más localizados simultáneamente.

Datos para estas pruebas

En esta sección se mostrarán varias propiedades de las representaciones ralas para el caso de análisis y descomposición de distintos fonemas del habla inglesa. Sólo se incluirán aquí los resultados derivados de utilizar diccionarios fijos y los métodos determinísticos para encontrar los coeficientes. Los experimentos se llevaron a cabo utilizando el conjunto de fonemas: */eh/*, */ih/*, */b/*, */d/*, */p/*, */t/*, */f/*, */s/*, correspondientes al hablante `\timit\train\dr1\fcjf0\` de la base de datos TIMIT. Este conjunto de fonemas es un poco diferente al utilizado en los experimentos de clasificación del Apéndice A, eliminándose */jh/* y agregándose algunos fonemas representativos de plosivas sordas y fricativas. Cada fonema se extrajo de acuerdo a las etiquetas fonéticas correspondientes, y su longitud fue ajustada para igualar la cantidad de muestras a la potencia de 2 más cercana, como lo requerían los algoritmos utilizados.

Evaluación de los diccionarios

Para decidir acerca de cuáles resultan los mejores diccionarios para representar la señal de voz se utilizaron las representaciones obtenidas mediante BP y MP. En este contexto se utilizaron algunos de los criterios mencionados en la Sección 2.4.2 para evaluar la calidad de la representación lograda mediante estos diccionarios. El primer criterio consistió en estimar la norma ℓ_0 de la representación de los fonemas seleccionados para diferentes diccionarios tipo WPT y CPT (se supuso como diferentes de cero a los coeficientes que superaban un umbral del 5 % del máximo valor absoluto de la representación). Los diccionarios utilizados poseían profundidad del árbol de filtros de 11 o 12 dependiendo de la longitud de las señales, lo que corresponde a un total de 11264 o 24576 elementos respectivamente.

Los resultados se muestran en las Figuras 7.12 y 7.13, para BP y MP respectivamente. Como puede apreciarse fácilmente en ambos gráficos los fonemas sonoros logran representaciones mucho más ralas con ambos métodos en forma relativamente independiente del diccionario utilizado. Ésto quiere decir que los fonemas plosivos y fricativos requieren un mayor número de elementos del diccionario para representarse adecuadamente. Para los fonemas sonoros también puede observase que los diccionarios con mayor número de momentos nulos dan representaciones más ralas, lo que se debe al mayor parecido de los átomos con este tipo de fonemas. Sin embargo no se observa una relación similar para el resto de los fonemas.

Ambos métodos, BP y MP, logran representaciones similarmente ralas para los fonemas considerados. Debe aclararse aquí que para el cálculo de MP el número de iteraciones se fijo de antemano en 1000, lo que fija también un techo para la dispersión debido a que los diccionarios cuentan con varios miles de átomos.

Luego se calculó el promedio de la norma ℓ_0 para todos los fonemas considerados y para cada uno de los métodos. Los resultados se muestran en la Figura 7.14. Se observa en la figura que el diccionario más ralo en promedio para BP es WPT Daubechies 12, mientras que con respecto a MP el más ralo es el de CPT 5. Un caso intermedio lo constituye el de WPT Symmlets 8.

En las pruebas anteriores se tuvo en cuenta sólo la dispersión de la representación

lograda. Sin embargo se requiere también analizar cuán bien representa a la señal cada uno de los diccionarios considerados. Para ello se calcularon los errores de aproximación utilizando sólo los 15 átomos más importantes de cada diccionario seleccionados por BP o MP (a partir del valor absoluto de los coeficientes respectivos). Esta cantidad de átomos representa entre el 0.06 % al 0.13 % del total de elementos del diccionario, es decir muchos menos elementos que los seleccionados en promedio por ambos métodos.

Los resultados se muestran en las Figuras 7.15 y 7.16. Como puede observarse en las figuras la capacidad de aproximación mediante BP y MP a partir de muy pocos elementos de los diccionarios varía para los distintos fonemas. A pesar de que las vocales poseían la representación más rala en las pruebas anteriores (Figuras 7.12 y 7.13) ahora poseen errores de aproximación mayores para esta cantidad tan pequeña de átomos. Esto se debe a la mayor energía de estos fonemas ya a que, a pesar de que los rasgos morfológicos importantes tienden a conservarse, los residuos no resultan de pequeña energía. Si se tiene en cuenta este efecto, de la diferencia de energía relativa, los errores resultan "mayores" para las consonantes, principalmente para las fricativas.

A continuación se procedió a calcular el promedio del MSE para todos los fonemas considerados y para cada uno de los métodos. Los resultados se muestran en la Figura 7.17. Se observa en la figura que el diccionario que aproxima mejor en promedio para ambos métodos es CPT 5, mientras que un caso intermedio lo constituye el de WPT Symmlets 8.

Inferencia mediante BP y MP

Como se mencionó en el capítulo anterior no parece haberse realizado un estudio sistemático de las representaciones logradas por BP y el resto de las técnicas relacionadas aplicada a datos del mundo real. Para comparar la dispersión de la representación obtenida mediante los distintos métodos se utilizó ahora el mismo diccionario sobrecompleto para todos los experimentos. Éste consistió en un diccionario tipo WPT basado en la ondita Symmlets con 8 momentos nulos (de profundidad 11 o 12 dependiendo de la longitud de las señales). La elección del diccionario se realizó a partir de las pruebas descriptas en la sección anterior debido a que las representaciones logradas con este diccionario mostraron un adecuado compromiso entre el porcentaje de coeficientes diferentes de cero y la capacidad de aproximación. Se aplicaron BP, MP, BOB y MOF a las señales extraídas de todos los fonemas.

Para ilustrar algunos conceptos se utilizará como ejemplo la señal correspondiente al fonema */p/*, aunque se obtuvieron resultados similares para el resto de los fonemas [171]. En la Figura 7.18 se muestra el sonograma de este fonema junto con su correspondiente espectrograma y un gráfico de la magnitud de los coeficientes de la representación. El espectrograma mostrado es de banda angosta lo que permite resolver adecuadamente la frecuencias de las secciones cuasi-estacionarias correspondientes a los fonemas vecinos. Sin embargo, puede observarse claramente como los eventos temporales importantes para la discriminación de este fonema, como por ejemplo el momento de la explosión, se "diluyen" completamente[7]. También puede apreciarse en la parte inferior de la gráfica

[7]Este efecto de "dilución" puede haberse acentuado en la gráfica debido a la interpolación bidimen-

que la representación lograda de esta forma no resulta rala debido a que prácticamente no existen coeficientes nulos.

En las Figuras 7.19, 7.20, 7.21 y 7.22 pueden observarse la representaciones tiempo-frecuencia del mismo fonema */p/* logradas por BP, MP, BOB y MOF respectivamente. Aquí puede observarse como BP, MP e inclusive BOB logran representaciones bastante ralas. Obsérvese por ejemplo la representación obtenida mediante BP (Figura 7.19), donde la porción inicial de la señal de naturaleza "cuasi-senoidal" ha sido detectada perfectamente a través del trazo horizontal correspondiente en el plano $t - f$. Asímismo los eventos temporales quedan también perfectamente descriptos sin pérdida de su localización. De esta manera se evidencia el comportamiento adaptativo al que se hizo referencia anteriormente, utilizando en la representación sólo aquellos elementos que mejor describen a la señal. Mediante este comportamiento la resolución tiempo-frecuencia en cada sector del plano se adapta en función de las características de la señal analizada. Se debe hacer notar nuevamente que MP fue utilizado con una opción para seleccionar solo los primeros 1000 coeficientes y ésto impone ciertas restricciones de dispersión sobre los resultados reportados con este método. Otro aspecto que debe mencionarse es que el comportamiento de las técnicas depende del diccionario particular seleccionado (o aprendido). En este caso se trata de un diccionario altamente sobrecompleto.

Resulta también interesante analizar algunos aspectos relacionados con la robustez de la representación. En la Tabla 7.8 se muestran los valores correspondientes a la norma ℓ_0 de los coeficientes de la representación para los diferentes métodos de inferencia, para el caso de señales limpias y contaminadas con ruido blanco a 10 dB SNR. En esta tabla puede verse como en todos los casos MOF es la que provee la representación menos rala. BP y MP proveen representaciones suficientemente ralas de los fonemas, con una adecuada localización de las pistas acústicas (ver también nuevamente Figuras 7.19, 7.20, 7.21 y 7.22). Los fonemas fricativos son los que aparecen como menos ralos, debido a que se requieren más elementos para describir sus características en términos del diccionario empleado, debido a su naturaleza "cuasi-ruidosa" y de banda ancha (Ver Figura 7.23). Para el caso de las señales contaminadas con ruido las representaciones logradas son un poco menos ralas, debido a que se utilizan algunos átomos también para describir el ruido.

En la Tabla 7.9 se muestran los resultados del error cuadrático medio (MSE) obtenido para cada fonema luego de aplicar cada método (nuevamente en los casos limpio y con ruido blanco a 10 dB SNR) para seleccionar los 15 átomos más significativos (como en las Figuras 7.24 y 7.25). Desde el punto vista de la exactitud de la aproximación, BP, MP y BOB resultan comparables. En los resultados del MSE mostrados en esta tabla puede apreciarse como los errores más grandes se cometen en la vocales. Esto se debe en realidad, como ya se ha explicado, a que son los fonemas que poseen mayor energía relativa. Para el resto de los fonemas, los fricativos poseen el mayor error de aproximación. También puede apreciarse aquí como el MSE disminuye un poco en el caso ruidoso, debido a que el error de reconstrucción se calcula sobre la señal ruidosa. El resto de las características son similares a las del caso limpio.

sional utilizada para representar el espectrograma.

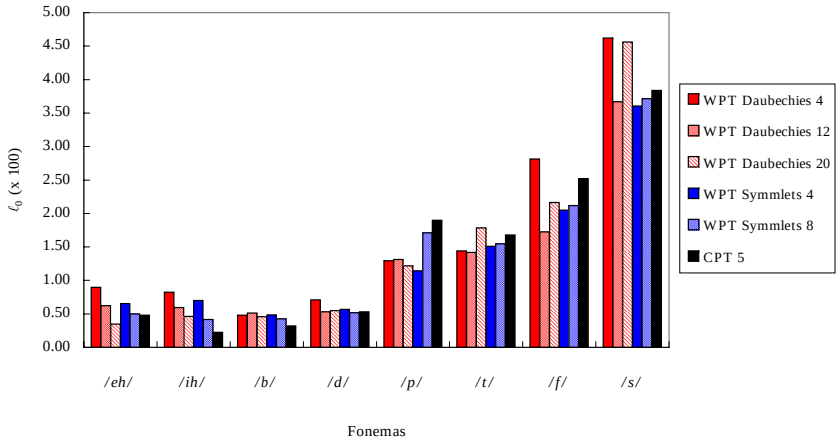

Figura 7.12: *Norma ℓ_0 de las representaciones de distintos fonemas obtenidas mediante BP a partir de diccionarios WPT y CPT.*

Tabla 7.8: *Norma ℓ_0 (x 100) de las representaciones de distintos fonemas obtenidas mediante BP, MP, BOB y MOF a partir del diccionario WPT Symmlets 8 (con profundidad 11 o 12), para señales limpias y contaminadas con ruido blanco a 10 dB SNR.*

SNR (dB)	Método	*/eh/*	*/ih/*	*/b/*	*/d/*	*/p/*	*/t/*	*/f/*	*/s/*
∞	BP	**0.501**	**0.417**	0.423	0.515	1.713	1.546	2.120	3.715
	MP	0.529	0.586	**0.370**	**0.497**	**1.651**	**1.306**	**1.851**	**2.897**
	BOB	0.574	0.559	0.533	0.630	2.228	1.758	2.201	3.984
	MOF	14.880	14.190	11.690	20.810	32.420	27.640	34.230	59.770
10	BP	**0.968**	**0.746**	1.322	**0.968**	2.752	2.987	2.779	4.761
	MP	1.363	1.252	**1.180**	1.696	**2.717**	**2.462**	**2.673**	**3.149**
	BOB	1.388	1.270	1.339	1.802	4.528	3.296	3.316	4.525
	MOF	22.970	20.790	19.660	30.270	48.020	45.730	50.650	67.320

Tabla 7.9: *Error cuadrático medio entre la señal original, correspondiente a distintos fonemas, y la aproximación obtenida a partir de conservar los 15 átomos más importantes seleccionados mediante BP, MP, BOB y MOF a partir del diccionario WPT Symmlets 8 (con profundidad 11 o 12), para señales limpias y contaminadas con ruido blanco a 10 dB SNR.*

SNR (dB)	Método	*/eh/*	*/ih/*	*/b/*	*/d/*	*/p/*	*/t/*	*/f/*	*/s/*
∞	BP	2.09E-03	3.67E-03	6.70E-04	7.52E-04	1.10E-03	1.74E-03	2.13E-03	2.26E-03
	MP	1.84E-03	3.28E-03	5.40E-04	6.38E-04	9.84E-04	1.63E-03	2.00E-03	2.20E-03
	BOB	1.84E-03	3.31E-03	6.37E-04	7.00E-04	1.05E-03	1.66E-03	2.04E-03	2.24E-03
	MOF	3.13E-03	6.09E-03	1.05E-03	1.49E-03	1.47E-03	2.04E-03	2.48E-03	2.45E-03
10	BP	1.42E-03	2.73E-03	4.69E-04	5.96E-04	7.26E-04	1.05E-03	1.28E-03	1.33E-03
	MP	1.35E-03	2.62E-03	4.30E-04	5.68E-04	6.71E-04	9.99E-04	1.23E-03	1.29E-03
	BOB	1.35E-03	2.65E-03	4.59E-04	5.88E-04	7.09E-04	1.02E-03	1.25E-03	1.31E-03
	MOF	1.87E-03	3.74E-03	6.30E-04	9.07E-04	8.72E-04	1.17E-03	1.44E-03	1.41E-03

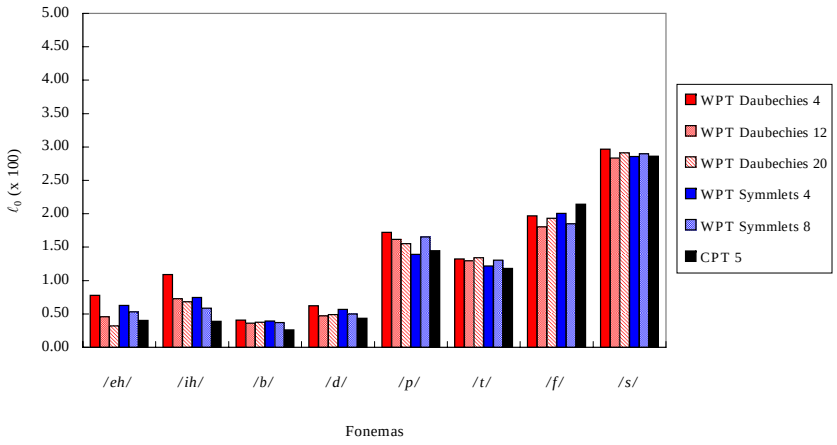

Figura 7.13: *Norma ℓ_0 de las representaciones de distintos fonemas obtenidas mediante MP a partir de diccionarios WPT y CPT.*

Para comprender mejor algunas características de la representación lograda, en la Figuras 7.24 y 7.25 se ha procedido a reconstruir la señal del fonema */p/*, para los casos limpio y con ruido respectivamente, utilizando sólo los 15 átomos más importantes (de un total de 11264). En la parte inferior de las figuras se pueden apreciar los átomos seleccionados por BP para realizar la síntesis en ambos casos. Nótese como los átomos empleados para el caso limpio y ruidoso resultan muy similares, variando sólo ligeramente el orden de importancia. Esto muestra que la representación obtenida logra preservar las características significativas aún en la presencia de ruido. Los métodos tradicionales para limpieza de ruido generalmente fallan en preservar algunas componentes importantes. En el caso del habla esto es de fundamental importancia para evitar artefactos que afecten la inteligibilidad de la misma. Estas propiedades se aprovechan en la sección siguiente para proponer un método heurístico de limpieza de ruido que preserva las pistas acústicas de la señal de voz.

Limpieza de ruido heurística

Las pruebas anteriores de representación rala de fonemas sugieren que las pistas acústico-fonéticas importantes pueden conservarse con tan sólo 15 átomos (de un total de más de 10000). Cuando se agregó ruido aditivo a las señales consideradas, prácticamente fueron seleccionados los mismos átomos (aunque en un orden diferente). En el filtrado tradicional basado en Fourier se asume que el espectro de la señal tiene poca superposición con el espectro del ruido y por consiguiente puede utilizarse un filtro lineal invariante en el tiempo. Esta aproximación de filtrado lineal no puede separar el ruido de la señal en las zonas dónde sus espectros de Fourier solapan. Además los métodos lineales tradicionales establecen un compromiso entre la supresión del ruido y un suavizado de las características de la señal. Las aproximaciones de limpieza de ruido basadas en umbrales son bastante diferentes debido a que resultan no lineales. Originalmente estas técnicas se aplicaron principalmente en el contexto de la DDWT, sin embargo para el caso ralo la aproximación resulta más general porque no se restringe el diccionario a una única

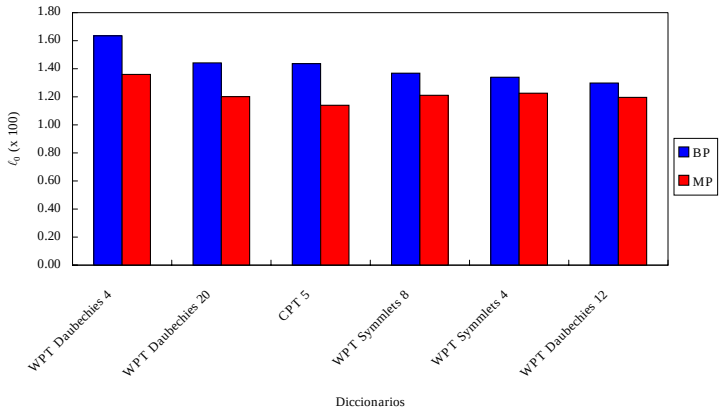

Figura 7.14: *Norma ℓ_0 promedio de las representaciones obtenidas para los diferentes fonemas mediante BP y MP a partir de diccionarios WPT y CPT.*

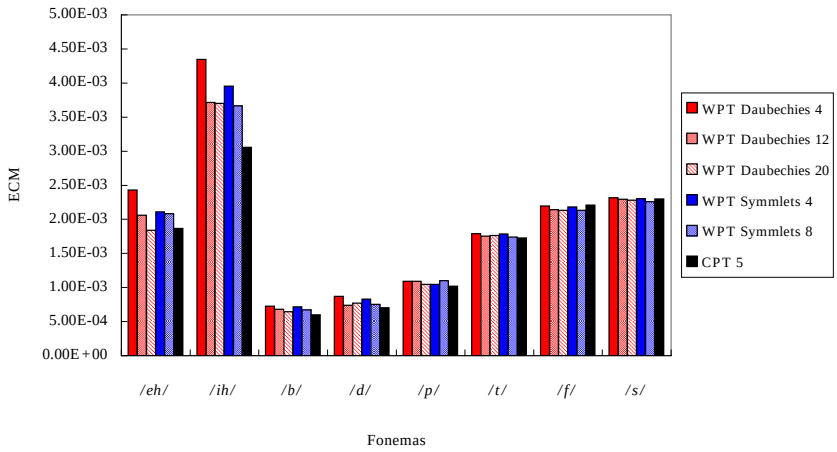

Figura 7.15: *Error cuadrático medio entre la señal original, correspondiente a distintos fonemas, y la aproximación obtenida a partir de conservar los 15 átomos más importantes seleccionados por BP a partir de diccionarios WPT y CPT.*

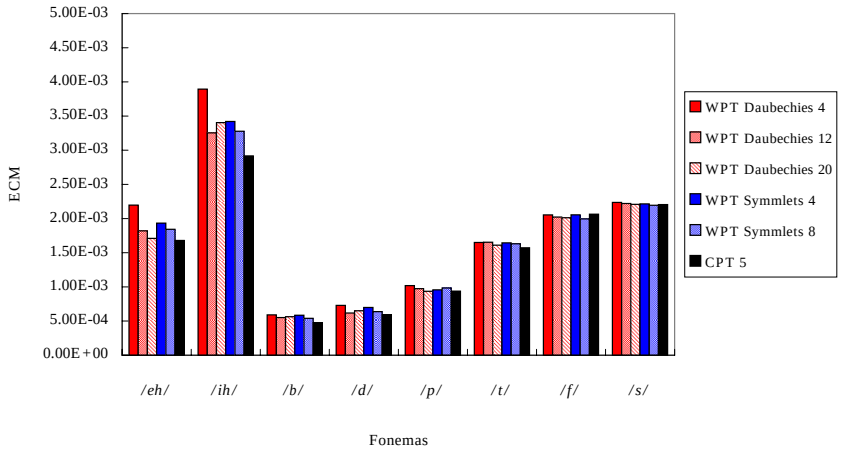

Figura 7.16: *Error cuadrático medio entre la señal original, correspondiente a distintos fonemas, y la aproximación obtenida a partir de conservar los 15 átomos más importantes seleccionados por MP a partir de diccionarios WPT y CPT.*

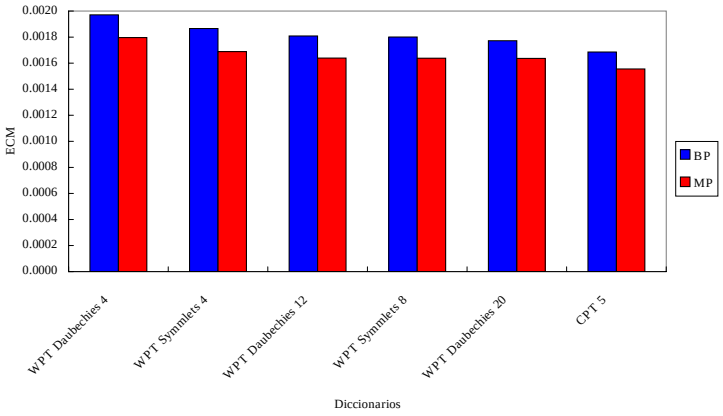

Figura 7.17: *Promedio sobre los distintos fonemas del MSE entre la señal original y la aproximación obtenida a partir de conservar los 15 átomos más importantes seleccionados por BP y MP a partir de diccionarios WPT y CPT.*

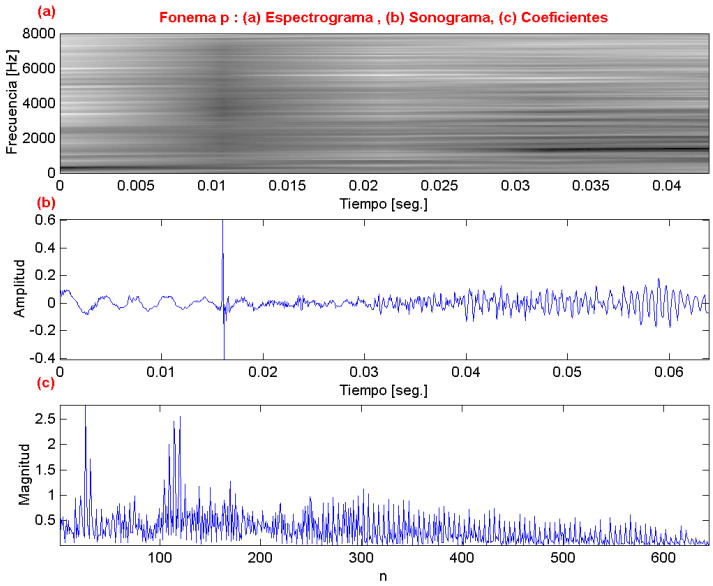

Figura 7.18: *Espectrograma(arriba), sonograma (centro) y magnitud de los coeficientes (abajo) para la señal correspondiente al fonema /p/.*

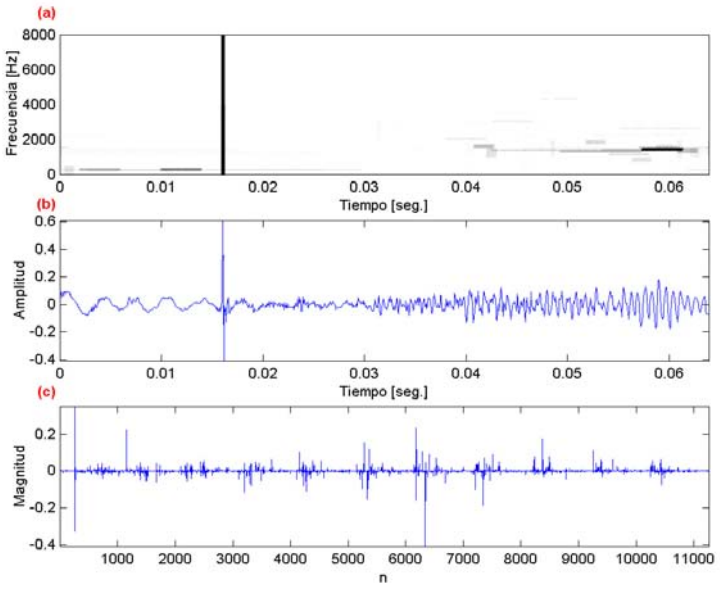

Figura 7.19: *Representación en el plano $t-f$ (arriba), sonograma (centro) y magnitud de los coeficientes correspondiente al fonema /p/ obtenidas a partir de BP y el diccionario WPT Symmlets 8 con profundidad 11.*

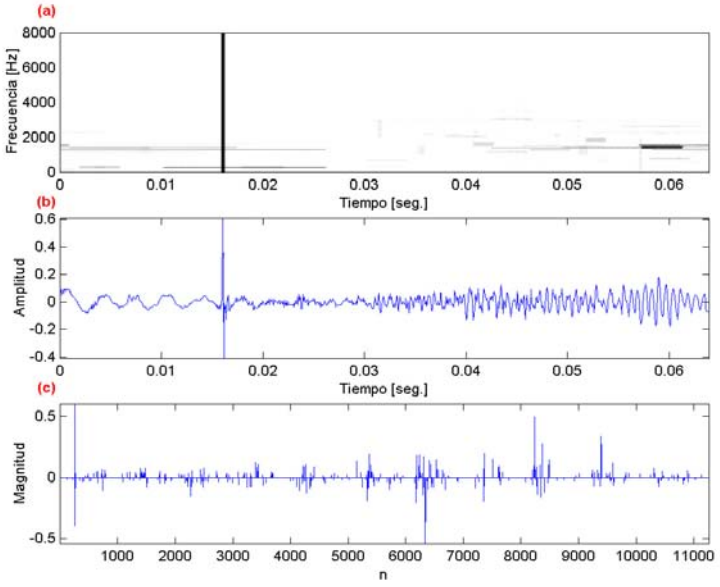

Figura 7.20: *Representación en el plano $t-f$ (arriba), sonograma (centro) y magnitud de los coeficientes correspondiente al fonema /p/ obtenidas a partir de MP y el diccionario WPT Symmlets 8 con profundidad 11.*

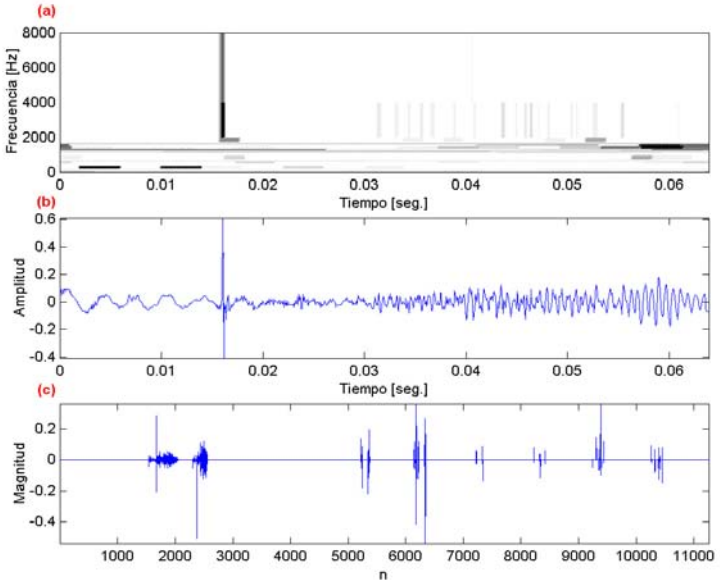

Figura 7.21: *Representación en el plano $t-f$ (arriba), sonograma (centro) y magnitud de los coeficientes correspondiente al fonema /p/ obtenidas a partir de BOB y el diccionario WPT Symmlets 8 con profundidad 11.*

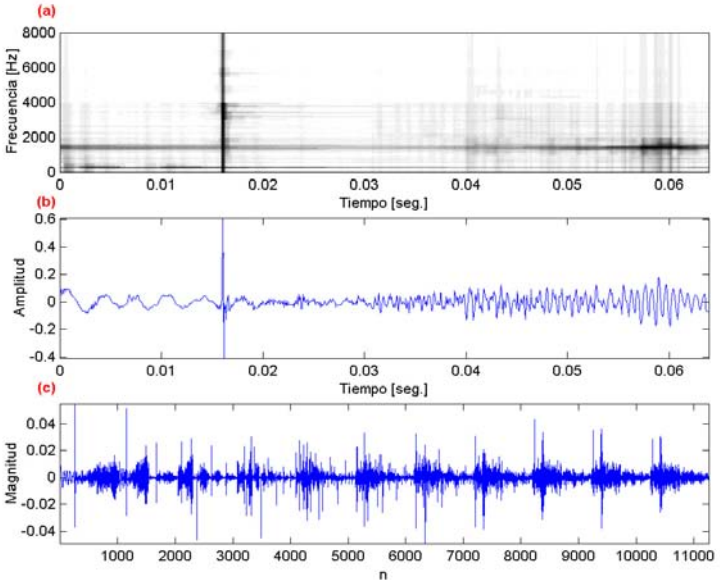

Figura 7.22: *Representación en el plano $t-f$ (arriba), sonograma (centro) y magnitud de los coeficientes correspondiente al fonema /p/ obtenidas a partir de MOF y el diccionario WPT Symmlets 8 con profundidad 11.*

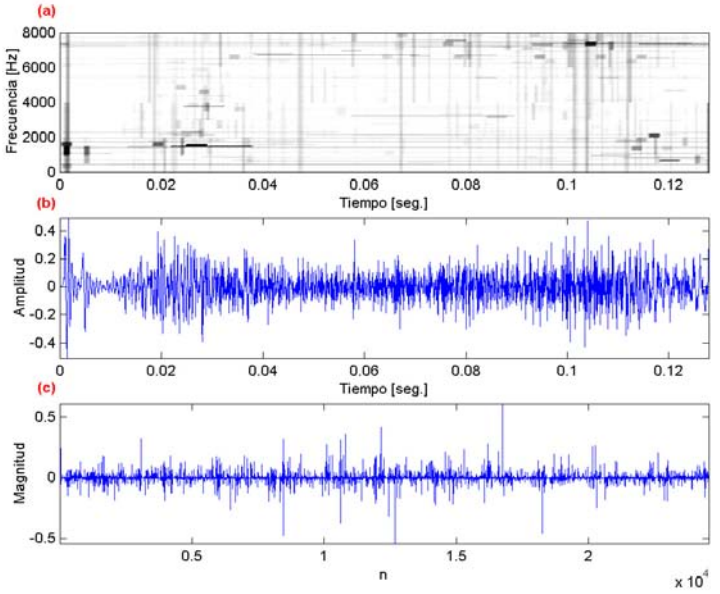

Figura 7.23: *Representación en el plano $t-f$ (arriba), sonograma (centro) y magnitud de los coeficientes correspondiente al fonema /s/ obtenidas a partir de BP y el diccionario WPT Symmlets 8 con profundidad 11. Comparar con la Figura 7.19.*

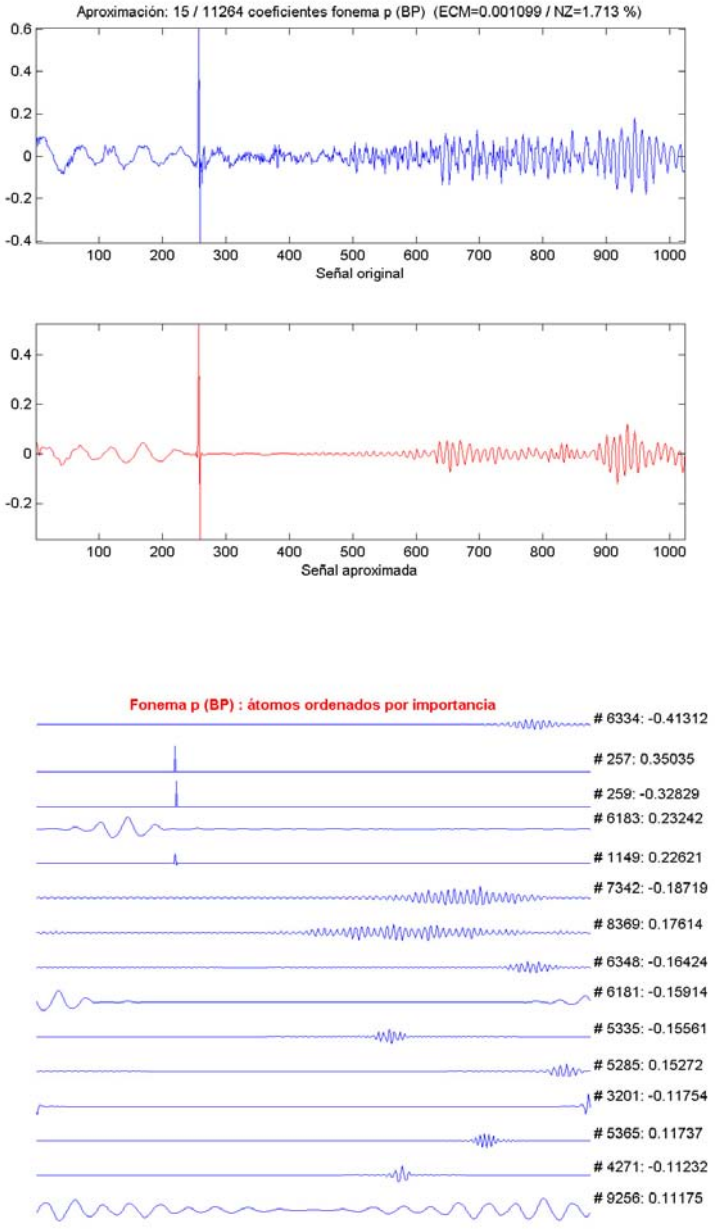

Figura 7.24: *Reconstrucción a partir de los átomos más importantes seleccionados por BP para el fonema /p/: Señal original limpia (arriba), aproximación (centro) junto con los átomos y coeficientes utilizados en la aproximación (abajo).*

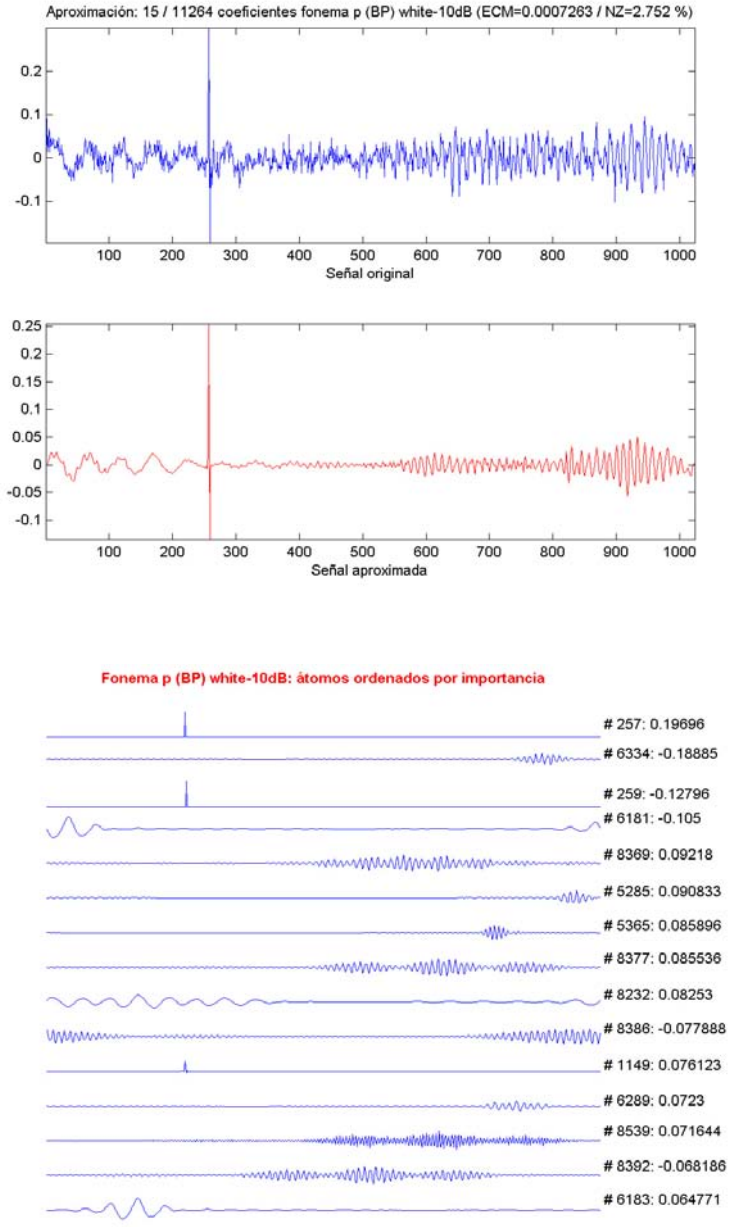

Figura 7.25: *Reconstrucción a partir de los átomos más importantes seleccionados por BP para el fonema /p/ contaminado con ruido blanco a 10 dB SNR: Señal original sucia (arriba), aproximación (centro) junto con los átomos y coeficientes utilizados en la aproximación (abajo).*

base ondita. De hecho tampoco queda restringido a una base ortogonal particular o una familia de bases ortogonales. Debido a ello existen varias formas de tomar ventaja de la robustez de estas representaciones, si el diccionario se elige para describir adecuadamente a la señal (y no al ruido).

En esta sección se propone un método sencillo para limpieza de ruido mediante representaciones ralas y se compara con otras técnicas como las que utilizan umbrales para la *limpieza de ruido mediante onditas* (WDN, ver Sección 6.4.2).

El método propuesto, se denominará *limpieza de ruido heurística* (HDN) y consiste en los siguientes pasos [173]:

1. Elegir un diccionario apropiado, preferentemente sobrecompleto.
2. Encontrar una representación rala de la señal (mediante BP o MP, HDN-BP o HDN-MP respectivamente).
3. Ordenar los átomos por el "tamaño" del coeficiente (valor absoluto).
4. Seleccionar aquellos coeficientes para los cuales el valor de la energía normalizada sea mayor que el MSE normalizado.
5. Igualar a cero el resto de los coeficientes (umbralamiento duro).
6. Reconstruir la señal definitiva a partir de los coeficientes retenidos.

El método propuesto representa un compromiso entre la calidad de la aproximación (en términos de la norma ℓ_2 del error) y la dispersión de la representación (a partir del número de átomos incluidos o ℓ_0). Es decir que se busca la menor cantidad de átomos que mejor contribuyan a la conformación de la señal. De hecho, ésto puede verse como una solución heurística para el problema de regularización planteado por la ecuación (6.21).

Para estos experimentos se utilizaron las señales de Albayzin contaminadas con ruido blanco y murmullo según se describe en el Apéndice B. El número de átomos retenidos se limitó entre 15 y 35. Las pruebas se realizaron con el diccionario CPT 5, que en las pruebas anteriores mostró ser una buena opción para utilizar en la representación.

En la parte izquierda de la Figura 7.26 se muestra un tramo de voz típico correspondiente a un fonema sonoro contaminado con ruido blanco a 10 dB, junto con la versión limpiada mediante HDN-BP. Luego de calcular la representación mediante BP, el número de átomos a retener se establece comparando el MSE relativo y la energía normalizada correspondiente a la señal reconstruida, como se visualiza a la derecha de la misma figura. En la Figura 7.27 se muestran, sobre la escala temporal mayor de una emisión, el sonograma y espectrograma resultado de la aplicación de HDN-BP. Se incluyen también para comparación las correspondientes versiones limpia y ruidosa de la señal. Se puede apreciar que después de la limpieza se han preservado las pistas acústicas importantes, como por ejemplo la liberación del fonema */k/*, las formantes y la duración de las vocales, y el "color" de las fricativa */s/*.

Finalmente se realizó una comparación de la relación señal ruido luego de la limpieza (SNR_{out}) para las diferentes técnicas ensayadas y bajo condiciones ruidosas diferentes.

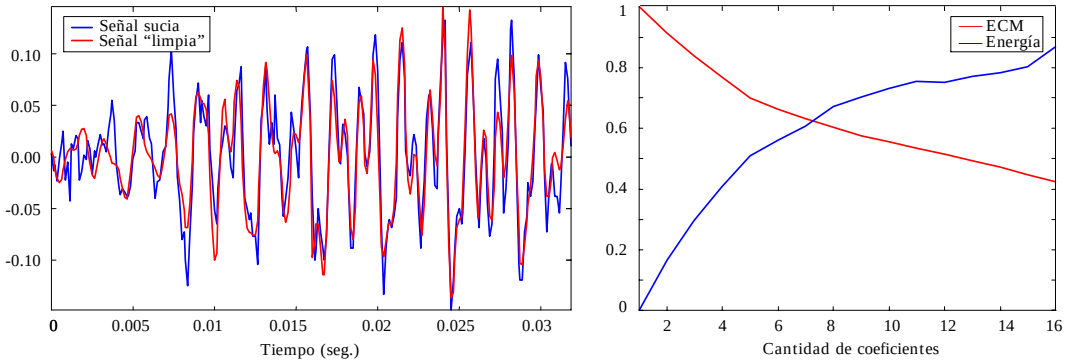

Figura 7.26: *Sonograma de un tramo de voz contaminado con ruido blanco a 10 dB y la correspondiente señal limpia estimada por el algoritmo (izquierda), y gráfico del MSE relativo y la energía normalizada de la señal reconstruida mediante los átomos encontrados por BP en función del número de coeficientes retenidos (ordenados por su importancia relativa, derecha).*

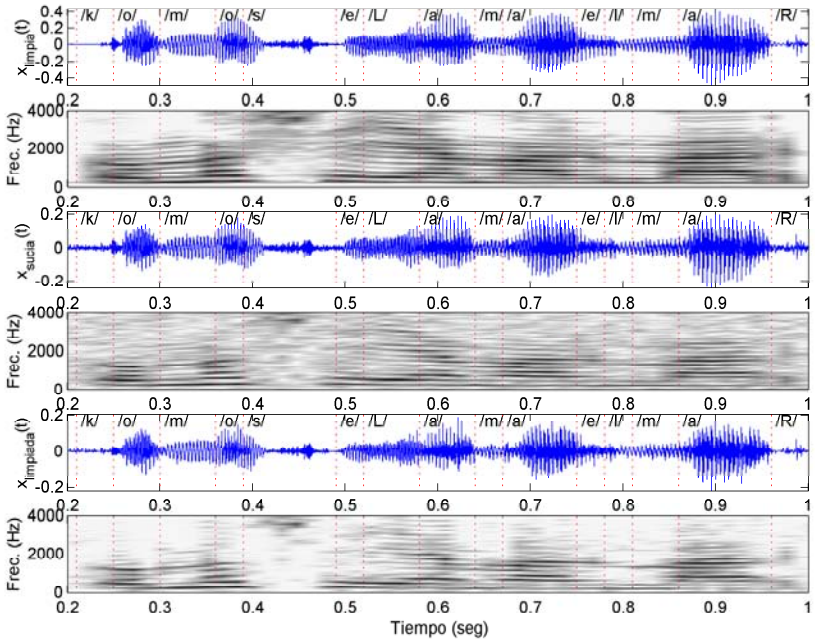

Figura 7.27: *Sonograma y espectrograma: Señal de voz limpia correspondiente al trozo "Cómo se llama..." (arriba), contaminada con ruido aditivo (SNR 10 dB, ruido blanco, centro) y limpiada mediante HDN-BP (abajo). Se puede observar claramente el aumento del contraste espectral en la señal limpiada respecto a la sucia y la preservación de las pistas acústicas significativas.*

Tabla 7.10: *Relación señal ruido de las señales luego de ser limpiadas por los distintos métodos (SNR_{out}), para distintas condiciones (SNR_{in}) y tipos de ruido.*

Tipo de Ruido	SNR_{in} (dB)	SNR_{out} (dB)			
		HDN-BP	HDN-MP	MPDN	WDN
Limpia	∞	**26.56**	14.33	4.00	17.43
Blanco	50	15.16	14.33	4.00	**17.42**
Blanco	25	15.05	14.19	4.01	**17.40**
Blanco	15	14.81	14.37	4.00	**16.22**
Blanco	10	12.18	**14.17**	3.96	14.09
Blanco	5	10.31	**13.18**	3.96	11.09
Blanco	0	5.24	**10.91**	3.16	6.75
Limpia	∞	**26.56**	14.33	4.00	17.43
Murmullo	50	15.13	14.46	4.00	**17.37**
Murmullo	25	15.19	14.23	4.00	**17.12**
Murmullo	15	14.89	13.55	3.89	**15.85**
Murmullo	10	**13.61**	12.24	3.78	13.18
Murmullo	5	**9.75**	8.43	3.42	7.78
Murmullo	0	**3.75**	1.63	2.01	-0.27

Los resultados se muestran en la Tabla 7.10. BPDN y BOBDN [21] se probaron pero no se incluyeron porque no convergieron en los datos utilizados. Puede observarse en la tabla que WDN y HDN tienen los mejores resultados en términos de SNR. En algunos casos WDN es incluso mejor que HDN, aunque ésto ocurre para SNRs relativamente altas[8]. En la mayoría de los casos dónde WDN superó a HDN se encontró distorsión acústica en la forma de "ruido musical" (evaluado mediante pruebas perceptuales subjetivas). Claramente, para una comparación objetiva más definitiva debe utilizarse alguna medida que incluya estos aspectos. Los valores del umbral y otros parámetros del método propuesto se mantuvieron fijos en forma independiente de la SNR_{in}, su adaptación mediante algún método de optimización podría mejorar el desempeño.

En esta sección (7.6.1) se han aplicado una serie de técnicas para lograr una representación rala de una señal. Mediante ejemplos de señales obtenidas a partir de distintos fonemas se han mostrado y discutido los resultados de aplicar BP, MP, BOB y MOF con diccionarios fijos a estas señales en las condiciones originales y luego de contaminarlas con ruido aditivo. En las representaciones tradicionales existe un importante compromiso en la resolución simultánea de eventos en el tiempo y la frecuencia. Esto puede esconder pistas acústicas presentes en la señal. En contraste con ello las técnicas aquí evaluadas proveen una primera aproximación a la solución de este problema, preservando las características importantes inclusive en presencia de ruido. Esto se aprovecha para plantear un método de limpieza de ruido para representaciones ralas con buenos resultados perceptuales. Aunque el método propuesto es simple, el punto importante para notar aquí es que este tipo de representaciones ofrecen una manera bastante directa de "ocuparse" del ruido en el procesamiento del habla (así como en otros campos).

Por supuesto que esta mejora en las capacidades es a costa de incrementar el costo computacional de las técnicas empleadas en el análisis con respecto a las técnicas convencionales, o incluso a las presentadas en la Sección 7.5. Una cuestión importante

[8]Se debe notar que los parámetros de HDN fueron ajustados para aumentar la inteligibilidad de habla limpiada, y se sabe bien que ésto no se correlaciona directamente con la SNR.

es la utilización de métodos de búsqueda automática de diccionarios óptimos, ya que éstos permiten encontrar los átomos que mejor representan a un conjunto particular de señales. Este enfoque se utilizará a continuación.

7.6.2. Diccionarios óptimos

En la sección anterior se desarrollaron una serie de técnicas para aprovechar las ventajas de obtener una representación rala de la señal de voz, basadas en la utilización de diccionarios fijos armados a mano a partir de familias de funciones con características conocidas. Un problema con este enfoque es que se requiere un número muy grande de elementos en el diccionario para asegurar que estén representados todos los comportamientos posibles de las señales a analizar. Sin embargo existen comportamientos o características que son más probables que aparezcan en el contexto de un conjunto de señales particulares. Es posible entonces que algunos elementos no se utilicen de manera importante para describir a la señal, o que varios de ellos puedan unirse para formar una característica más compleja y representativa. También es posible que, para representar características significativas de la señal, algunos átomos requieran ajustes fuera de los límites que imponen las fórmulas que los generaron y que dan la estructura al diccionario. Por ello lo ideal consiste en recurrir a los datos como una forma de estimar los átomos que componen el diccionario de manera óptima. Ésto permite otra vez un enfoque desde la perspectiva de que el diccionario a utilizar permita establecer un modelo de la señal que permita describirla de forma adecuada.

Experimentos iniciales

El problema del aprendizaje automático de los diccionario es un problema bastante más demandante de recursos que el de la inferencia ya discutido. Muchos de los experimentos presentados requirieron de varias semanas de tiempo de máquina para alcanzar los resultados mostrados utilizando computadoras de última generación[9]. Por esta razón se comenzó trabajando con un conjunto relativamente pequeño de señales correspondiente a 10 emisiones de voz tomadas de un único hablante de la base Albayzin.

Para estas pruebas iniciales se utilizó la alternativa propuesta por Olshausen y Field [145, 143], y posteriormente la implementada por Lewicki y Sejnowski [116]. La primera alternativa presentó problemas de convergencia sobre los datos utilizados, debidos a las aproximaciones realizadas para poder resolver el planteo de la regla de actualización del diccionario $\mathbf{\Phi}$ en (6.39). Por ello se muestran en esta sección sólo los resultados del segundo método que resulta un poco más general y con mejores resultados en este contexto. Este método permite encontrar una solución para el caso de *ICA sobrecompleto y con ruido* (NOCICA)[10].

En Figura 7.28 se presentan resultados de la aplicación de esta técnica para el caso de un diccionario sobrecompleto de 64x128, correspondiendo entonces a cada átomo una duración de unos 8 mseg. A partir de esta figura se puede apreciar como los elementos

[9]Basadas en procesadores tipo Pentium IV, 2.8 GHz de velocidad de reloj.

[10]Aquí se utilizó el código desarrollado por Lewicki y descripto en [116]

del diccionario tienden a funciones que hacen recordar fácilmente a fonemas o trozos de fonemas[11]. Se debe remarcar aquí que en los diccionarios tradicionales existían parámetros específicos para cada átomo que permitían realizar algún tipo de ordenamiento de los mismos que le confería un significado a su posición relativa dentro del diccionario. Por ejemplo en los diccionario tipo WPT los distintos parámetros permitían la organización de los elementos a partir de su frecuencia, escala o localización temporal. Para los diccionarios óptimos este último aspecto ya no está presente, por lo que para otorgarles cierto significado o estructura al diccionario deben emplearse otros métodos. En el caso de la Figura 7.28 se los ha ordenado utilizando un *mapa auto-organizativo* (SOM) [103].

Se observa también una distribución de varios de los átomos con localización precisa en el plano tiempo frecuencia, lo que se evidencia en la Figura 7.29. En esta figura se muestra un diagrama de la cobertura del plano $t - f$ para el diccionario considerado, utilizando un método similar al descripto en [114]. Cada elipse da cuenta de la cobertura de un átomo individual. La extensión temporal de cada átomo se indica utilizando el ancho necesario para cubrir el 95 % de su energía. La extensión en la frecuencia se indica mediante el ancho de banda correspondiente, 10 dB debajo del pico principal. Los átomos que no presentaban una buena localización se omitieron del gráfico (es decir aquellos donde el pico espectral principal representa menos del 50 % de la energía total). Se puede decir que casi la mitad de los átomos del diccionario poseen comportamientos complejos que no han podido ser incluidos en este gráfico. Sin embargo el resto se distribuye de forma bastante uniforme en el plano $t - f$.

Como es de esperar, la representación lograda por este diccionario resulta bastante rala. Ésto puede observarse en los histogramas de activación de los coeficientes correspondientes a algunos átomos que se muestran en la Figura 7.30 (esta dispersión se cuantificará en los experimentos de las secciones siguientes).

En la Sección 6.5.2 se han presentado varios métodos que permiten estimar un diccionario óptimo a partir de señales con características generales. En esta sección se mostraron algunos resultados iniciales para un diccionario estimado a partir de señales de habla utilizando un método estadístico.

Sin embargo, el interés aquí está centrado en aprovechar algunas características especiales de las señales de voz. La mayoría de las aproximaciones existentes no tienen en cuenta, por ejemplo, las importantes correlaciones temporales que existen en el habla. Se sabe que estas correlaciones pueden ser aproximadas mediante modelos lineales[12]. El hecho de incorporar este tipo de conocimiento a priori acerca de la señal puede facilitar la búsqueda de una solución conveniente al problema de encontrar un diccionario óptimo. Además ésto puede también ayudar con la interpretación de la representación lograda.

[11] De hecho es posible "escuchar" estos átomos y la mayoría suenan en forma similar a fonemas vocálicos.

[12] Esto constituye el fundamento de la técnica de LPC tan difundida en el anásis del habla y que fue presentada en el Capítulo 4.

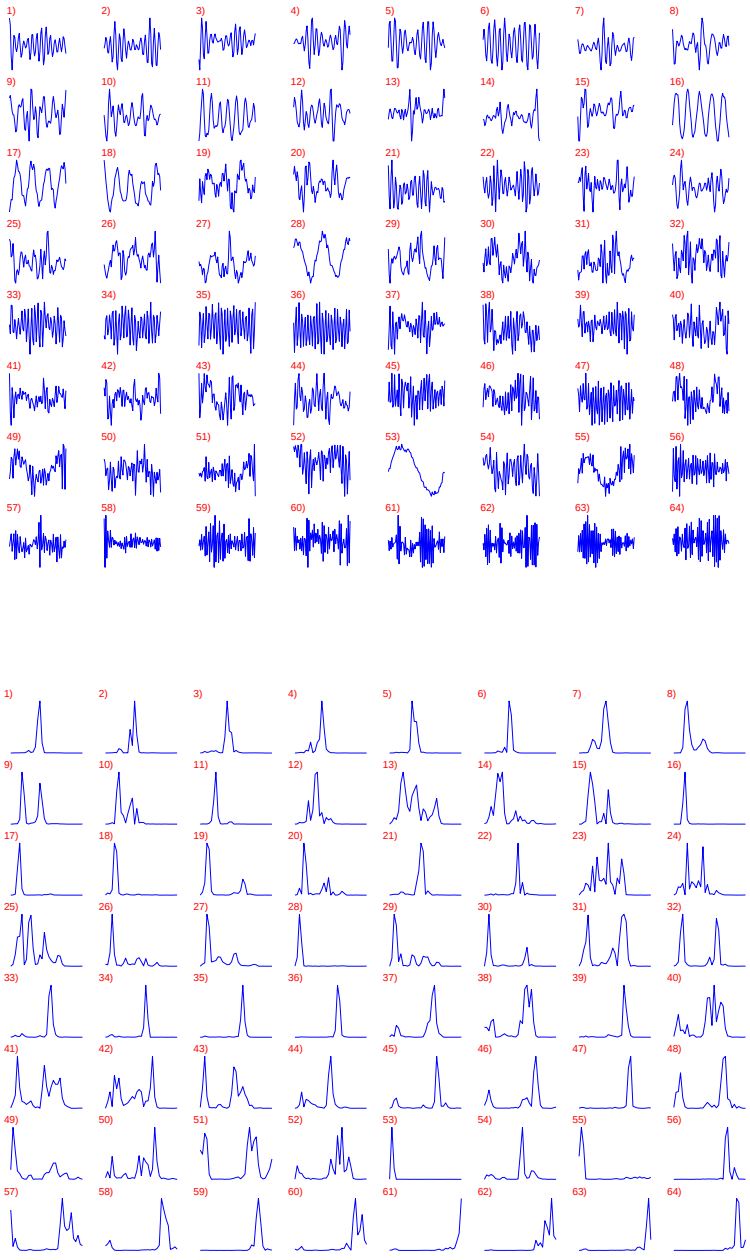

Figura 7.28: *Algunos átomos del diccionario (arriba) y sus correspondientes espectros (abajo) encontrados mediante el método planteado en [116], a partir de señales de voz de un hablante de la base de datos Albayzin. El diccionario completo posee 128 átomos de 64 muestras cada uno. Es posible observar que algunos átomos se parecen a tonos puros, mientras que otros poseen una estructura armónica más compleja que recuerda a la de algunos de los fonemas contenidos en los datos.*

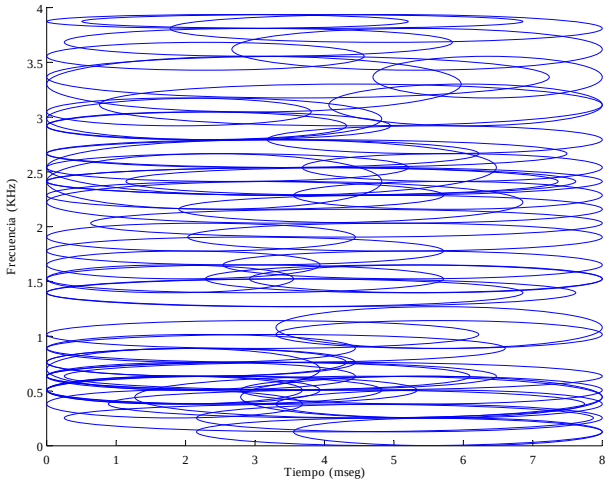

Figura 7.29: *Distribución en el plano $t - f$ de los átomos del diccionario óptimo estimado a partir de los datos de un hablante de Albayzin. Cada elipse representa la cobertura aproximada de uno de los átomos del diccionario calculada a partir de la señal temporal y su respectivo espectro de magnitud. Se han graficado sólo aquellos átomos con una localización tiempo-frecuencia marcada, descartándose aquellos que presentan un comportamiento más complejo (que corresponden a 61 átomos del total de 128).*

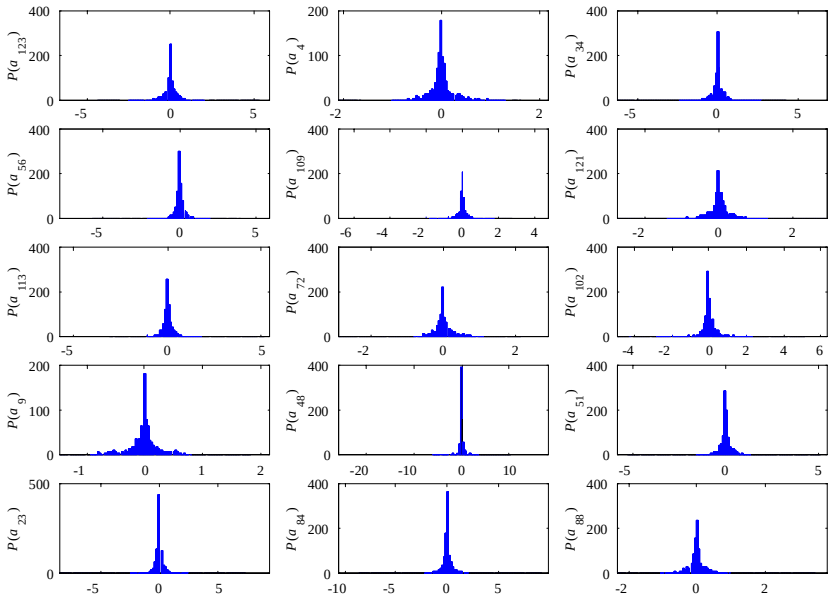

Figura 7.30: *Histogramas de activación estimados para los coeficientes correspondientes a diferentes átomos del diccionario de la Figura 7.28, obtenidos a partir de tramos de voz con los cuales fue entrenado. Se observa a simple vista que todos resultan con curtosis positivo y similares a fdp laplacianas.*

Aprendizaje mediante LP-ICA

El problema de modelar la señal de voz con métodos generales, como el utilizado en la sección anterior, es que ignoran toda la información acerca de la correlación temporal que existe entre las muestras de la señal discreta. En la presente sección se propone un nuevo método para obtener una representación rala de la señal de voz que utiliza un modelo generativo "paramétrico". Este nuevo método consiste en una modificación de la técnica estadística utilizada en la sección anterior, que resuelve el problema ruidoso y sobrecompleto de análisis de componentes independientes, para incluir un modelo lineal de los elementos del diccionario.

Se puede asociar a los átomos $\boldsymbol{\phi}_j$ con los estados característicos de un modelo lineal del tracto vocal para diferentes fonemas. De esta forma un trozo de una señal de voz particular puede obtenerse "sumando" las características más importantes. Para implementar esta idea, en la presente sección se aproxima a las formas de onda utilizadas para el diccionario $\boldsymbol{\Phi}$ en (6.1) por medio de:

$$\hat{\phi}_{i,j} = -\sum_{q=1}^{Q} \phi_{i-q,j}\, c_{q,j} + \delta_i\, \mathrm{g}_j, \tag{7.1}$$

donde $c_{q,j}$ son los coeficientes del predictor lineal, y g_j es el coeficiente de ganancia correspondiente para una entrada tipo delta de Dirac. Esta restricción permite la inclusión explícita de la correlación temporal de las muestras de cada átomo por medio de los coeficientes $c_{q,j}$[13].

Esto significa que el problema a ser resuelto puede expresarse como uno de ICA sobrecompleto ruidoso y con ciertas restricciones en la matriz de mezcla. Estas restricciones incluyen la aproximación por medio de un modelo de predicción lineal para las columnas de esta matriz. Por ello el método propuesto se denominará *ICA por predicción lineal* (LP-ICA) y representa un caso particular de mezclas convolutivas en el dominio del tiempo.

El modelo también puede formularse en el dominio z y entonces la convolución entre $\phi_{i,j}$ y $\delta_i\, \mathrm{g}_j$ se convierte en un producto (simplificado en este caso porque $\mathsf{Z}\{\delta_i\} = 1$). Los átomos $\boldsymbol{\phi}_j$ pueden expresarse de esta manera como una función de z:

$$\Phi_j(z) = \frac{\mathrm{g}_j}{C_j(z)}, \tag{7.2}$$

donde $C_j(z) = 1 + \sum_q c_{q,j}\, z^{-q}$ corresponde a un polinomio de orden Q en z, con coeficientes o parámetros $c_{q,j}$. El modelo generativo correspondiente se muestra en la Figura 7.31.

Para resolver este problema de ICA paramétrico, los problemas de encontrar los coeficientes de la representación, las formas de onda y la aproximación paramétrica pueden manejarse separadamente. La aproximación seguida aquí para encontrar los coeficientes

[13] Aquí el índice temporal avanza en la dirección de las filas i de $\boldsymbol{\Phi}$, es decir a lo largo de cada columna o átomo j.

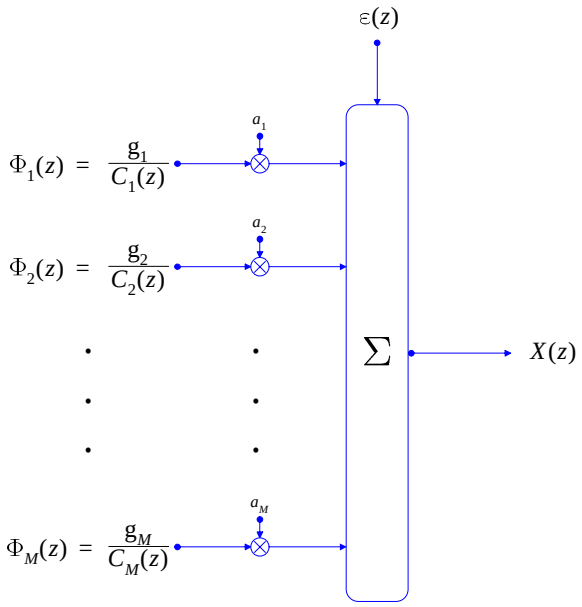

Figura 7.31: *Diagrama del modelo generativo utilizado para las señales de voz. Éste constituye un caso particular de mezclas convolutivas.*

a_j y las formas de onda $\phi_{i,j}$ modeladas paramétricamente es utilizar las técnicas descritas en la Sección 6.5.2 ([116], [115]), incluyendo un paso de aproximación paramétrica [174].

Aquí la matriz $\mathbf{\Phi}$ debe satisfacer simultáneamente las restricciones impuestas sobre las columnas por (7.1) y la maximización de la verosimilitud en (6.36). Una vez estimada $\mathbf{\Phi}$ en cada paso del algoritmo los coeficientes $c_{q,j}$ pueden calcularse por medio de[14]:

$$\frac{\partial\ \mathcal{E}\left[\left\|\boldsymbol{\phi}_j - \hat{\boldsymbol{\phi}}_j\right\|_2\right]}{\partial c_{q,j}} = 0, \tag{7.3}$$

lo que implica minimizar el MSE entre cada átomo $\boldsymbol{\phi}_j$ y su versión $\hat{\boldsymbol{\phi}}_j$ aproximada mediante (7.1). Para la solución de (7.3), se usa el método de Prony [148] debido a su habilidad para recuperar la respuesta al impulso que coincide con una secuencia dada, y porque se comportó mejor que el método de autocorrelacón considerado en la versión inicial [174]. El diccionario $\mathbf{\Phi}$ se reemplaza entonces por su versión paramétrica $\hat{\mathbf{\Phi}}$. Para asegurarse que este cambio no resulte disruptivo durante los primeros pasos del algoritmo, se disminuye la complejidad del modelo gradualmente utilizando el orden Q (mientras que $\log P(\mathbf{x}|\mathbf{\Phi})$ se incrementa). Además, los átomos cuyo error de aproximación excede un umbral predeterminado permanecen inalterados. Resumiendo, la solución del problema puede describirse en términos del algoritmo LP-ICA que se muestra en la Figura 7.32.

Para evaluar el método propuesto en esta sección se realizaron dos tipos de experimentos: uno utilizando datos artificiales, y el otro con datos de habla real. En el primer

[14]Utilizando la hipótesis usual de estacionariedad por tramos de la señal de voz.

```
Inicializar Φ aleatoriamente
Inicializar el orden Q = Q_ini de la aproximación paramétrica
REPETIR
  Inicializar a mediante (6.16), la solución basada en Φ̃
  REPETIR
    Calcular Δa usando (6.31)
    a = a + Δa
  HASTA la condición de finalización
  Calcular ΔΦ mediante (6.39)
  Φ = Φ + ΔΦ
  Calcular c_{q,j} utilizando (7.3)
  Calcular g_j igualando la energía de φ_j y φ̂_j
  Calcular Φ̂ para orden Q usando (7.1)
  SI MSE(φ_j, φ̂_j) >= ϑ ENTONCES φ̂_j = φ_j
  SI |log P(x|Φ̂) − log P(x|Φ)| < ζ ENTONCES Φ = Φ̂, SI NO Q = Q − 1
  SI Q < Q_min ENTONCES Q = Q_min
HASTA la condición de finalización
```

Figura 7.32: *Pseudocódigo del algoritmo LP-ICA para la búsqueda del diccionario óptimo para representar señales de voz. Las constantes predefinidas ϑ y ζ controlan la rapidez y el grado de la aproximación paramétrica. Además existen constantes predefinidas Q_{ini} y Q_{min} que fijan el valor inicial y el valor mínimo permitido para Q respectivamente. Ambas condiciones de finalización se cumplen cuando se alcanza un número predeterminado de iteraciones.*

caso, el proceso de síntesis o "problema directo" puede controlarse usando el modelo generativo de la Figura 7.31 de manera tal que la solución del "problema inverso" resulta conocida de antemano. Para los datos de habla reales, las diferentes representaciones y diccionarios obtenidos para cada tipo de fonemas se compararon con las características importantes de cada clase fonética. Esto puede resultar de particular interés si este enfoque se utiliza para modelar los diferentes fonemas[15].

Experimentos con datos artificiales Se generó un conjunto de datos artificiales a partir de la versión paramétrica del modelo generativo (6.1). De esta forma es posible comprobar el efecto de la utilización de conocimiento a priori acerca de la estructura del modelo generativo en los métodos para estimar el diccionario. El método paramétrico propuesto, LP-ICA, y el método NOCICA descripto en [115] se aplicaron a los datos, y varias de las medidas descriptas en la Sección 2.4.2 fueron calculadas. Un objetivo importante consistió en que estos datos artificiales se parecieran lo más posible a tramos de fonemas vocálicos, de acuerdo con la interpretación del modelo generativo como un sintetizador de voz. Por consiguiente, los elementos del diccionario se escogieron como funciones con 2 polos en el dominio z. Para la elección de los coeficientes $c_{q,j}$, se tomó el valor de la frecuencia de las dos primeras formantes de las cinco vocales españolas, pronunciadas en forma aislada y sostenida por diferentes hablantes [5]. De esta manera, los átomos constituyen sinusoides amortiguadas con frecuencias que son equivalentes a las resonancias características del tracto vocal para la producción de estas vocales. La frecuencia de muestreo utilizada fue de 8000 Hz. El caso considerado consistió en 64 átomos de 64 muestras cada uno (64x64). Con el diccionario armado de esta forma, se generaron los coeficientes a partir de distribuciones laplacianas independientes, y los átomos se mezclaron utilizando (6.1), produciendo el conjunto de datos o señales para los experimentos (un total de 1000 tramos con 64 muestras cada uno). Una cantidad pequeña de ruido fue agregada, con una distribución de gaussiana y media cero (SNR 80 dB)[16]. En las Figuras 7.33 y 7.34 pueden observarse ejemplos de los átomos y de las señales generadas, respectivamente.

Una vez generados los datos, se llevaron a cabo pruebas sobre los coeficientes utilizados originalmente para generar las señales artificiales, los coeficientes estimados a partir de estas señales por diferentes métodos pero con el diccionario original, y los estimados utilizando las bases coseno y ondita discreta diádica. Estos resultados se muestran en la Tabla 7.11. En esta tabla "original" hace referencia al hecho de que se utilizaron los coeficientes y el diccionario que ayudaron generar los datos artificiales en el "problema directo" mediante (6.1). Para las pruebas mediante NOCICA, BP y MP se utilizó el diccionario original, pero los coeficientes se calcularon a partir de los datos artificiales mediante estos métodos. Para las pruebas mediante DCT y DDWT se utilizaron nuevamente los datos artificiales para calcular la representación en términos de la transformada *coseno discreta* y *onditas discreta diádica* (con ondita madre Symmlet 8), respectivamente.

[15]Podrían emplearse por ejemplo en modelos de estadísticos de observaciones como los planteados en [113], para reemplazar a los de mezclas gaussianas actualmente utilizados en los HMMs para ASR.

[16]Esta pequeña cantidad de ruido asegura cierta robustez en la estimación de los átomos del diccionario

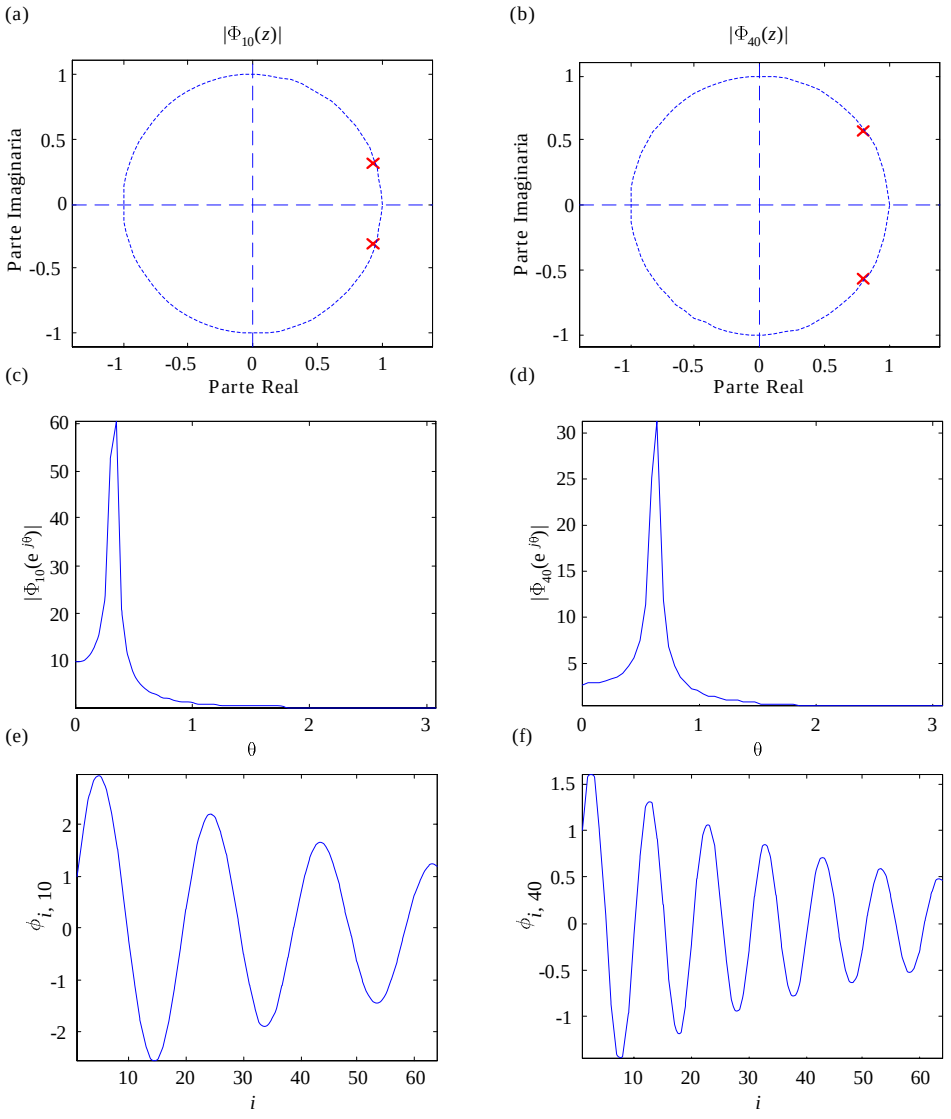

Figura 7.33: *Diagramas de polos y ceros (arriba), espectros (centro) y señales temporales correspondientes (abajo) para dos átomos del diccionario utilizado para generar los datos artificiales con el modelo generativo (izquierda y derecha).*

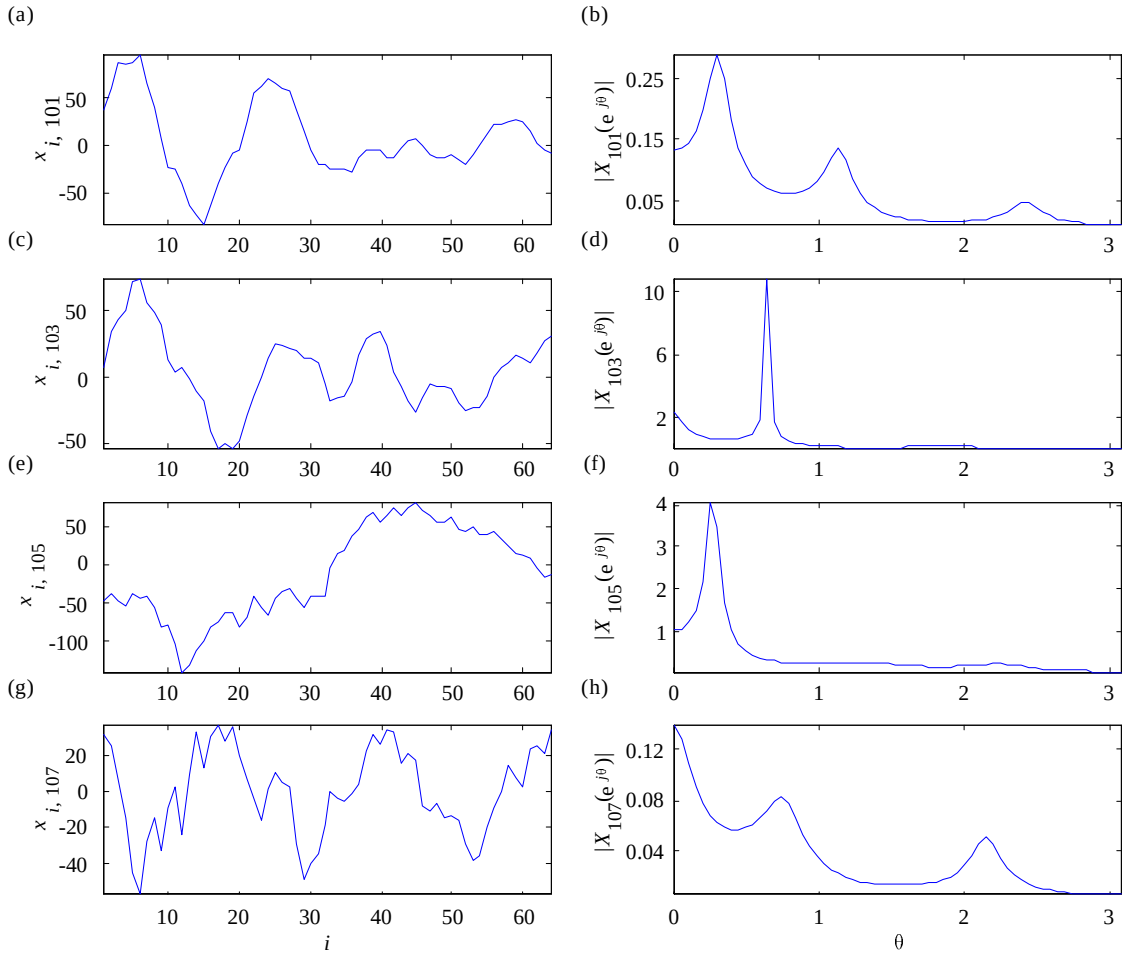

Figura 7.34: *Ejemplos de señales (izquierda) y sus espectros correspondientes (derecha) generados utilizando el diccionario ilustrado en la Figura 7.33.*

Pueden realizarse algunas observaciones a partir de esta última tabla. Entre las transformaciones que utilizan bases ortogonales, la representación menos rala es la generada mediante DDWT. Ésto se debe al hecho de que los elementos de la base son bastante diferentes a los átomos utilizados para generar los datos (teniendo de esta manera que utilizar muchos más elementos para representar cualquiera de las señales). También debe notarse que la DDWT es la que requiere un número más grande de bits para codificar los coeficientes. Lo contrario ocurre con la DCT, debido al hecho de que los átomos son bastante similares a las funciones coseno (aunque las frecuencias para el ejemplo fueron escogidas especialmente). Los coeficientes originales se ubican en una posición intermedia. Entre los métodos específicos para lograr representaciones ralas, puede verse en la tabla que BP logra la representación más rala y requiere un número menor de bits para codificar los coeficientes, aunque con un margen de error mayor que los otros métodos[17]. NOCICA y MP resultan similares. Estos métodos específicos parecen encontrar representaciones incluso más ralas que la original. Un análisis alternativo de la dispersión de las representaciones que confirma algunas de estas observaciones puede realizarse a partir de la Figura 7.35. Aquí se grafica la media de los coeficientes, ordenados y normalizados con el valor máximo para las diferentes representaciones mostradas en la Tabla 7.11. Ésta gráfica provee una forma bastante directa de valorar la dispersión de cada representación a partir de la velocidad de decaimiento de los coeficientes.

El problema inverso se resolvió entonces utilizando ambos algoritmos, el paramétrico o LP-ICA y el no paramétrico, para que pudieran compararse los diccionarios estimados por estos métodos con el original (qué para este caso artificial resulta conocido). La Tabla 7.12 muestra los resultados obtenidos por los diferentes métodos. Como puede verse, todas las medidas excepto ℓ_1, favorecen al método propuesto. Puede decirse que este método logra una representación más rala y con un margen menor de error. Se incluye otra columna en la tabla con el promedio MSE entre el diccionario original y el encontrado por ambos métodos. El resultado muestra cómo el método paramétrico LP-ICA estima mejor el diccionario original que generó los datos que el no paramétrico. Esto puede corroborarse inspeccionando la Figura 7.36 donde se presenta una comparación entre algunos de los átomos obtenidos por ambos métodos con los originales (en el dominio del tiempo y de la frecuencia). Como puede observarse, el método NOCICA tiende a encontrar átomos con más picos espectrales que los originales. Puede apreciarse también como el método paramétrico LP-ICA encuentra átomos que son más similares a los originales y logra una representación más rala que el método de NOCICA. Esto se debe a que se beneficia del conocimiento a priori de que la estructura temporal de los átomos puede describirse en términos de un modelo paramétrico simple (qué es precisamente el caso seleccionado para este ejemplo).

Experimentos con datos reales Para estos experimentos se utilizaron emisiones de distintos hablantes tomadas de un subconjunto del corpus Albayzin [17] (Ver Apéndice B). De la información de segmentación, se extrajeron tramos de 128 muestras para las

[17]En realidad para la reconstrucción mediante la DCT existe un error muy pequeño del orden de la precisión de la máquina que se ha despreciado en la Tabla 7.11 (4.48E-16).

Tabla 7.11: *Medidas de dispersión y costos de codificación obtenidos a partir de las diferentes representaciones de los datos artificiales con el diccionario fijado de antemano.*

Representación	ℓ_0	*minvol*	$\mathcal{K}$	*#bits*	MSE($\mathbf{x} - \mathbf{\Phi a}$)
Original	0.45	0.67	2.89	3.02	2.28E-03
NOCICA (Ec. (6.31))	0.23	0.74	0.98	3.32	**1.26E-03**
BP	**0.05**	**0.51**	**62.55**	**1.64**	4.68E-02
MP	0.28	0.44	8.30	2.38	2.90E-03
DCT	0.21	0.53	0.88	3.37	0.00E+00
DDWT	0.47	0.92	0.39	3.53	0.00E+00

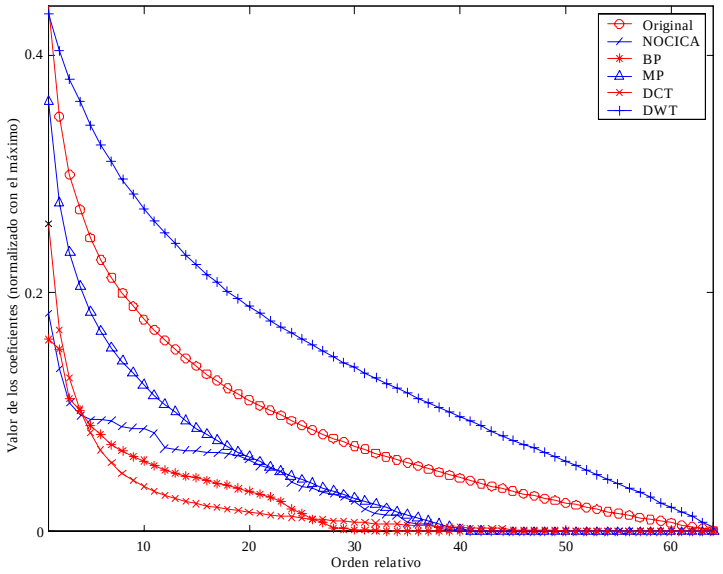

Figura 7.35: *Valor medio de los coeficientes ordenados por magnitud y normalizados con el valor máximo, para las diferentes representaciones mostradas en la Tabla 7.11 obtenidas a partir de un diccionario fijo (64x64).*

Tabla 7.12: *Medidas de dispersión y costos de codificación obtenidos a partir de las representaciones de los datos artificiales usando los diferentes métodos (incluyendo la estimación del diccionario).*

Método	ℓ_0	*minvol*	ℓ_1	$\mathcal{K}$	*#bits*	MSE($\mathbf{x} - \mathbf{\Phi a}$)	MSE($\mathbf{\Phi}, \hat{\mathbf{\Phi}}$)
NOCICA	0.45	0.63	0.60	1.14	3.37	1.28E-04	1.3634
LP-ICA	**0.20**	**0.34**	**0.85**	**21.86**	**2.39**	**5.38E-06**	**1.0902**

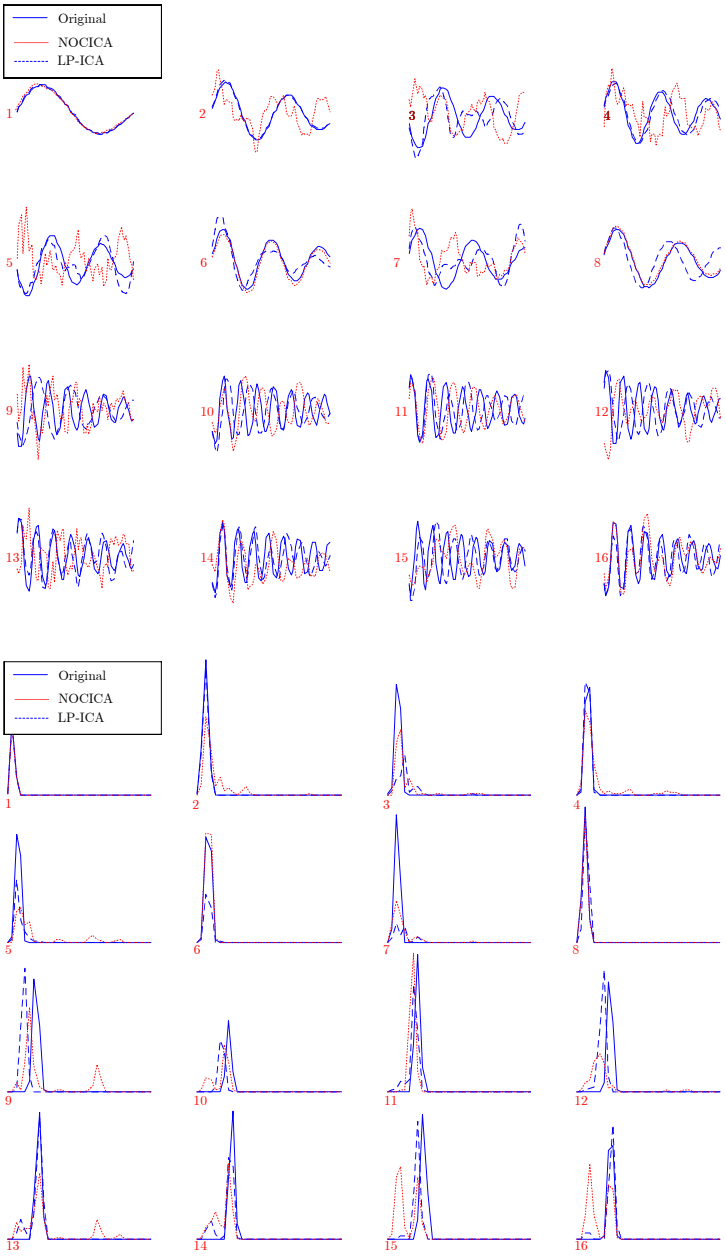

Figura 7.36: *Comparación entre algunos átomos del diccionario original (64x64) y los encontrados por los diferentes métodos a partir de los datos artificiales, en el dominio de tiempo (arriba) y en el de la frecuencia (abajo).*

5 vocales (*/a/*, */e/*, */i/*, */o/* y */u/*) y 2 consonantes (*/s/* y */k/*), obteniendo aproximadamente 2000 tramos de cada uno. Otra vez el subconjunto fue seleccionado para incluir diferentes clases fonéticas diferentes pero manteniendo un tamaño relativamente pequeño. El método paramétrico propuesto LP-ICA y la versión NOCICA se aplicó a estos datos, y se calcularon los valores de medidas descriptas en la Sección 2.4.2 para los casos completo (128x128) y sobrecompleto (128x256). Se realizaron experimentos diferentes, entrenando los métodos a paritr de los datos para cada fonema en forma aislada, y a partir de todos los datos juntos (el caso denominado como "todos"). Los mismos datos se utilizaron para entrenar los diccionarios, y para calcular la dispersión y los costos de codificación.

Los resultados obtenidos para los dos métodos se muestran en las Tablas 7.13 y 7.14 para cada caso. Las últimas filas de las tablas muestran los valores medios y la desviación estándar obtenidas para las columnas correspondientes promediadas sobre todos los fonemas. En el caso paramétrico el orden final medio Q encontrado fue de 29 y 22, para el caso completo y sobrecompleto respectivamente. Como puede observarse en estas tablas el método propuesto LP-ICA da en general una representación más rala, con un número menor de bits y con un MSE similar. Esta diferencia es más pronunciada en el caso del sobrecompleto.

A modo de ejemplo, las formas de onda de algunos de los átomos encontrados para el caso de "todos" los fonemas (128x256) (los datos mezclados de todas las clases) se muestran en la Figura 7.37, para NOCICA y el método paramétrico propuesto respectivamente. A simple vista los diccionarios encontrados parecen similares y para este caso las medidas favorecen sólo ligeramente al método paramétrico .

Para una mejor comprensión acerca de por qué los diccionarios aprendidos reflejan los rasgos más importantes de los diferentes tipos de fonemas, se realizó un análisis cualitativo en algunos de ellos. La Figura 7.38 muestra los espectogramas obtenido de los átomos del diccionario aprendidos para la vocal */a/* (128x256) con ambos métodos. Para el cálculo de estos espectrogramas se realizó un compromiso entre el ancho de la ventana de tiempo y el solapamiento para poder identificar simultáneamente eventos en el tiempo y la frecuencia. Los espectrogramas se ordenaron con un SOM unidimensional de manera que los más similares aparecieran juntos. Finalmente algunos de los átomos intermedios fueron eliminados para mostrar sólo los más importantes.

Entre las diferencias encontradas en los espectrogramas, es posible ver cómo NOCICA logra una representación que no sólo abarca las distintas frecuencias involucradas, sino que además algunos átomos responden a la "fase" de eventos temporales específicos. Por otro lado, dado que el método paramétrico LP-ICA asume que los átomos constituyen respuestas al impulso de filtros AR, el aspecto de la fase relativa se ignora y sólo aparecen átomos sintonizados en frecuencias específicas. Esto indicaría una relativa insensibilidad a la fase que podría ser un rasgo deseable si uno quiere utilizar el diccionario como un detector de eventos independiente al desplazamiento.

La Figura 7.39 muestra los espectogramas obtenidos de los átomos del diccionario aprendidos para el fonema */s/* (128x256) para ambos métodos. Aquí puede realizarse un análisis similar al del fonema */a/* notando que hay una diferencia más marcada en el número de átomos sintonizados en una frecuencia principal. Esto se debe a que para

Tabla 7.13: *Medidas de dispersión y costos de codificación obtenidos a partir de las representaciones de los datos de habla reales para diferentes fonemas utilizando NOCICA.*

Experimento	ℓ_0	*minvol*	ℓ_1	$\mathcal{K}$	$\mathcal{H}$	*#bits*	MSE($\mathbf{x} - \mathbf{\Phi a}$)
/*a*/(128x128)	0.17	0.26	0.54	25.96	1.06	1.86	6.19E-04
/*e*/(128x128)	0.15	0.23	0.47	35.62	0.94	1.79	6.07E-04
/*i*/(128x128)	0.12	0.17	0.32	46.79	0.70	1.33	6.17E-04
/*o*/(128x128)	0.11	0.16	0.35	70.70	0.76	1.24	5.80E-04
/*u*/(128x128)	0.08	0.10	0.17	105.89	0.38	0.74	5.48E-04
/*s*/(128x128)	0.10	0.15	0.32	37.61	0.64	1.05	7.36E-04
/*k*/(128x128)	0.38	0.51	0.56	11.49	0.77	1.63	1.20E-03
Promedio	0.16	0.23	0.39	47.72	0.75	1.38	7.01E-04
Desv. Std	0.10	0.14	0.14	31.48	0.22	0.41	2.28E-04
/*a*/(128x256)	0.22	0.27	0.27	12.99	0.52	1.96	5.55E-04
/*e*/(128x256)	0.09	0.12	0.21	61.15	0.48	1.12	6.07E-04
/*i*/(128x256)	0.06	0.08	0.15	103.55	0.35	0.76	6.06E-04
/*o*/(128x256)	0.04	0.07	0.16	110.18	0.39	0.78	5.86E-04
/*u*/(128x256)	0.04	0.05	0.08	134.24	0.18	0.48	5.41E-04
/*s*/(128x256)	0.17	0.21	0.31	17.19	0.54	1.25	9.30E-04
/*k*/(128x256)	0.21	0.24	0.23	27.17	0.42	1.04	7.80E-04
Todos (128x256)	0.12	0.20	0.51	32.86	1.03	1.29	8.69E-04
Promedio	0.12	0.16	0.24	62.42	0.49	1.09	6.84E-04
Desv. Std	0.07	0.09	0.13	47.42	0.25	0.45	1.52E-04

Tabla 7.14: *Medidas de dispersión y costos de codificación obtenidos a partir de las representaciones de los datos de habla reales para diferentes fonemas utilizando LP-ICA.*

Experimento	ℓ_0	*minvol*	ℓ_1	$\mathcal{K}$	$\mathcal{H}$	*#bits*	MSE($\mathbf{x} - \mathbf{\Phi a}$)
/*a*/(128x128)	0.16	0.24	0.50	32.46	1.00	1.58	6.03E-04
/*e*/(128x128)	0.02	0.11	0.76	173.23	1.92	0.85	1.09E-03
/*i*/(128x128)	0.11	0.16	0.33	61.82	0.75	1.16	7.85E-04
/*o*/(128x128)	0.03	0.09	0.50	201.22	1.22	0.86	8.63E-04
/*u*/(128x128)	0.06	0.08	0.16	157.96	0.36	0.68	6.52E-04
/*s*/(128x128)	0.11	0.18	0.40	36.50	0.82	1.16	8.25E-04
/*k*/(128x128)	0.03	0.06	0.19	143.69	0.50	0.59	7.01E-04
Promedio	0.07	0.13	0.41	115.27	0.94	0.98	7.88E-04
Desv. Std	0.05	0.06	0.21	69.88	0.52	0.34	1.63E-04
/*a*/(128x256)	0.07	0.12	0.30	34.75	0.62	1.31	5.71E-04
/*e*/(128x256)	0.06	0.10	0.28	79.81	0.61	0.99	7.80E-04
/*i*/(128x256)	0.04	0.07	0.21	125.19	0.49	0.71	7.02E-04
/*o*/(128x256)	0.06	0.09	0.17	87.28	0.38	0.76	5.87E-04
/*u*/(128x256)	0.00	0.02	0.15	797.53	0.33	0.20	6.16E-04
/*s*/(128x256)	0.10	0.13	0.20	34.75	0.41	0.88	7.20E-04
/*k*/(128x256)	0.04	0.06	0.13	120.10	0.34	0.56	7.02E-04
Todos (128x256)	0.12	0.19	0.45	36.78	0.92	1.20	8.24E-04
Promedio	0.06	0.10	0.24	164.52	0.51	0.83	6.88E-04
Desv. Std	0.04	0.05	0.10	258.39	0.20	0.36	9.06E-05

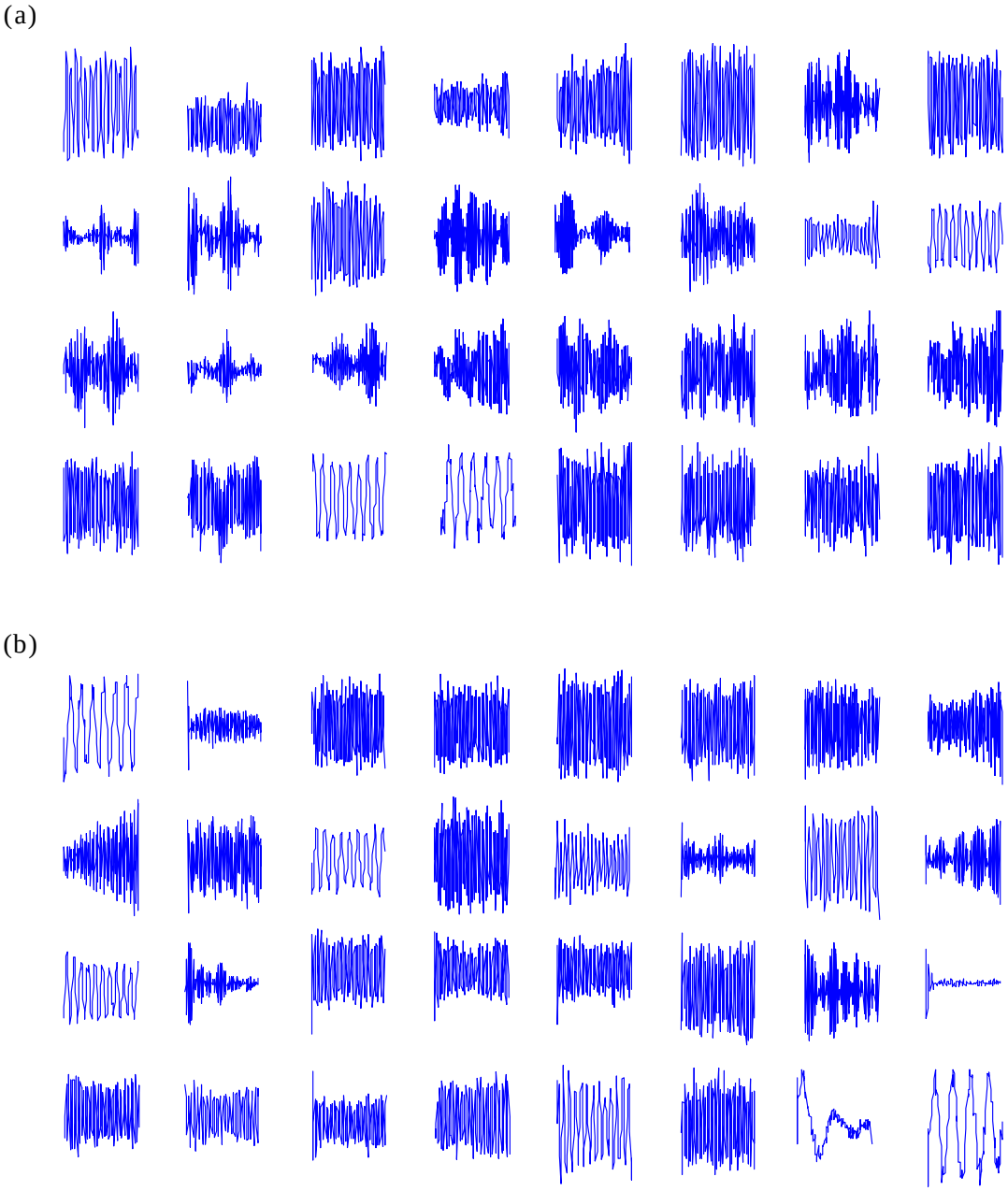

Figura 7.37: *Algunos de los átomos obtenidos a partir de los datos de "todos" los fonemas utilizando: el método de NOCICA (arriba), el método paramétrico propuesto LP-ICA (abajo). Los átomos se presentan en su orden natural.*

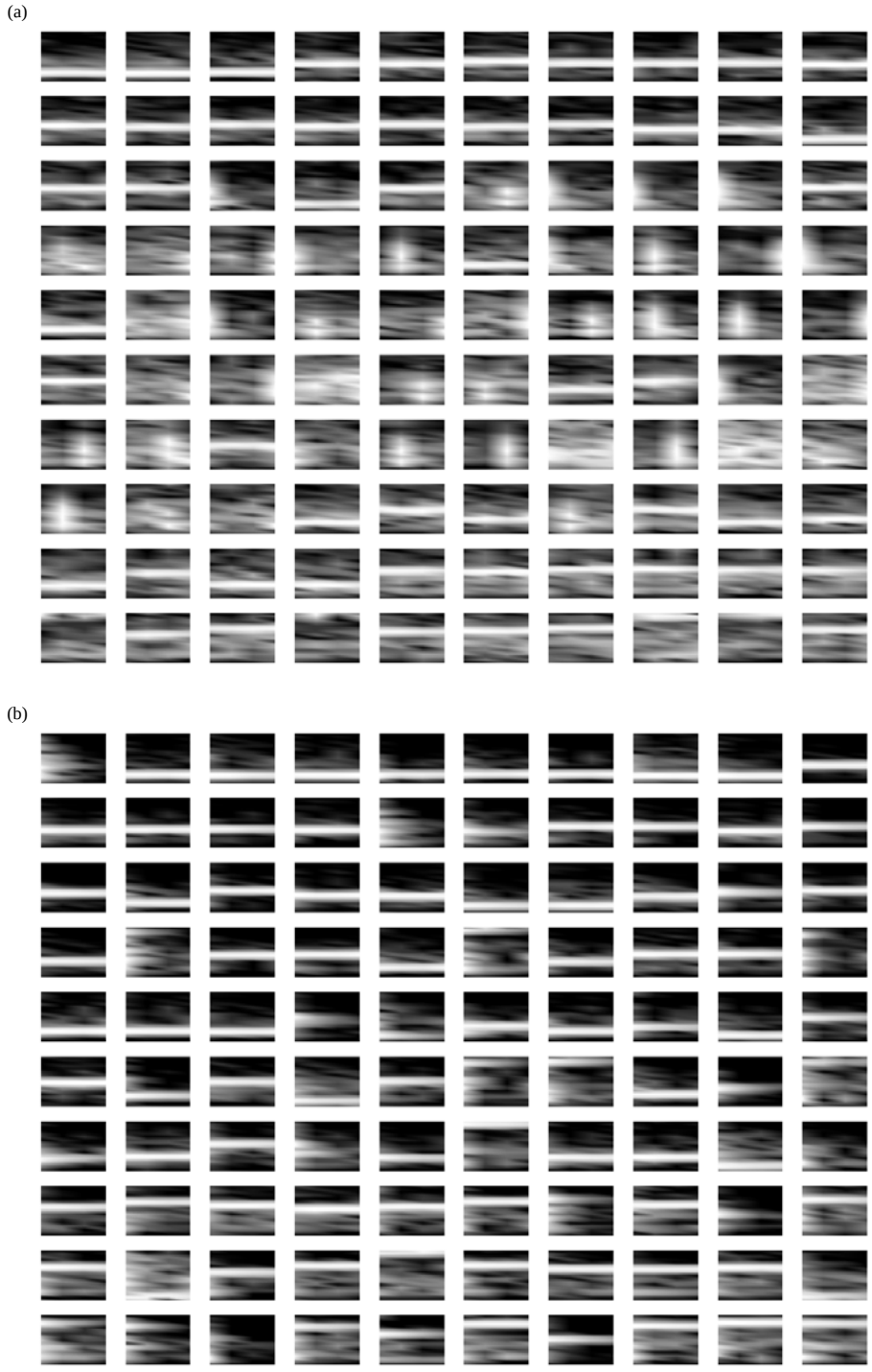

Figura 7.38: *Espectrogramas de los átomos del diccionario aprendido para la vocal /a/ (128x256): NOCICA (arriba), LP-ICA (abajo). El ancho temporal y altura para el eje de frecuencias es de 16 ms y 4 KHz respectivamente para cada átomo.*

lograr anchos de banda grandes, deben usarse modelos más complejos de mayor orden; así el método encuentra una solución más simple que a su vez resulta ser aun más rala (Ver Tablas 7.13 y 7.14 para este caso).

La Figura 7.40 ilustra estos aspectos, dando también una idea más "global" acerca de la distribución de los átomos en el plano $t - f$. Se puede observar en esta figura la cobertura $t - f$ de cada uno de los átomos del diccionario mediante elipses, según se describió en la Sección 7.6.2. Puede verse claramente cómo, en el caso de */a/*, el método propuesto LP-ICA ofrece una mayor resolución en frecuencia para las frecuencias bajas, particularmente en la zona que corresponde al formantes (ver más adelante). También se corrobora el aspecto de fase única ya mencionado. Por otro lado, en el caso de */s/* puede observarse cómo la mayoría de los átomos encontrados por el método propuesto se sintoniza en una frecuencia principal, con mayor resolución en la zona de frecuencias altas.

Si los diccionarios encontrados por ambos métodos se analizan mejor, se puede apreciar que para el caso de las vocales, aunque la mayor parte de la energía de los átomos se localiza alrededor de una frecuencia principal, también aparecen picos de menor magnitud a otras frecuencias . Esto significaría que se ha codificado otra información relevante en los átomos y que no resulta evidente de los análisis anteriores (aunque puede notarse después de una cuidadosa inspección de los espectrogramas en la Figura 7.38). Se ha visto que las formantes son importantes para distinguir entre las vocales, tanto en el caso aislado como en el discurso continuo. Aunque en este último caso se deben rastrear también cambios en los patrones formánticos temporales debido a que las clases no aparecen tan claramente separadas [77]. Debido a su rango de frecuencia esta información extra podría asociarse con las formantes del habla.

Esto significa que estos métodos pueden encontrar información relevante a los fines de la discriminación utilizando sólo los datos de entrenamiento, y en el caso paramétrico LP-ICA, esta información parece estar mejor representada.

Inferencia mediante MP a partir de diccionarios LP-ICA

Una vez obtenido el diccionario paramétrico en base a los datos, el siguiente paso consiste en realizar con él un análisis de las señales de voz, e interpretar los resultados de este análisis en términos de la extracción de características importantes de los fonemas o señales considerados. En este sentido existen varias posibilidades, que consisten en utilizar alguno de los diferentes métodos para la solución del problema de inferencia. En la sección anterior se utilizó el método derivado de la ecuación (6.31). Interesa aquí aplicar alguna alternativa que aproveche alguna característica especial del diccionario aprendido y que también contemple el ruido de manera explícita. Otro aspecto a considerar es la rapidez para encontrar una representación, aunque ésta resulte aproximada.

Teniendo en cuenta estos aspectos se propone utilizar el algoritmo de MP como se expone a continuación. Cada uno de los átomos del diccionario aprendido mediante LP-ICA puede interpretarse como la respuesta al impulso de un filtro de coincidencia (en inglés *matched filter*, MF) de tipo autoregresivo[18]. Los filtros de coincidencia son filtros

[18]Los átomos deben ser previamente normalizados para evitar sesgos en la detección debidos a las

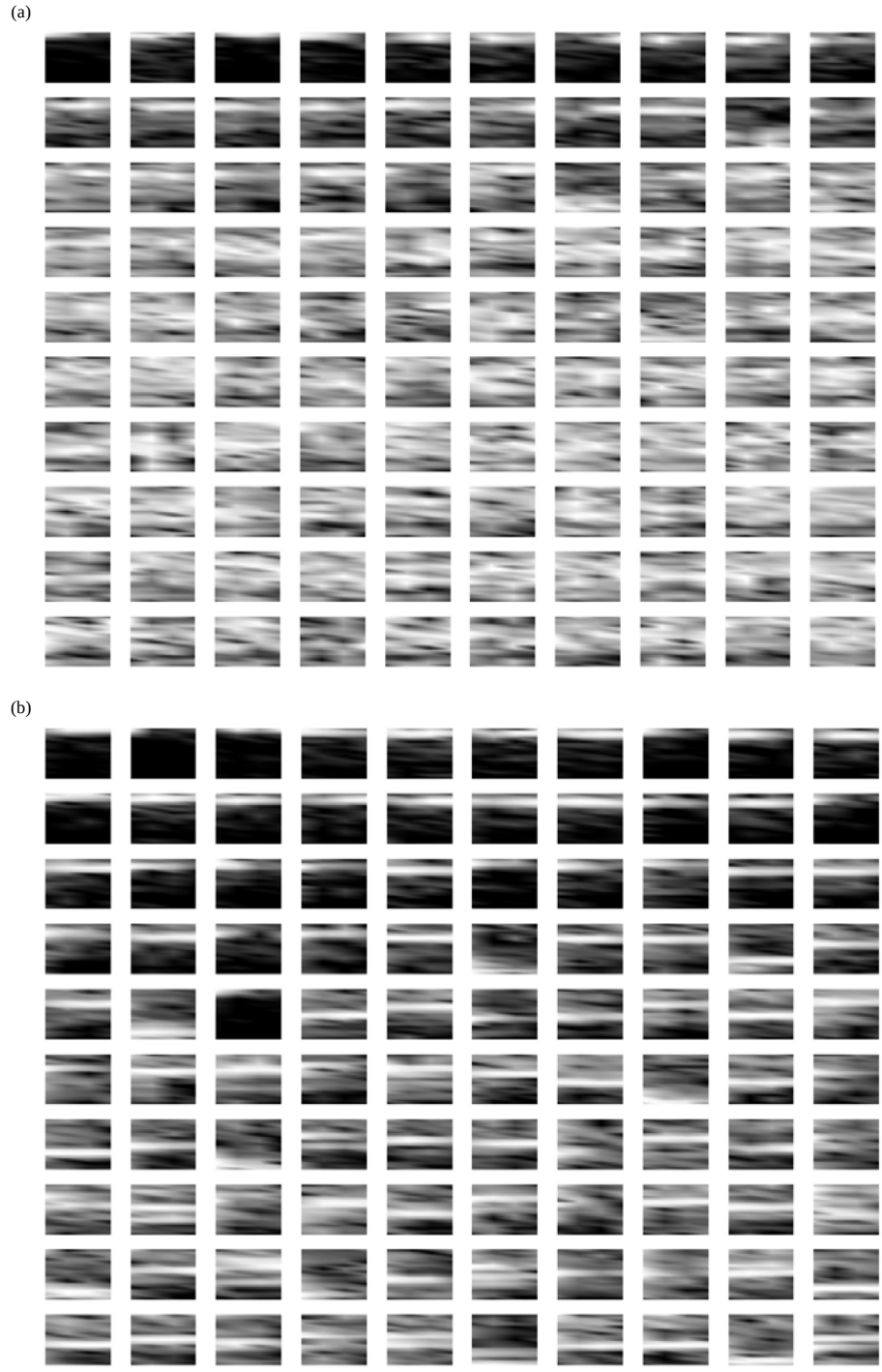

Figura 7.39: *Espectrogramas de los átomos del diccionario aprendido para el fonema fricativo /s/ (128x256): NOCICA (arriba), LP-ICA (abajo). El ancho temporal y altura para el eje de frecuencias es de 16 ms y 4 KHz respectivamente para cada átomo.*

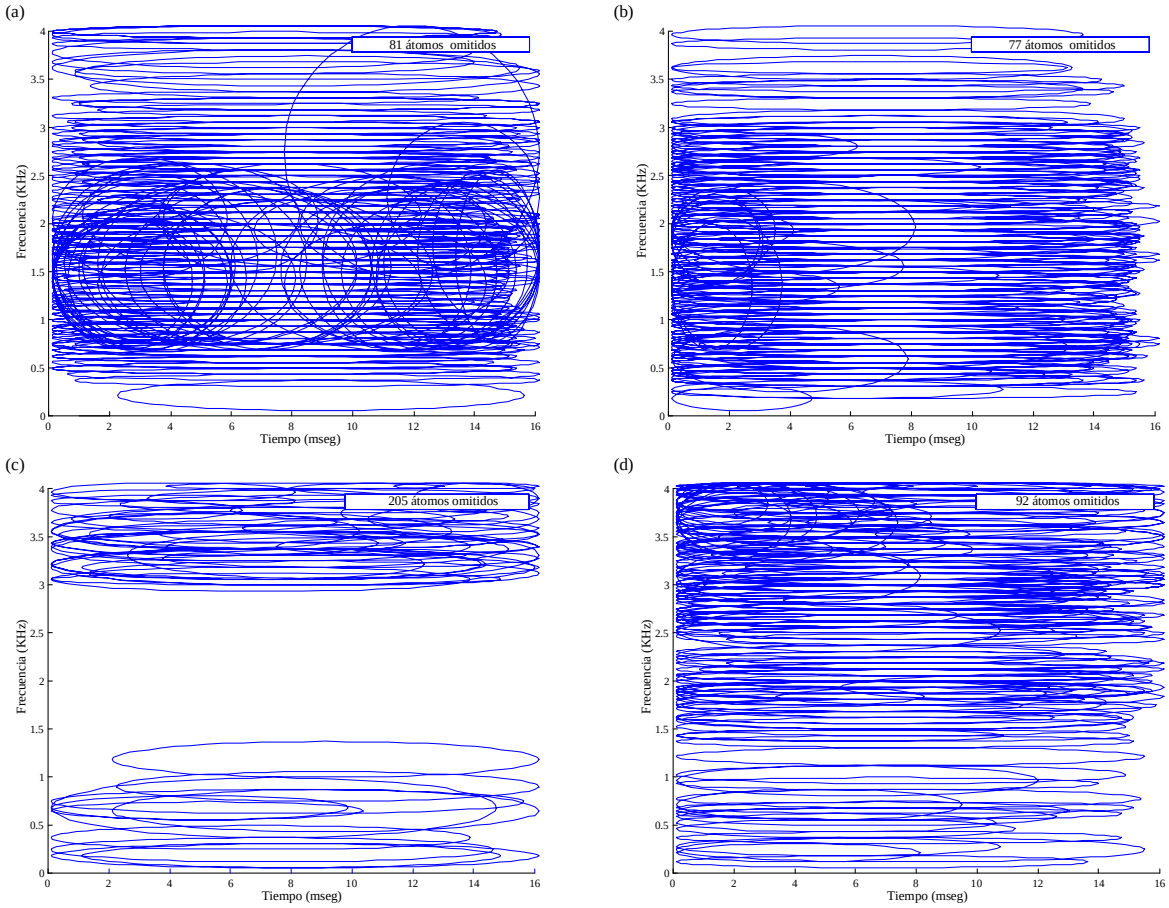

Figura 7.40: *Diagrama de cobertura del plano $t-f$ del diccionario aprendido para: vocal /a/ NOCICA (128x256, arriba izquierda), vocal /a/ LP-ICA (128x256, arriba derecha), fricativo /s/ NOCICA (128x256, abajo izquierda), fricativo /s/ LP-ICA (128x256, abajo derecha). El número de átomos que se omitieron debido a una mala localización se indica en cada gráfico.*

óptimos para la detección de una señal conocida sumergida en ruido blanco, maximizando la SNR para la señal considerada [97]. El algoritmo de MP puede verse entonces como una versión especial de un banco de filtros MF, aplicado iterativamente en la búsqueda de las coincidencias con los rasgos significativos de la señal. Cuando el diccionario ha sido entrenado mediante LP-ICA estos rasgos significativos pueden asociarse con los estados característicos de un modelo lineal del tracto vocal, según se explicó en la sección anterior.

Por supuesto que la representación mediante MP utilizando átomos arbitrarios puede ser también interpretada en términos de filtros de coincidencia. Sin embargo en el caso general ésto no resulta muy útil, salvo desde el punto de vista didáctico. En la sección anterior se mostró como los diccionarios aprendidos mediante LP-ICA presentaban diferencias importantes con los aprendidos sin la restricción del modelo paramétrico. Estas diferencias favorecían en la mayoría de los casos a este enfoque, mostrándolo como un mejor modelo para la señal de voz. De este modo el diccionario LP-ICA debería preservar mejor las pistas significativas del habla incluso en condiciones de ruido[19]. Se puede implementar la limpieza de ruido incluyendo algún tipo de umbralamiento de los coeficientes en el proceso, o limitando simplemente el número de iteraciones del algoritmo. En este sentido se ha comprobado en la Sección 7.6.1 que, si el diccionario es adecuado, muy pocos átomos son necesarios para preservar las pistas acústicas importantes de la señal de voz.

Experimentos de análisis de voz Resulta interesante analizar como, este último enfoque para el cálculo de las activaciones, refleja las características de los fonemas específicos con los cuales se entrenó el diccionario mediante LP-ICA. En la Figura 7.41 se muestra el sonograma, el espectrograma y los correspondientes coeficientes de la representación obtenida para una emisión de voz a partir del diccionario paramétrico entrenado con los datos correspondientes a la */a/* (como en la sección anterior).

Como se puede observar las zonas donde los coeficientes se activan con mayor frecuencia corresponden a aquellas donde aparece el fonema con el cuál se entrenó el diccionario. Para el resto de los fonemas de esta emisión la activación resulta bastante menor, y guarda relación con el parecido relativo entre el fonema analizado y el fonema con el cuál se entrenó el diccionario. Se debe recordar que la posición relativa de cada coeficiente no tiene un significado específico en este análisis. Para solucionar este inconveniente es posible agrupar los átomos correspondiente mediante un SOM, como se mostró anteriormente para el diccionario de la Figura 7.28. También es posible realizar un agrupamiento de acuerdo al grado de activación conjunta de los coeficientes como en el denominado ICA topográfico [89].

Si se estima la probabilidad condicional de activación de los coeficientes del diccionario dado el fonema, o $P(act(\mathbf{a})|/\cdot/)$, ésta puede dar una idea cuantitativa de lo men-

diferencias de energía entre ellos.

[19]Este enfoque permitiría también utilizar simplificaciones similares a las propuestas por Goodwin en [61] para átomos del tipo de sinuoides amortiguadas. Estas simplificaciones pueden acelerar los tiempos requeridos para el cálculo de las correlaciones necesarias para la representación mediante MP, ya sea en el tiempo o en la frecuencia.

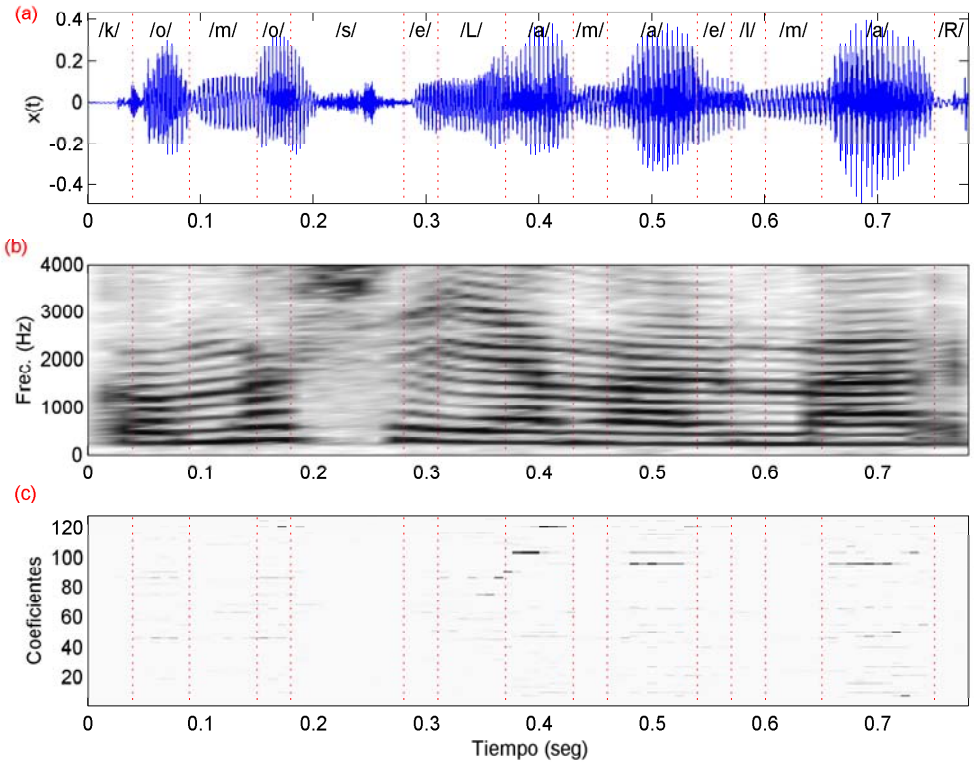

Figura 7.41: *Análisis MP basado en un banco de filtros de coincidencia AR de una emisión de voz: Sonograma (arriba), espectrograma (centro) y coeficientes correspondientes para el diccionario paramétrico (128x128) entrenado a partir de datos del fonema /a/ de un subconjunto de Albayzin.*

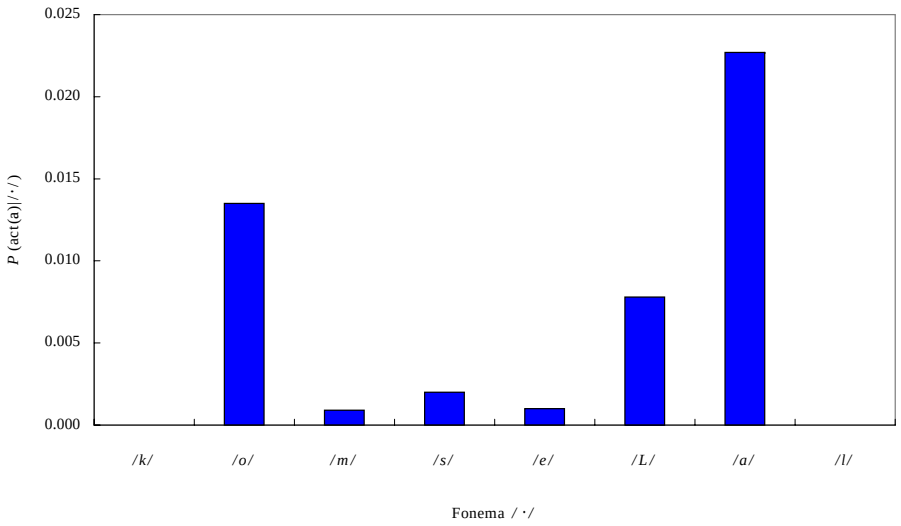

Figura 7.42: *Histograma de la probabilidad condicional de activación de los coeficientes dado el fonema analizado en función de los diferentes fonemas que conforman la emisión, calculados a partir del análisis mostrado en la Figura 7.41.*

cionado anteriormente[20]. En la Figura 7.42 se observa un gráfico de esta probabilidad en función de cada fonema considerado, calculado a partir de los coeficientes del análisis mostrados en la Figura 7.41. Como se puede observar la probabilidad de activación es más grande para el fonema con el cual se entrenó el diccionario[21].

Experimentos de clasificación de fonemas Los experimentos iniciales de clasificación de fonemas utilizando la representación directa mediante los coeficientes MP a partir de diccionarios NOCICA y LP-ICA no resultaron satisfactorios. Una posible causa para éste inconveniente puede ser el hecho de que el diccionario utilice muchos elementos para codificar aspectos relacionados con la fase relativa de ciertos eventos acústicos respecto de la ventana a través de la cuál se extraen los tramos de voz para entrenar el diccionario. Esta información está presente en los datos pero no resulta útil a los fines de la discriminación fonética. En la sección anterior se mostró que los diccionarios aprendidos por LP-ICA codificaban en mucho menor grado esta fase que los aprendidos por NOCICA.

Es posible también relacionar éste comportamiento con el efecto denominado "curso de la dimensionalidad", que es ampliamente conocido y afecta a la mayoría de los clasificadores tradicionales. Como se puede apreciar en el Apéndice A, los experimentos de

[20]Esta probabilidad se estima a partir del histograma de la cantidad de casos donde el valor absoluto de los coeficientes supera un umbral prefijado, en relación con la cantidad de tramos que corresponden a un fonema determinado.

[21]Ésto coincide con lo esperado y sugiere la utilización de esta representación como un modelo específico de las observaciones para cada fonema con un enfoque similar al utilizado en [113], lo que podría resultar útil en el contexto de los HMMs.

clasificación con las técnicas convencionales favorecieron en general a aquellas representaciones con un menor número de dimensiones[22]. Ésto plantea un posible problema para las representaciones consideradas en esta sección, en las cuales la cantidad de dimensiones de los patrones generados puede ser relativamente grande. Por el otro lado la posibilidad de representación de las señales en estos espacios de grandes dimensiones es lo que permitiría separar más fácilmente a la señal del ruido (que aparecen más "mezclados" en pocas dimensiones). En la Sección 7.5.2 la integración de los coeficientes derivados de la WPT permitió disminuir las dimensión efectiva de los patrones (y también el efecto de la fase de los átomos), sin perder información significativa, y por lo tanto mejorar los resultados. Para las representaciones basadas en los diccionarios óptimos y estimadas a partir de MP es posible realizar algo similar a través del cálculo de los coeficientes MFCC. Esto puede verse como una forma de lograr una reducción dimensional aprovechando a su vez las ventajas para el caso de la señal de voz de esta representación convencional.

Para calcular los MFCC a partir de los coeficientes de la representación lograda existen dos caminos diferentes. El más simple, aunque más exigente en recursos computacionales, consiste en re-sintetizar la señal a partir de un pequeño número de coeficientes y calcular entonces los MFCC en la forma tradicional. Como al realizar esta síntesis se realiza una limpieza del ruido presente en la señal, se aprovechan también las características de robustez de las representaciones ralas. Otra posibilidad consiste en realizar el filtrado por coincidencias directamente en el dominio espectral[23]. Ésto resulta particularmente apropiado para los diccionarios aprendidos mediante LP-ICA, donde es posible estimar el espectro de los átomos utilizando los coeficientes del modelo de predicción lineal. A partir del espectro re-sintetizado (y limpiado) de la señal mediante MP es posible ahora completar el resto de los pazos para el cálculo de los MFCC como en el caso tradicional.

En la Tabla 7.15 se muestran los resultados de clasificación de fonemas de TIMIT correspondientes al Apéndice A para las representaciones MFCC generadas mediante MP. Para este caso se utilizaron sólo la mitad de los átomos de los diccionarios obtenidos con NOCICA y LP-ICA a partir de las señales de voz. Nuevamente se han repetido los resultados de referencia en la tabla para facilitar la comparación. Como se puede observar los resultados favorecen a la representaciones obtenidas de los diccionarios LP-ICA. Sin embargo las diferencias no resultan significativas.

Para evaluar la robustez de estas representaciones se agregó ruido aditivo sobre los datos de prueba y se volvió a clasificar los tramos en fonemas. Los resultados para ruido blanco y murmullo a diferentes SNR se muestran en las Figuras 7.43 y 7.44. Es posible apreciar claramente que prácticamente todos los experimentos superan al de referencia en casi todas las condiciones y tipos de ruido (salvo para ruido murmullo y MFCC MP NOCICA 128x128). Para el caso del ruido blanco las mejoras son en general más importantes. Si se comparan los resultados para los diccionarios completos (128x128) se

[22]Ésto ocurre cuando se utilizan clasificadores tradicionales, como los estadísticos o las redes neuronales clásicas. Sin embargo para otro tipo de clasificadores, como los SVM, podría ocurrir que la dimensión de los patrones no resultara una limitación tan importante.

[23]La realización de la operación de filtrado en términos de la magnitud espectral de los átomos convierte a los detectores de coincidencia en independientes de la fase, y la representación debe poseer ahora sólo coeficientes positivos lo que cambia ligeramente la implementación del algoritmo de MP.

Tabla 7.15: *Resultados de experimentos de clasificación de fonemas con redes neuronales y las representaciones generadas mediante MP utilizando la mitad de los átomos de los diccionarios obtenidos con NOCICA y LP-ICA a partir de las señales de voz.*

Nº	Experimento	Estr. Red	TRN	TST	/b/	/d/	/jh/	/eh/	/ih/
1	MFCC MP NOCICA (128x256)	14+14/28/5	77.96	76.26	40.00	56.72	88.64	74.21	79.82
2	MFCC MP LP-ICA (128x256)	14+14/28/5	**78.52**	**78.12**	**31.25**	**59.46**	**91.86**	**76.06**	**82.66**
3	MFCC MP NOCICA (128x128)	14+14/28/5	77.40	75.55	30.00	80.00	84.52	75.70	76.16
4	MFCC MP LP-ICA (128x128)	14+14/28/5	76.15	77.75	28.57	75.00	88.89	76.45	80.33
9	Mel Cepstra+En. (128,14)	14+14/28/5	77.39	77.28	46.51	75.38	91.11	80.56	74.40

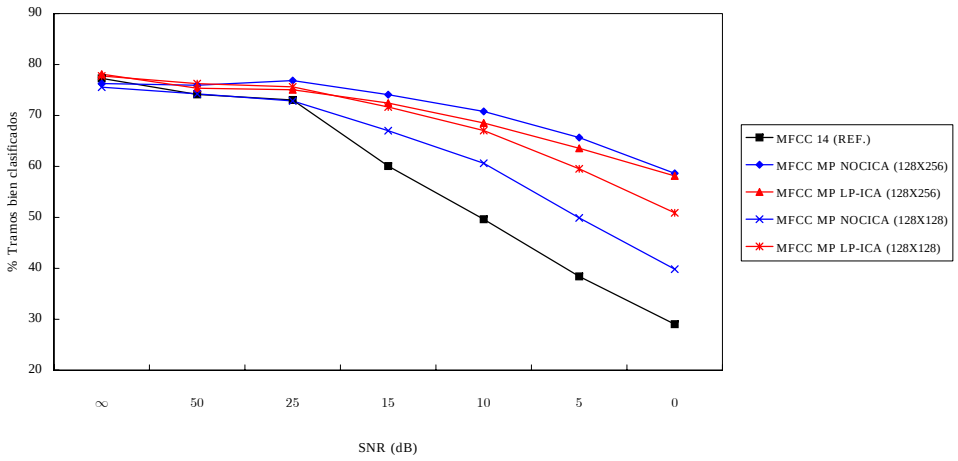

Figura 7.43: *Comparación del porcentaje de tramos bien clasificados entre las representaciones MFCC tradicional y las generadas mediante MP de los diccionarios obtenidos con NOCICA y LP-ICA a partir de las señales de voz, para los experimentos de clasificación de fonemas con redes neuronales con habla limpia y contaminada con diferentes cantidades (SNR) de ruido blanco. En todos los casos se han utilizado sólo la mitad de los átomos del diccionario seleccionados mediante MP.*

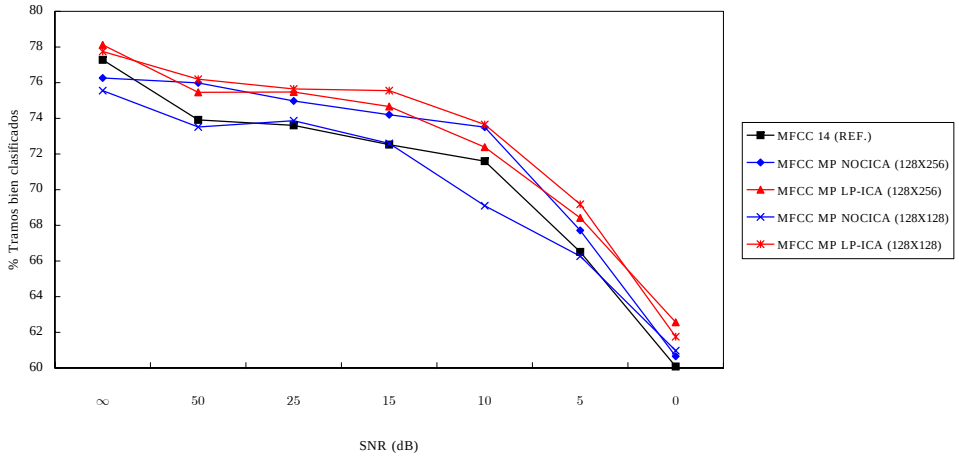

Figura 7.44: *Comparación del porcentaje de tramos bien clasificados entre las representaciones MFCC tradicional y las generadas mediante MP de los diccionarios obtenidos con NOCICA y LP-ICA a partir de las señales de voz, para los experimentos de clasificación de fonemas con redes neuronales con habla limpia y contaminada con diferentes cantidades (SNR) de ruido murmullo. En todos los casos se han utilizado sólo la mitad de los átomos del diccionario seleccionados mediante MP.*

puede observar que los resultados favorecen ampliamente a las representaciones obtenidas de los diccionarios LP-ICA. Para los diccionarios sobrecompletos (128x256) los resultados en general se invierten, aunque no presentan diferencias tan marcadas. En el caso del ruido murmullo los mejores resultados corresponden nuevamente a las representaciones obtenidas de los diccionarios LP-ICA.

En esta sección se han mostrado diferentes formas de aprovechar las ventajas de las representaciones ralas y/o independientes en problemas concretos de limpieza de ruido y clasificación. En la sección siguiente se presenta otra alternativa con un enfoque biológicamente inspirado que permite a su vez encarar algunos de los problemas prácticos mostrados en éstos últimos experimentos.

Diccionarios basados en campos receptivos auditivos

En el Capítulo 3 se presentaron las bases fisiológicas de la comunicación humana. Se mostró como el oído interno, a nivel de la cóclea, realiza un complejo análisis tiempo-frecuencia y codifica una serie de pistas significativas en las descargas del nervio auditivo. Estas representaciones auditivas tempranas, o espectrogramas auditivos, han sido extensamente estudiadas y se dispone de modelos matemáticos y computacionales que permiten estimarlas adecuadamente. Uno de estos modelos es el desarrollado por Shamma [186] (que se ha utilizado para generar la Figura 3.22). Es posible entonces utilizar este tipo de representaciones auditivas como punto de partida para lograr una representación rala que describa a la señal en términos de características un nivel más alto que las derivadas de la estadística temporal.

El sistema auditivo codifica aspectos importantes para la discriminación fonética en los espectrogramas auditivos. En esta representación de más alto nivel se han eliminado

también algunos aspectos "superfluos" de la señal de variación temporal de la presión sonora que llega al tímpano. Entre estos aspectos superfluos se encuentra la fase relativa de algunas ondas a nivel acústico [107]. Es por ello que la estimación de un diccionario bidimensional de características tiempo-frecuencia a partir del espectrograma auditivo resulta de interés para intentar solucionar algunas de las limitaciones descriptas en las secciones anteriores.

En la Sección 3.4.5 se discutieron algunos aspectos de los denominados campos receptivos espectro-temporales y su relación con las representaciones del habla generadas a nivel de la corteza auditiva. También se mencionó que han podido generarse representaciones con características muy similares a las sensoriales a partir de modelos estadísticos que maximizan la dispersión y la independencia estadística [104]. Sería posible interpretar a algunos de los espectrogramas de los átomos aprendidos mediante las técnicas de la Sección 7.6.2 como si fueran equivalentes a las STRF[24]. Sin embargo, el hecho de que estos átomos hayan sido calculados originalmente a partir de sonogramas temporales hace que el diccionario aprendido sea completamente distinto al que se esperaría partiendo de una representación más "auditiva". Ésto es así debido a que las regularidades estadísticas capturadas en los diccionarios dependen de la estructura de los datos con los cuales fueron entrenados.

Experimentos de clasificación de fonemas De acuerdo con las consideraciones anteriores se procedió a diseñar un experimento que permitiera generar representaciones con una interpretación más directa en términos de modelos corticales o STRF. Para ello se utilizaron los datos de habla correspondientes a la región DR1 de TIMIT para todos los fonemas, previa eliminación de los trozos de silencio. Para cada una de las emisiones se calculó el espectrograma auditivo correspondiente, al que posteriormente se le redujo la resolución frecuencial. De esta manera se obtuvieron espectrogramas auditivos con un total de 32 o 16 coeficientes frecuenciales en total para cada instante de tiempo. Finalmente, mediante una ventana deslizante de 160 mseg (16 muestras) corrida a intervalos de 10 mseg (1 muestra), se obtuvo el conjunto de patrones espectro-temporales que servirían de base para la estimación de los diccionarios. En la Figura 7.45 se pueden apreciar algunos de los pasos de este proceso. De acuerdo con la resolución frecuencial de los patrones generados se obtuvieron dos conjuntos de datos: uno basado en patrones de 32x16 (512 dimensiones) y otro en patrones de 16x16 (256 dimensiones).

A partir de estos patrones se entrenaron diccionarios bidimensionales mediante el algoritmo de NOCICA. Se corrieron varias pruebas para los casos completos y sobrecompletos de cada configuración, con una duración de unas 2 o 3 semanas de tiempo total de cálculo para cada una de ellas[25]. En algunos casos se realizó previamente un blanqueo de los datos para facilitar la convergencia de los algoritmos (Ver Sección 2.5.1).

Uno de los diccionarios aprendidos para el caso de los patrones de 32x16 se muestra en la Figura 7.46. En esta figura se pueden observar varios comportamientos característicos útiles para la discriminación entre los diferentes fonemas que formaron parte del material

[24]De hecho esta fue una comparación que se realizó en el Capítulo 1 (Ver Figura 1.7).

[25]En una computadora personal con procesador tipo Pentium IV y 2.8 GHz de velocidad.

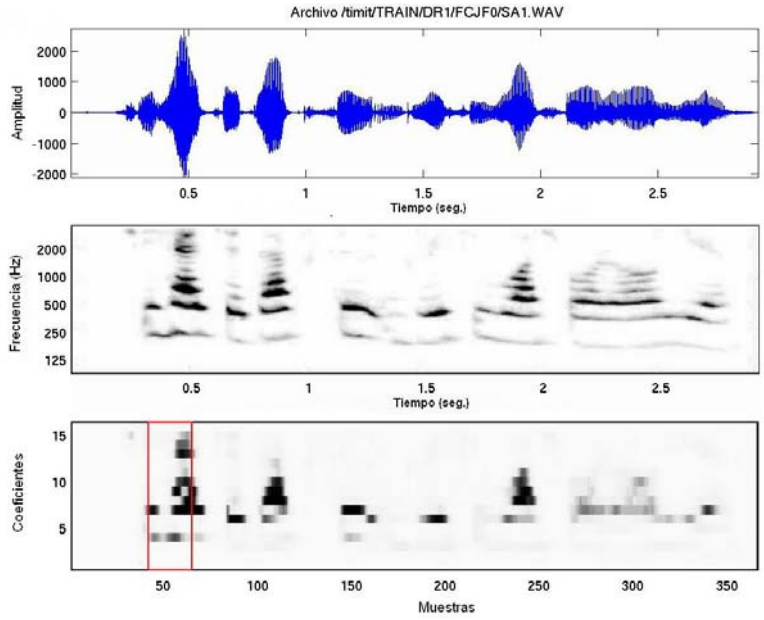

Figura 7.45: *Algunos pasos del proceso para generar los patrones espectro-temporales que sirven de base para estimar los STRF: Sonograma (arriba), espectrograma auditivo original (centro) y espectrograma de baja resolución (abajo). En este último se ha remarcado la sección correspondiente a la ventana deslizante a partir de la cual se genera cada uno de los patrones espectro-temporales.*

empleado para el aprendizaje.

Mediante las STRF así obtenidas se procedió a calcular las activaciones correspondientes utilizando el método de inferencia derivado de la ecuación (6.31). Con estas activaciones como patrones se entrenaron clasificadores de fonemas similares a los descriptos en el Apéndice A. Los resultados de estos experimentos de clasificación de fonemas se pueden observar en la Tabla 7.16. Como se puede apreciar en esta tabla, los resultados de clasificación sobre los datos de entrenamiento para la representación cortical son mejores que los de utilizar la representación auditiva directa. De hecho estos resultados son los mejores reportados hasta ahora en este trabajo y para esta tarea. Otro aspecto importante es que los resultados se mantienen inclusive para estructuras de red relativamente pequeñas en relación con el tamaño de los patrones. Ésto corrobora la hipótesis de que las clases se encuentran mejor separadas en este espacio de grandes dimensiones y por lo tanto un clasificador más sencillo puede completar con éxito la tarea. Sin embargo no resulta un comportamiento similar con respecto a los datos de prueba, que resultan incluso peor que para las representaciones auditivas directas. Una posible explicación de este comportamiento es la siguiente. La cantidad de datos utilizada para entrenar los STRF y/o las redes puede no ser suficiente. Ésto puede crear detectores demasiado específicos de características presentes principalmente en los datos de entrenamiento. El hecho de que la representación resulte demasiado rala implica que los coeficientes se activan muy pocas veces cada uno, lo que puede requerir una mayor cantidad de datos para lograr aprender una regla general de clasificación de los patrones y no sólo los ejemplos. De

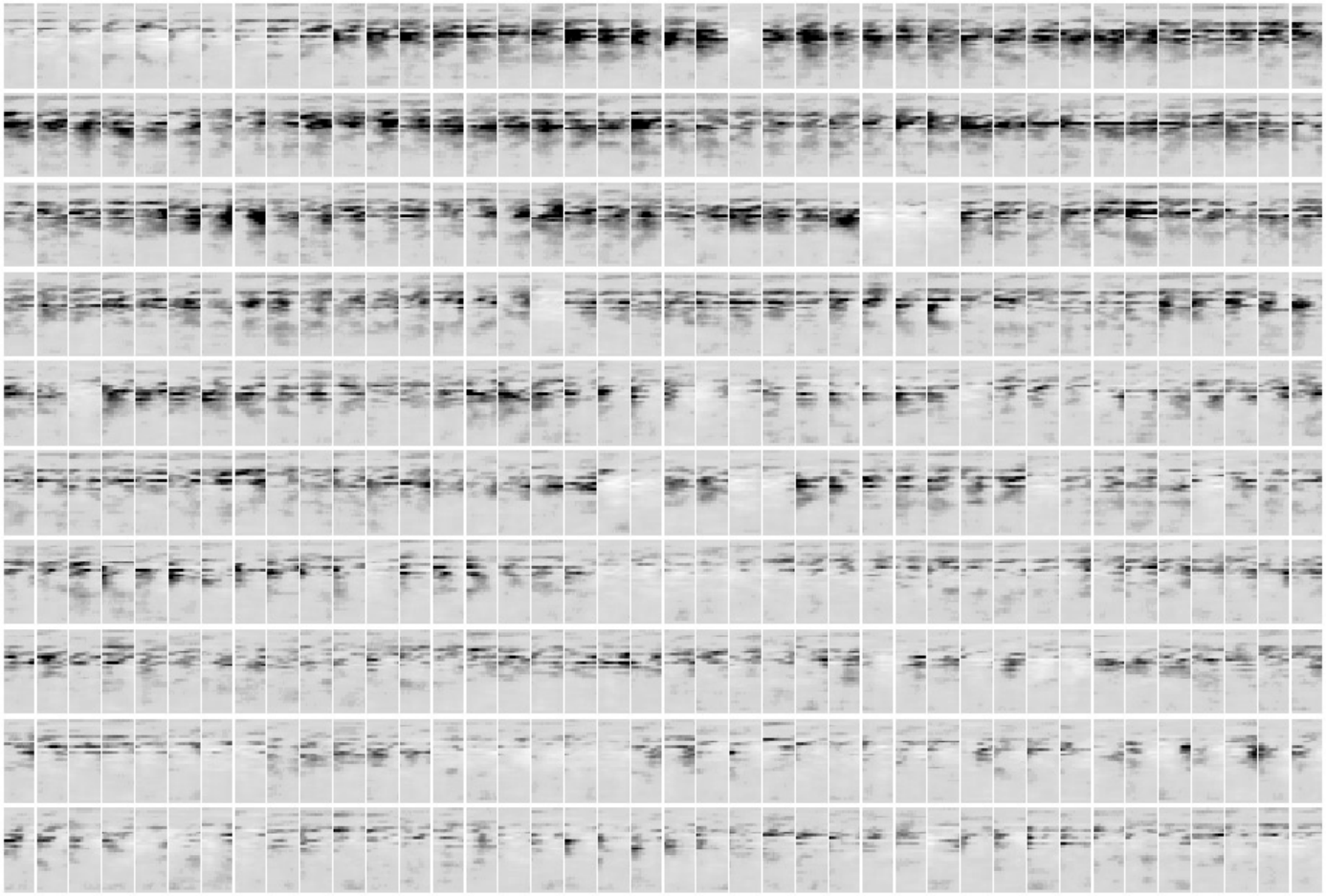

Figura 7.46: *Campos receptivos espectro-temporales estimados a partir de los patrones de 32x16 obtenidos de las representaciones auditivas tempranas de las emisiones de la región DR1 de TIMIT (TRAIN). La posición relativa de cada elemento del diccionario tiene que ver con su similitud respecto a los demás elementos (en términos de la norma ℓ_2 de sus diferencias). Este caso corresponde al diccionario completo $\mathbf{\Phi} \in \mathbb{R}^{512x512}$. Es posible observar STRF que actúan como detectores de diversas características significativas como por ejemplo: frecuencias únicas, patrones formánticos estables, cambios de formantes, componentes ruidosas o fricativas, patrones bien localizados en tiempo o en frecuencia.*

Tabla 7.16: *Resultados de experimentos de clasificación de fonemas con redes neuronales y las representaciones generadas mediante los modelos auditivos tempranos, y los corticales derivados de la activación de los STRF. Estos STRF se estimaron a partir de los datos de habla de TIMIT DR1 (TRAIN).*

Nº	Experimento	Estruct. Red	TRN	TST	/b/	/d/	/jh/	/eh/	/ih/
1	Auditivo 16x16	256/32/5	81.05	**79.10**	64.15	95.12	68.42	81.48	74.14
2	Cortical 16x16	256/32/5	**87.07**	69.08	60.00	72.37	64.71	62.86	77.48
3		256/16/5	82.99	69.59	63.27	69.23	43.75	67.29	78.26
4		256/10/5	84.82	69.17	72.55	74.03	50.00	63.81	72.07
5		256/5/5	83.66	68.48	52.00	75.32	76.47	63.89	74.14
6	Auditivo 32x16	512/64/5	81.38	**80.64**	98.11	67.07	84.21	76.64	85.34
7	Cortical 32x16	512/64/5	91.37	75.20	63.27	81.82	62.50	70.37	82.05
8		512/32/5	92.28	75.14	50.00	85.53	53.33	74.07	83.48
9		512/16/5	94.23	74.05	53.85	84.62	58.82	75.93	76.52
10		512/8/5	91.99	74.93	50.98	85.90	58.82	73.15	82.30
11		512/5/5	**95.05**	75.13	57.69	82.28	52.63	75.00	81.90
3	Mel Fourier (256, 20)	20+20+20/135/5	**82.56**	**81.83**	**72.95**	**79.28**	**90.19**	**84.74**	**78.33**

hecho todos los entrenamientos se detuvieron en el pico de generalización respecto del archivo de prueba, pero si se continuaban entrenando llegaban a clasificar correctamente casi el 100 % de los patrones.

7.7. Comentarios de cierre del capítulo

En este capítulo se han presentado una variedad de métodos posibles para representar el habla mediante técnicas no convencionales. Además, se los ha comparado con las representaciones logradas por algunas técnicas convencionales, a través de análisis cualitativos, cuantitativos y de experimentos de clasificación de fonemas, reconocimiento y limpieza del habla. En el capítulo siguiente se presentarán las conclusiones de este trabajo y las posibilidades de desarrollo de líneas de investigación futuras a partir del mismo.

Capítulo 8

Conclusiones y trabajos futuros

"Apareció en el cielo una señal grande, una mujer envuelta en el sol, con la luna debajo de sus pies, y sobre la cabeza una corona de doce estrellas, ..."

(Apocalipsis 12,1)

Contenido

8.1. Conclusiones generales

En este trabajo se han desarrollado y presentado una serie de alternativas para el análisis y la representación de la señal de voz. Éstas se basan en la utilización de técnicas no convencionales que incorporan aspectos no contemplados explícitamente por las técnicas más tradicionales. Varios de estos aspectos están presentes en los sistemas neurosensoriales humanos y parecen ser responsables de algunas de las extraordinarias capacidades de decodificación del mensaje contenido en el habla.

Las alternativas desarrolladas están orientadas principalmente a la representación de señales basadas en diccionarios discretos de familias de onditas o en formas de onda más generales, que pueden inclusive ser estimadas a partir de los datos de interés. Estas representaciones codifican los rasgos importantes de la señal en términos de unos pocos coeficientes significativos, dando lugar a códigos ralos y/o factoriales.

En la tesis se han presentado, mediante un enfoque unificado e integrador, los fundamentos conceptuales de estas alternativas, junto con el correlato neurofisiológico correspondiente. Éste último ha servido como fuente de inspiración para proponer soluciones

a los problemas planteados. Además se ha evaluado el desempeño de las metodologías propuestas a partir de diversos experimentos de análisis cualitativos y cuantitativos de las representaciones de la señal de voz obtenidas, en el contexto de aplicaciones relacionadas con limpieza de ruido, sistemas de clasificación de fonemas y reconocimiento automático del habla. Se ha demostrado de esta forma la existencia de propiedades útiles en éstas representaciones tales cómo: buena capacidad de aproximación de señales, super-resolución de eventos en el tiempo y en la frecuencia y robustez al ruido. Sin embargo, se debe aclarar que en general éstas propiedades son a costa de un aumento en el tiempo de cálculo de la representación respecto a las alternativas convencionales.

Los experimentos presentados aquí permiten corroborar las ventajas de este tipo de representaciones para el caso de la señal de voz, aunque se requiere todavía un esfuerzo importante para integrarlas de manera definitiva en los sistemas artificiales actuales que intentan emular la comunicación humana.

El enfoque resulta novedoso, ya que prácticamente no se ha aplicado a la representación de señales de voz. Los métodos desarrollados y las propiedades mostradas en este trabajo presentan a las técnicas no convencionales tratadas como una interesante alternativa frente a las técnicas clásicas.

8.2. Conclusiones específicas

Se presenta aquí un resumen de las conclusiones específicas acerca de los distintos aspectos considerados en este trabajo.

Inclusión de cambios complejidad: se introdujeron medidas de complejidad como parte de la representación utilizada por un sistema de ASR. La motivación original para esta aproximación fue la relación observada entre los cambios en la evolución de las medidas de complejidad y la segmentación de la señal de voz. Se obtuvieron mejoras significativas en la cantidad de palabras bien reconocidas para señales de voz inmersas en ruido estacionario con una SNR de entre 10 y 15 dB. Aunque este rango constituye una condición ruidosa "moderada", estos resultados sugieren que las medidas de complejidad proveen información útil para los sistemas de ASR por el hecho de evidenciar los cambios de dinámica del aparato fonador.

Transformada ondita discreta diádica: como continuación del trabajo comenzado en la tesis de Maestría [167] se exploraron y evaluaron una variedad de alternativas basadas en la transformada ondita discreta diádica encontrándose un desempeño relativamente bueno para la clasificación de fonemas transitorios del inglés [169, 170]. Para los fonemas estables resultaron mejores los resultados de los enfoques tradicionales. La variedad de comportamientos, tanto transitorios como estables, presentes en la voz inspiró la búsqueda de otras bases o diccionarios relacionados con familias de onditas más "generales".

Transformada paquetes de onditas: mediante la utilización de criterios originados en la discriminación de frecuencias para el sistema auditivo se diseño un conjunto

de nuevas bases, dentro de las posibilidades ofrecidas por las familias de bases ortogonales de paquetes de onditas [202, 172, 201]. Éste enfoque logró mejorar de manera significativa los resultados a nivel de clasificación de fonemas respecto a los obtenidos con la transformada ondita diádica (del orden de un 16 % en valores absolutos).

Representación rala mediante diccionarios fijos: se evaluó la capacidad de representación con distintos diccionarios basados en familias de funciones con características controladas. Se comprobó que diccionarios sobrecompletos, como los formados por familias de paquetes de onditas y cosenos, son adecuados para la representación de la señal de voz. Esto se tradujo en términos de la dispersión de las representaciones obtenidas y la buena aproximación de las características significativas con muy pocos elementos, inclusive en presencia de ruido [171, 173].

Limpieza de ruido heurística a partir de representaciones ralas: a partir de los resultados anteriores se propuso un método sencillo (HDN) para limpieza de ruido utilizando diccionarios sobrecompletos adecuados. Éste se comparó con otros métodos, como los basados en onditas diádicas, mostrando mejores resultados en cuanto a la preservación (en la señal limpiada) de las pistas acústicas para la discriminación de los fonemas [173].

Representación rala e independiente mediante diccionarios óptimos: se estimaron diccionarios óptimos para tramos de señales de voz y se analizó su capacidad para expresar las características de los distintos fonemas mediante los átomos aprendidos [174]. Se desarrolló un algoritmo específico (LP-ICA) para encontrar el diccionario óptimo que toma en cuenta la correlación temporal presente entre muestras sucesivas de la señal de voz de tiempo discreto. Se realizaron pruebas con datos artificiales y tramos de fonemas reales del castellano donde se mostraron las ventajas del método propuesto. Se propuso una forma de aprovechar la robustez de estas representaciones para mejorar significativamente los resultados de clasificación de fonemas del inglés inmersos en ruido.

Representación mediante modelos de procesamiento cortical: basándose en las analogías establecidas con los sistemas neurosensoriales se estimaron diccionarios óptimos a partir de las representaciones auditivas tempranas correspondientes a señales de voz. Los átomos encontrados, que se identifican con los campos receptivos espectro-temporales de la corteza auditiva, mostraron poder actuar como detectores de características importantes a este nivel. Pueden mencionarse por ejemplo la detección de trozos estacionarios, distintos tipos de evolución de formantes y zonas "ruidosas". Mediante estas representaciones, basadas en características bidimensionales, se entrenó un sistema de clasificación de fonemas. Si bien las capacidades de generalización fueron regulares, se obtuvieron resultados excelentes respecto a los datos de entrenamiento.

8.3. Aportes originales

A continuación se destacan los aportes originales de esta tesis en forma resumida y se discuten las diferencias con otros trabajos relacionados:

1. El análisis de las diferentes *formas de medir la eficacia de una representación* mediante un enfoque unificado que contempla los aspectos relacionados con diferentes normas, los estadísticos y los de teoría de información, resulta novedoso. Harpur realiza un planteo semejante pero centrado sólo en los dos últimos aspectos [71].

2. La inclusión de la *evolución de las medidas de complejidad temporal* en tareas de reconocimiento del habla para incrementar la robustez es original. Se han utilizado algunas de estas medidas a nivel espectral pero en problemas más sencillos como los de detección de voz [81].

3. No se ha reportado en la bibliografía ningún estudio sistemático acerca del desempeño de sistemas de clasificación de fonemas basados en la *transformada ondita diádica* y la influencia de las diferentes familias y parámetros como el que aquí se presenta.

4. Para el caso de la transformada paquetes de onditas, el enfoque que culmina en el desarrollo de la *transformada paquetes de onditas orientada perceptualmente* resulta también original en este contexto. Existen otros trabajos que utilizan la *transformada ondita continua muestreada*, con un costo computacional sustancialmente mayor [47]. Posteriormente a nuestros primeros artículos han aparecido aportes equivalentes que aprovechan también algunas de las propiedades de la transformada para realizar una limpieza de ruido [46].

5. No se ha encontrado un estudio sistemático acerca de las propiedades de las representaciones de la señal de voz mediante métodos "ralos" como el de *búsqueda de bases* y diferentes diccionarios. Chen y colaboradores dan varios ejemplos con señales artificiales sencillas que muestran los beneficios de este método comparado con las representaciones correspondientes encontradas por otros métodos [20].

6. Si bien existen trabajos recientes sobre limpieza de ruido mediante diccionarios sobrecompletos, regularización y esquemas relacionados [190], no se han hallado trabajos que analicen la preservación de las pistas acústicas de la señal de voz fundamentales para la inteligibilidad.

7. Las técnicas disponibles para encontrar diccionarios óptimos, en cuanto a su dispersión e independencia estadística, no incluyen restricciones acerca de la estructura temporal entre muestras sucesivas de la señal de voz como las utilizadas en el método de *ICA por predicción lineal* aquí propuesto. En [87] se presenta un método denominado *búsqueda por complejidad* que incluye ideas relacionadas con un modelo de predicción lineal pero utilizando la *complejidad de Kolmogoroff*. Existen

también algunos enfoques que incorporan información relacionada con la estructura temporal de los coeficientes de la representación en diferentes instantes, como *ICA sensible al contexto* [150, 149].

8. La utilización de representaciones ralas e independientes en el contexto de clasificación de fonemas también resulta novedosa, particularmente respecto a la utilización de las propiedades de robustez para estimar los *coeficientes cepstrales* limpios. Existen trabajos recientes donde se integran algunas de las características de estas representaciones en un sistema de reconocimiento del habla, pero sólo para el caso limpio y con resultados todavía preliminares [107].

9. El empleo de diccionarios que emulan a los *campos receptivos espectro temporales* de la corteza auditiva y métodos ralos para clasificación de fonemas es original y compatible con el enfoque de sistemas *biológicamente inspirados*. En [104] se realiza un análisis cualitativo de diccionarios obtenidos de manera similar y se comparan sus propiedades con las de los campos receptivos reales. Harpur realiza experimentos sencillos de clasificación de fonemas pero utilizando códigos de baja entropía, con coeficientes sólo positivos, generados a partir de bancos de filtros [71].

10. Finalmente, el análisis comparativo para el caso de la señal de voz de todas las alternativas planteadas mediante un enfoque unificado que contempla aspectos fisiológicos, de identificación y modelado de sistemas, teoría de señales e información, resulta también un aporte.

8.4. Trabajos futuros

Existen varias direcciones posibles para continuar con la investigación en este área, entre las que se destacan:

Inclusión de cambios de complejidad: mediante el análisis multiresolución es posible estudiar los cambios de complejidad de la señal de voz a diferentes escalas [203]. A partir de esta información se pueden proponer representaciones alternativas que utilicen una estimación más robusta de los cambios de complejidad.

Métodos específicos para la búsqueda de diccionarios óptimos: se pueden desarrollar nuevos métodos para estimar diccionarios óptimos que incorporen características adicionales de la señal de voz o del sistema auditivo. Por ejemplo, el método de LP-ICA no utiliza información respecto a la correlación existente entre los coeficientes de la representación correspondientes a diferentes tramos temporales [23].

Diccionarios óptimos mediante estrategias evolutivas: es posible utilizar métodos de búsqueda de diccionarios óptimos basados en algoritmos genéticos o evolutivos [78, 131, 60]. Estos métodos de optimización pueden encontrar por ejemplo el diccionario que logre la mejor discriminación para un conjunto de fonemas, dentro

de las familias paquetes de onditas o cosenos, con o sin la restricción de ortogonalidad [55, 189]. La ventaja de este enfoque radica en la posibilidad de maximizar varios criterios simultáneamente, como por ejemplo discriminación, dispersión e independencia estadística. Ésto resulta difícil de lograr con métodos analíticos, sin tener que recurrir a demasiadas simplificaciones [178, 180].

Tratamiento del ruido: es importante proponer esquemas que incluyan el tratamiento explícito de otros tipos de ruidos, como por ejemplo no gaussianos y/o no estacionarios. Para ello se requieren plantear y resolver nuevos problemas de regularización y limpieza de ruido con enfoques aún más generales [182].

Medidas de calidad de la representación: se requiere encontrar nuevos criterios y medidas que permitan valorar cuán buena resulta una representación en términos de discriminación, dispersión, error de aproximación, etc. Ésto es particularmente necesario cuando las medidas no incluyen aspectos subjetivos relacionados con la percepción, como en el caso que nos ocupa [173].

Integración con sistemas artificiales actuales: las técnicas convencionales funcionan muy bien con los métodos actuales de modelado o clasificación acústica, debido a una sintonización conjunta que llevó bastante tiempo. Sin embargo éstos no se ajustan adecuadamente a algunas de las características de las nuevas representaciones [107]. Un ejemplo de ello es la forma de las distribuciones de probabilidad de las representaciones ralas que resultan con curtosis positivos relativamente grandes y por lo tanto no son bien modeladas por las mezclas de gaussianas utilizadas en los modelos ocultos de Markov convencionales. Por lo tanto se necesita adaptarlos en diversos aspectos para que aprovechen completamente las nuevas características (Ver Figura 8.1).

Métodos de modelado o clasificación acústica: como otra alternativa al planteo anterior se puede pensar en métodos que compartan algunos de los principios de funcionamiento de estas representaciones. Por ejemplo, como se dijo anteriormente las representaciones ralas son fácilmente codificables en términos de trenes de pulsos similares a los utilizados a nivel de las neuronas biológicas. Para convertir estos trenes de pulsos en clases fonéticas es posible utilizar redes neuronales pulsadas con sinapsis dinámicas, que son un tipo de redes que se ajustan mejor al comportamiento biológico que las tradicionales [92]. También es posible utilizar clasificadores basados en máquinas de soporte vectorial que comparten propiedades importantes con los métodos para lograr representaciones ralas [208, 57, 58, 59].

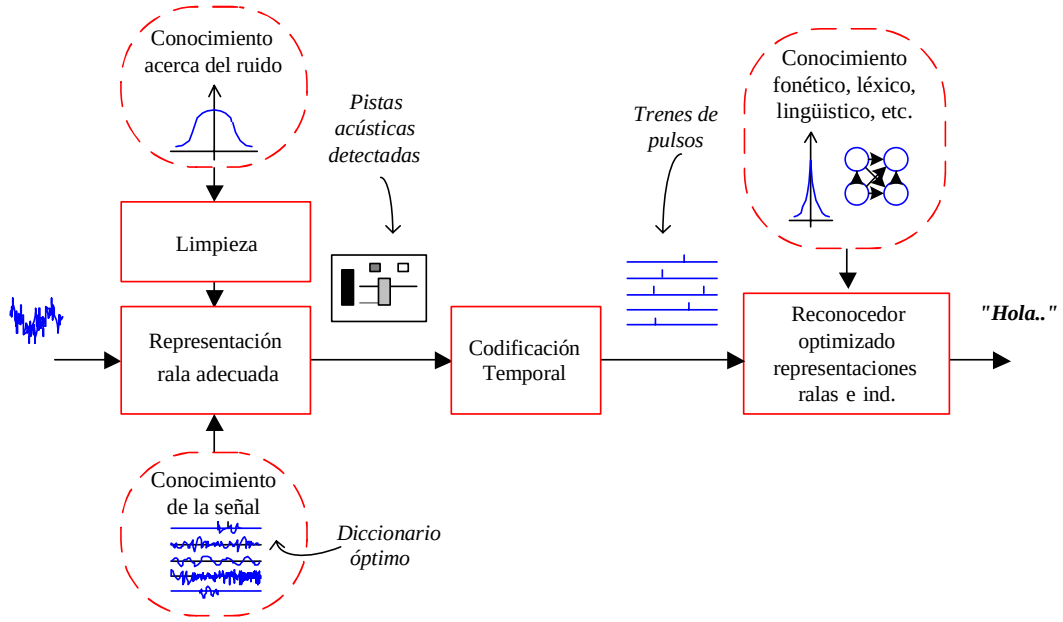

Figura 8.1: *Esquema conceptual de un posible sistema de reconocimiento automático del habla que contemple las características especiales de las representaciones ralas e independientes.*

Apéndice A

Clasificación de fonemas

Contenido

A.1. Introducción

PARA comparar diferentes representaciones se requiere establecer un contexto de aplicación de las mismas. En este apéndice se presentan los detalles constructivos de un sistema de clasificación de fonemas extraídos a partir de habla continua en idioma inglés. Como los resultados de este sistema se tomarán como referencia para las alternativas propuestas en este trabajo, es que se utilizarán algunas representaciones clásicas para el análisis del habla.

Las características principales del sistema de clasificación desarrollado son las siguientes:

- Representación basada en DFT y MFCC en escala frecuencial lineal y de mel con diferentes configuraciones [37].
- Clasificación tramo a tramo de fonemas mediante redes neuronales con retardos temporales, con una capa oculta y entrenadas por el algoritmo de retropropagación [210].
- Entrenamiento y prueba con un subconjunto de fonemas de la región “dr1” de la base de datos TIMIT [56].

- Evaluación de la robustez del sistema mediante el agregado de ruido aditivo de la base de datos NOISEX [209].
- Validación simple sobre el conjunto de prueba original de TIMIT [132, 66].

A.1.1. Importancia de los datos

Gran parte de los resultados obtenidos en esta sección depende de los datos o muestras de voz etiquetadas (corpus) utilizados para generar los patrones para entrenamiento y prueba de los clasificadores. La gran influencia sobre los resultados del corpus empleado tiene principalmente tres razones:

- La cantidad de emisiones, cantidad de hablantes y diversidad fonética determinan la complejidad de la tarea de clasificación o reconocimiento.
- De la fiabilidad de los mismos depende en gran medida la validez de los resultados obtenidos.
- La disponibilidad y difusión de la base de datos empleada repercute sobre la posibilidad de comparación con otras estrategias pasadas o futuras.

De acuerdo con estos puntos se decidió utilizar una base de datos standard del tipo citado en la bibliografía especializada y los artículos del área. Este enfoque tiene la ventaja de poder comparar los resultados con los reportados previamente. Sin embargo estas bases de datos están disponibles generalmente en idioma inglés y los resultados no son directamente extrapolables a otros idiomas (en el Apéndice B se realizan experimentos de ASR con un corpus en idioma español). Estos corpora se pueden obtener en CD-ROM a través de distintas instituciones[1]. Existen dos bases de datos "clásicas" muy utilizadas, una de ellas es TIMIT [48] para discurso continuo y la otra es NIST TI-46 [47] para palabras aisladas. Esta última es una base multi-hablante y se han reportado gran cantidad de resultados para un subconjunto denominado *E-set*, por tratarse de un conjunto de palabras altamente confundible entre sí. Este conjunto está compuesto por las palabras correspondientes al alfabeto inglés que tienen como segunda letra a "e" (como "be", "de", "ge", etc.). La baja energía y corta duración de las consonantes en relación a la vocal hacen que sea un conjunto difícil de clasificar. La base TIMIT es también multi-hablante, pero bastante mayor en tamaño, y es una de las más empleadas en el ámbito del discurso continuo por ser la más grande, completa y mejor documentada de su tipo. Por esta raxón se seleccionó, para la tarea de clasificación de fonemas extraídos del discurso continuo, se eligió TIMIT. Esta base o corpus posee una gran cantidad de

[1] Como por ejemplo:

- Linguistic Data Consortium, University of Pennsylvania, Philadelphia, USA, `ftp://www.cis.upenn.edu/pub/ldc_www/hpage.html`.
- Oregon Graduate Institute, Center for Spoken Language Understanding, Portland, USA, `http://www.cse.ogi.edu/CSLU/corpora/isolet2.html`.

fonemas en diversos ambientes y pronunciados por más de 600 hablantes diferentes. Esto constituye un total de unas 5 horas de material hablado etiquetado y casi 650 MBytes de información para procesar, lo que da una idea de las dificultades involucradas en su manipulación y utilización, así como también de la complejidad de la tarea de clasificación de los fonemas contenidos en la señal. Se podría decir que se trata de un problema de identificación de fonemas inmersos en discurso continuo e independiente del hablante [167]. Otro punto importante de mencionar es que, debido a los procedimientos de grabación, la señal registrada está prácticamente libre de ruido, situación que difícilmente se dé en condiciones distintas a las de laboratorio. Es por ello que a los fines de probar la robustez del sistema implementado se realizaron también pruebas contaminando las señales de esta base con ruido aditivo.

En lo que se sigue se describirán las características principales de TIMIT, así como su organización y tipos de archivos. Luego se explicará como se escogieron los hablantes y se detallarán las condiciones utilizadas durante la grabación de los datos. Siguiendo se describirá el tipo de texto empleado en las oraciones y la separación de los datos en entrenamiento y prueba. A continuación se presentarán los fonemas registrados y los símbolos utilizados para representarlos. Se explicarán también los criterios empleados para elegir el conjunto de emisiones que se emplearon en este sistema. Posteriormente se introducirán las características del clasificador empleado, además de las correspondientes a las representaciones utilizadas. Por último se describirán algunos detalles de implementación, mostrando finalmente los resultados del sistema que serán utilizados como referencia en los otros experimentos.

A.2. Descripción del corpus: TIMIT

Aquí se describirán las características principales del corpus elegido de manera de definir perfectamente los alcances y complejidad de la tarea de clasificación de fonemas extraídos del mismo, para más detalles remitirse a la documentación suministrada con la base de datos [56]. Esta base de datos ha sido confeccionada en forma conjunta por Texas Instruments (TI) y el Massachusetts Institute of Technology (MIT). Consiste en una serie de emisiones de voz grabadas a través de la lectura de diversos textos por un conjunto de hablantes. Esta base ha sido diseñada para la adquisición de conocimiento acústico-fonético a partir de los datos de voz y para el desarrollo y evaluación de sistemas de ASR. TIMIT contiene la voz de 630 hablantes representando las 8 mayores divisiones dialécticas del *inglés americano*, cada uno pronunciando 10 oraciones fonéticamente diversas. El corpus TIMIT incluye la señal de voz correspondiente a cada oración hablada, así como también transcripciones ortográficas, fonéticas y de palabras alineadas temporalmente. Además los datos vienen ya divididos en subconjuntos de entrenamiento y prueba balanceados para cobertura dialéctica y fonética, lo que facilita también la comparación de resultados. La versión de la base de datos empleada en este trabajo es la 1-1 (1 de Octubre de 1990).

TIMIT contiene un total de 6300 oraciones, 70 % de los hablantes son masculinos y 30 % son femeninos. El material de texto consiste de 2 *oraciones de dialecto* (SA), 450

oraciones fonémicamente compactas (SX), y 1890 *oraciones fonéticamente diversas* (SI). Cada hablante lee las 2 SA, 5 de las SX y 3 de las SI.

A.2.1. Organización de los datos

El CD-ROM contiene una estructura de árbol de directorios jerárquica que permite fácil acceso a los datos en forma automática (por medio del programa de procesamiento). La estructura de este árbol es la siguiente :

```
/<Corpus>/<Uso>/<Dialecto>/<Sexo><Hablante>/<Oración>.<Tipo_Archivo>
```

donde,

```
CORPUS = timit
USO = train|test (entrenamiento|prueba)
DIALECTO = dr1|dr2|dr3|dr4|dr5|dr6|dr7|dr8 (regiones dialécticas)
SEXO = f|m
HABLANTE = <INICIALES><DÍGITO>
```

donde,

```
INICIALES = Iniciales del Hablante (3 letras)
DÍGITO = número 0-9 para diferenciar hablantes iniciales iguales
ORACIÓN = <TIPO_TEXTO><NÚMERO_ORACIÓN>
```

donde,

```
TIPO_TEXTO = sa|si|sx
NÚMERO_ORACIÓN = 1 ... 2342
TIPO_ARCHIVO = wav|txt|wrd|phn (datos de voz|texto|palabras|fonemas)
```

Por ejemplo : `/timit/train/dr1/fcjf0/sa1.wav`, corresponde al corpus TIMIT, conjunto de entrenamiento, región dialéctica 1, sexo femenino, hablante "cjf0", texto oración "sa1", archivo señal de voz.

A.2.2. Tipos de archivo

TIMIT incluye varios archivos asociados con cada emisión (Tabla A.1). Así mismo se incluye un archivo diccionario con todas las palabras contenidas en el corpus y su transcripción fonética (léxico), otro con todas las oraciones empleadas, y otro con información específica de los hablantes. Los archivos de voz poseen el formato con cabecera NIST SPHERE que ha sido diseñado para facilitar el intercambio de datos de señales de voz en CD-ROM. La cabecera NIST es una estructura orientada a objetos de 1024 bytes que precede a los datos propiamente dichos. En ella se almacena información acerca de la emisión como ser frecuencia de muestreo, bits por muestra, identificación del hablante y la oración, etc.

Los archivos de transcripción tienen la siguiente forma :

Tabla A.1: *Tipos de archivo asociados a cada emisión del corpus de TIMIT.*

Tipo de archivo	Descripción
.wav	Archivo de voz con cabecera tipo SPHERE.
.txt	Transcripción ortográfica de las palabras dichas por el hablante.
.wrd	Transcripción de palabras alineada temporalmente con archivo de voz.
.phn	Transcripción fonética alineada temporalmente con archivo de voz.

```
<MUESTRA_COMIENZO><MUESTRA_FINAL> <TEXTO><nueva-línea>
.
.
.
<MUESTRA_COMIENZO><MUESTRA_FINAL> <TEXTO><nueva-línea>
```

donde,

```
MUESTRA_COMIENZO = Muestra inicial del segmento (número entero >=0)
MUESTRA_FINAL = Muestra final del segmento (número entero <= última
                muestra)
TEXTO = <ORTOGRAFÍA> | <ETIQUETA_PALABRA> | <ETIQUETA_FONÉTICA>
```

donde,

```
ORTOGRAFÍA = Transcripción ortográfica completa del texto.
ETIQUETA_PALABRA = Una palabra de la transcripción ortografía.
ETIQUETA_FONÉTICA = Un código de transcripción fonética.
```

Por ejemplo las transcripciones de la emisión en `timit\test\dr1\mjsw0\si1640.wav` serían :
Ortografía (.txt):

```
0 26112 How did one join them?
```

Palabras (.wrd) :

```
2276 5111 how
5111 8003 did
8003 12560 one
12560 19174 join
19174 22747 them
```

Fonética (.phn) :

```
0 2276 h#
2276 3320 hh
3320 5111 aw
5111 5931 dcl
```

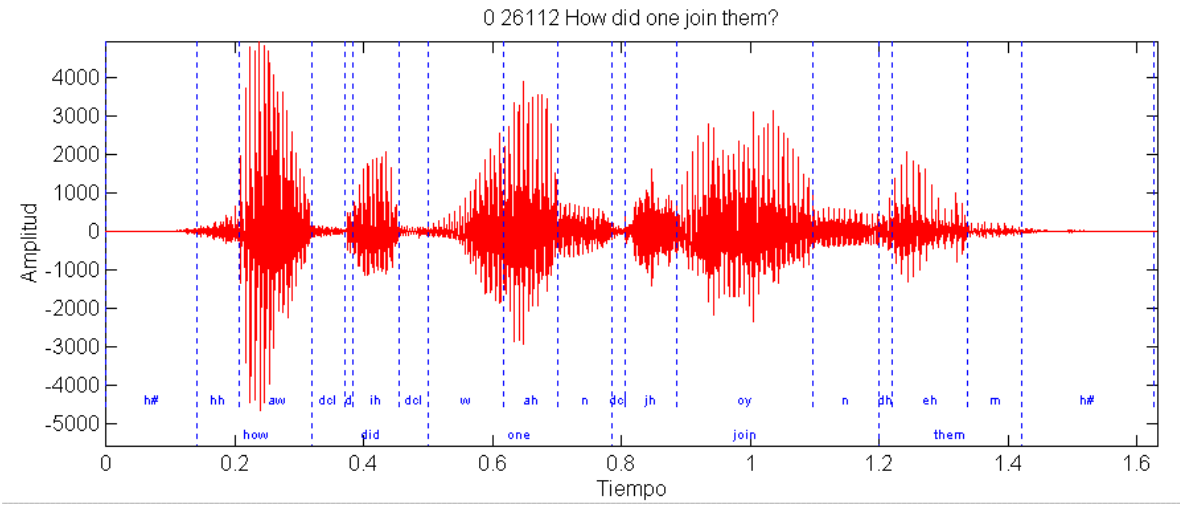

Figura A.1: *Señal de voz con etiquetas de palabras y fonemas*


```
5931 6143 d
6143 7272 ih
7272 8003 dcl
8003 9869 w
9869 11224 ah
11224 12560 n
12560 12879 dcl
12879 14160 jh
14160 17560 oy
17560 19174 n
19174 19533 dh
19533 21400 eh
21400 22747 m
22747 26000 h#
```

En la Figura A.1 se puede apreciar la emisión correspondiente con la superposición de las etiquetas de palabras y fonemas alineadas temporalmente con la misma.

A.2.3. Selección de hablantes

Las 10 oraciones leídas por cada uno de los 630 hablantes representan aproximadamente 30 segundos de voz por hablante. En total el corpus contiene aproximadamente 5 horas de voz. Todos los participantes seleccionados fueron hablantes nativos de inglés americano. Además todos fueron calificados como sin patologías clínicas del habla por un especialista del área. Se detectaron en algunos sujetos pequeñas anormalidades en el habla o la audición que fueron anotadas en los archivos de información de los hablantes que acompañan la base de datos. Los hablantes fueron seleccionados para ser representativos de diferentes regiones dialécticas geográficas de los Estados Unidos de acuerdo

Tabla A.2: *Distribución de los hablantes por región dialéctica en el corpus de TIMIT.*

Región Dialéctica Nombre	Código	Nº Hablantes masculinos (%)	Nº Hablantes femeninos (%)	Nº Total de hablantes
New England	1	31 (63)	18 (27)	49 (8)
Northern	2	71 (70)	31 (30)	102 (16)
North Midland	3	79 (67)	23 (23)	102 (16)
South Midland	4	69 (69)	31 (31)	100 (16)
Southern	5	62 (63)	36 (37)	98 (16)
New York City	6	30 (65)	16 (35)	46 (7)
Western	7	74 (74)	26 (26)	100 (16)
Army Brat	8	22 (67)	11 (33)	33 (5)
Nº Total de hablantes		438 (70)	192 (30)	630 (100)

con la región donde vivieron en su niñez. En la Tabla A.2 se presenta la distribución de los hablantes en cada región dialéctica.

A.2.4. Condiciones de grabación

Las grabaciones fueron hechas en una cabina de grabación aislada de ruidos usando un sistema semiautomático para la presentación del texto al hablante y la grabación. Los datos fueron digitalizados a una frecuencia de muestreo de 20 KHz (16 bits) con un filtro anti-alias en 10 KHz. La voz fue filtrada digitalmente, nivelada (debiased) y submuestreada a 16 KHz [48]. A los sujetos se los estimuló con una señal de ruido de fondo de bajo nivel a través de auriculares para suprimir la inusual calidad de voz producida por el efecto de aislación de la cabina. También se les pidió que leyeran el texto con "voz natural".

A.2.5. Texto del corpus

Las oraciones SA fueron diseñadas para exponer las diferencias dialécticas y fueron leídas por todos los hablantes. Las oraciones SX fueron diseñadas a mano para proveer una buena cobertura en cuanto a pares de fonemas, con ocurrencias extra de contextos fonéticos difíciles o de interés particular. Cada hablante leyó 5 de estas oraciones y cada una fue leída por 7 hablantes. Las oraciones SI fueron seleccionadas de fuentes de texto existentes para agregar diversidad en los tipos de oraciones y los contextos fonéticos. El criterio de selección maximiza la variedad de contextos alofónicos encontrados en los textos. Cada hablante leyó 3 de estas oraciones y cada una fue leída solo una vez. En la A.3 se muestra la distribución del material de texto del corpus.

A.2.6. Subdivisión en entrenamiento y prueba

Existen diferentes métodos para estimar la capacidad de generalización de un clasificador [132]. Es ampliamente conocido que las tasas de error tienden a sesgarse si se

Tabla A.3: *Material de texto del corpus TIMIT*

Tipo Oración	Nº Oraciones	Nº Hablantes /oración	Total	Nº Oraciones /Hablante
Dialecto (SA)	2	630	1260	2
Compactas (SX)	450	7	3150	5
Diversas (SI)	1890	1	1890	3
Total	2342		6300	10

estiman a partir de los mismos datos que se utilizaron en el proceso de aprendizaje o entrenamiento del clasificador. Una forma muy sencilla (y difundida) de abordar el problema consiste en separar los datos en un conjunto de entrenamiento y otro de prueba. En la sección sobre el clasificador se volverá sobre este punto. Sin embargo, aquí es importante notar que la cantidad de datos involucrados en este problema hace difícil la utilización de métodos más precisos de validación para estimar el error.

El material contenido en TIMIT fue dividido en conjuntos de entrenamiento y prueba siguiendo los siguientes criterios :

- Del 20 al 30 % del corpus sería usado para propósitos de prueba dejando el restando 70 a 80 % para entrenamiento.
- Ningún hablante debería aparecer en ambos conjuntos.
- Todas las regiones dialécticas deberían estar representadas en ambos conjuntos, con al menos un hablante masculino y uno femenino de cada dialecto.
- La cantidad de material de texto repetido en ambos conjuntos debería minimizarse o, en lo posible, eliminarse.
- Todos los fonemas deberían estar cubiertos en el material de prueba, preferiblemente en diferentes contextos.

Estos criterios, junto con lo que se mencionó en la introducción, hacen que el problema de reconocimiento o clasificación sea independiente del hablante y del texto, lo que implica un grado de complejidad apreciable teniendo en cuenta la cantidad de material disponible.

A.2.7. Códigos de símbolos fonémicos y fonéticos

En la Tabla A.4 se muestran los símbolos fonémicos y fonéticos usados en el léxico de TIMIT y en las transcripciones fonéticas. Estos incluyen marcadores de intensidad (stress) {1,2} encontrados solo en el léxico y los siguientes símbolos que ocurren solo en las transcripciones:

1. Los intervalos de cierre u oclusión de las oclusivas los cuales se distinguen de la liberación o explosión de las mismas. Los símbolos de la oclusión para */b/*, */d/*,

/g/, */p/*, */t/*, */k/* son */bcl/*, */dcl/*, */gcl/*, */pcl/*, */tck/*, */kcl/*, respectivamente. Las porciones de oclusión de */jh/* y */ch/*, son */dcl/* y */tcl/*.

2. Alófonos que no ocurren en el léxico. El uso de determinado alófono puede depender del hablante, del dialecto, la velocidad de emisión y el contexto fonémico entre otros factores. Dado que el uso de estos alófonos es difícil de predecir no han sido usados en las transcripciones fonéticas del léxico.

3. Otros símbolos incluyen dos tipos de silencio, pau indicando una pausa, y epi, denotando el silencio epentético que es frecuentemente encontrado entre una fricativa y una semivocal o nasal, además de h#, usado para marcar el silencio y/o no aparición de eventos de voz encontrado al principio o al final de la señal.

La cantidad total de símbolos que se pueden utilizar en la clasificación es de 52. Estos se distribuyen como 8 tipos de fonemas oclusivos, 2 africados, 15 fricativos, 7 semivocales y glides y 20 vocales.

A.3. Datos elegidos para los experimentos

Como se puede apreciar la cantidad de símbolos o fonemas a clasificar y la de emisiones y hablantes es demasiado grande para intentar realizar los todos los experimentos con toda la base de datos. Por esta razón se debieron establecer algunos criterios para utilizar un subconjunto menor de los fonemas y hablantes de la base y sin embargo poder "extrapolar" los resultados a todo el conjunto. Estos criterios fueron :

- Utilizar un subconjunto de fonemas de relativa dificultad de diferenciación.
- Cubrir los tipos de fonemas más importantes.
- Disminuir la cantidad de hablantes y la diversidad de dialectos.

En base a estos criterios el subconjunto de fonemas elegido fue: */b/*, */d/ /jh/ /eh/ /ih/*. Las razones de esta elección se exponen brevemente a continuación. Se sabe por experimentos psico-acústicos que las consonantes */b/* y */d/* del tipo oclusivo son difíciles de distinguir en varios contextos. Por otra parte el fonema */jh/* es africado con lo que se incluyen las características especiales de este grupo (que posee una componente oclusiva seguida de una fricativa). Además, éstas son algunas de las consonantes iniciales del conjunto E-Set del corpus de palabras aisladas TI-46 ya mencionado, que ha probado también ser un subconjunto de palabras difícil de clasificar por medios automáticos. Para agregar a este grupo algunas vocales se eligieron */eh/* e */ih/* cuya distancia en el espacio de formantes es muy pequeña. Esto las convierte en otro grupo altamente confundible. En la Figura A.2 se puede observar la distribución de las vocales del inglés en función de F_1 y F_2. De esta manera el subconjunto está formado por 5 fonemas (que constituyen un 10 % del total).

Siguiendo los criterios expuestos se eligió la región "dr1" que posee casi 50 hablantes (ver Tabla A.2). Se respetaron las divisiones en conjuntos de entrenamiento y prueba

Tabla A.4: *Símbolos fonéticos utilizados en la transcripción*

Tipo	Símbolo	Ejemplo	Transcripción fonética o descripción
Oclusivas	b	Bee	BCL B iy
	d	Day	DCL D ey
	g	gay	GCL G ey
	p	pea	PCL P iy
	t	tea	TCL T iy
	k	key	KCL K iy
	dx	muddy, dirty	m ah DX iy, dcl d er DX iy
	q	bat	bcl b ae Q
Africadas	jh	joke	DCL JH ow kcl k
	ch	choke	TCL CH ow kcl k
Fricativas	s	sea	S iy
	sh	she	SH iy
	z	zone	Z ow n
	zh	azure	ae ZH er
	f	fin	F ih n
	th	thin	TH ih n
	v	van	V ae n
	dh	then	DH e n
Nasales	m	mom	M aa M
	n	noon	N uw N
	ng	sing	s ih NG
	em	bottom	b aa tcl t EM
	en	button	b ah q EN
	eng	washington	w aa sh ENG tcl t ax n
	nx	winner	w ih NX axr
Semivocales y	l	lay	L ey
Glides	r	ray	R ey
	w	way	W ey
	y	yacht	Y aa tcl t
	hh	hay	HH ey
	hv	ahead	ax HV eh dcl d
	el	bottle	bcl b aa tcl t EL
Vocales	iy	beet	bcl b IY tcl t
	ih	bit	bcl b IH tcl t
	eh	bet	bcl b EH tcl t
	ey	bait	bcl b EY tcl t
	ae	bat	bcl b AE tcl t
	aa	bott	bcl b AA tcl t
	aw	bout	bcl b AW tcl t
	ay	bite	bcl b AY tcl t
	ah	but	bcl b AH tcl t
	ao	bought	bcl b AO tcl t
	yo	boy	bcl b YO
	ow	boat	bcl b OW tcl t
	uh	book	bcl b UH kcl k
	uw	boot	bcl b UW tcl t
	ux	toot	tcl t UX tcl t
	er	bird	bcl b ER dcl d
	ax	about	AX bcl b aw tcl t
	ix	debit	dcl d eh bcl b IX tcl t
	axr	butter	bcl b ah dx AXR
	ax-h	suspect	s AX-H s pcl p eh kcl k tcl t
Otros	pau		pausa
	epi		silencio epentético
	h#		marcador de comienzo / fin
	1		marcador de stress primario
	2		marcador de stress secundario

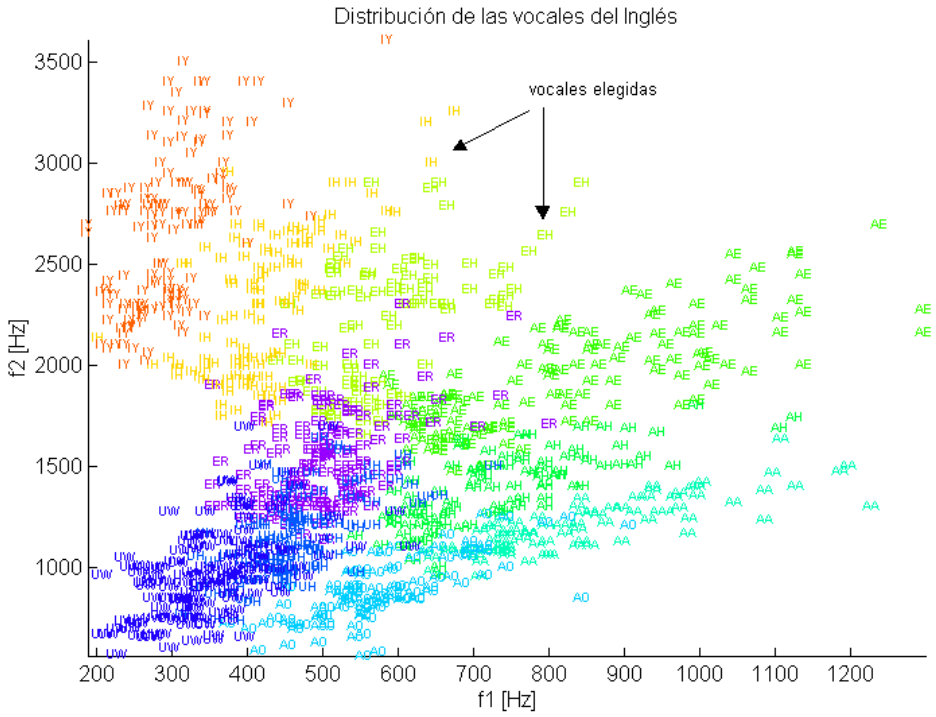

Figura A.2: *Distribución de las vocales del inglés. Los datos utilizados para construir la gráfica fueron tomados de [152].*

propuesta en TIMIT de manera que la distribución final de los fonemas elegidos en cada región se puede apreciar en la Tabla A.5.

Se debe aclarar que en primera instancia se incluyó la oclusión de */b/*, */d/* y */jh/* (símbolos fonéticos BCL y DCL en la Tabla A.4), pero debido a que constituían prácticamente silencios se quitaron en los experimentos definitivos.

Aquí interesa seguir la evolución temporal de distintas características de la señal por lo que varias de las técnicas utilizan análisis por tramos, a partir de una ventana deslizante sobre la señal. El ancho de esta ventana, y por lo tanto la duración de cada patrón generado, es del orden de los 10 mseg.

Tabla A.5: *Distribución de los fonemas elegidos en entrenamiento y prueba.*

Fonema	Nº Entrenamiento (%)	Nº Prueba (%)	Total
/b/	183 (14.4)	59 (15.9)	242
/d/	300 (23.6)	90 (24.2)	390
/jh/	104 (8.2)	20 (5.3)	124
/eh/	316 (24.8)	93 (25.1)	409
/ih/	370 (29.0)	109 (29.4)	479
Total	1273	371	1644

Tabla A.6: *Duración promedio de los fonemas elegidos en dr1.*

Fonema	/b/	/d/	/jh/	/eh/	/ih/
Muestras	300	378	916	1419	1218
Tiempo (mseg.)	19	24	57	89	76
Tramos	2	3	7	11	10

En la Tabla A.6 se muestra la duración promedio de los fonemas elegidos en la región "dr1". Se puede apreciar gran variación de duración para los distintos fonemas. Por ejemplo */b/* y */d/* son generalmente muy cortas frente a */eh/*, */ih/* o incluso */jh/*, esto junto con sus respectivas distribuciones (Tabla A.5) lleva a que una vez procesadas la cantidad de tramos o patrones generados para cada clase sea muy diferente, produciendo algunos problemas sobre las clases menos representadas. Por otra parte la diferencia en duración también plantea problemas en cuanto a los procesos involucrados en la clasificación dinámica de los patrones.

Luego de ser entrenado con habla limpia, el clasificador fue probado con habla contaminada con diferentes tipos de ruido a distintas relaciones señal ruido (SNR), para determinar la robustez del mismo. Se utilizó ruido blanco y tipo murmullo de la base NOISEX-92 [209]. Los datos de ruido blanco fueron digitalizados de un generador de ruido analógico de alta calidad (Wandel & Goltermann) a 19.98 KHz y 16 bits, con igual energía en todo el ancho de banda. El ruido de conversación se grabó mediante un dispositivo DAT (*digital audio tape*) equipado con un micrófono tipo condensador. La fuente del "murmullo" fueron 100 personas hablando en una cantina. El radio del cuarto fue de unos dos metros; por lo que, la voces individuales son ligeramente audibles. El nivel de sonido durante el proceso de grabación fue de 88 dB SPL.

El ruido se remuestreó a 16 KHz y se mezcló con las señales de habla de TIMIT en distintas proporciones para lograr distintas SNR. Se debe aclarar que en todos los casos se ha obviado el conocido efecto Lombard (por el cual el hablante modifica en forma importante su fonación en presencia de ruido ambiente [121, 108, 95]) y se ha procedido por simple adición de las señales.

A.4. Redes neuronales con retardos temporales

Existen muchas técnicas de clasificación que se han aplicado al caso de los fonemas [37]. En este sistema se utiliza una *red neuronal con retardos temporales* (TDNN) [210] para realizar una clasificación de los fonemas escogidos, luego de obtener la representación con algunas de las técnica planteadas. Una razón para elegir este tipo de clasificador es que permite la utilización del clásico algoritmo de retropropagación casi sin modificaciones. En pruebas de comparación iniciales [167] este tipo de redes funcionó mejor que otras arquitecturas de redes recurrentes simples como las de Jordan y Elman [45]. Las redes neuronales con retardos consisten en unidades elementales similares a las de un perceptrón multicapa, pero modificadas a fin de que puedan procesar información generada en distintos instantes. A las entradas sin retardos de cada neurona, se les agregan

las entradas correspondientes a instantes distintos. De este modo, una unidad neuronal de estas características es capaz de relacionar y procesar conjuntamente la entrada actual con eventos anteriores.

En todos los experimentos se ajustaron pesos y umbrales con el algoritmo de retropropagación con momento. Los parámetros de aprendizaje se mantuvieron fijos durante el entrenamiento. Luego de bastante experimentación los siguientes valores parecieron adecuados. La tasa de aprendizaje se estableció en 0.1, el factor de momento también en 0.1 y los pesos iniciales aleatorios entre -0.3 y 0.3. Entradas y salidas se normalizaron entre 0 y 1 utilizando los valores máximos y mínimos correspondientes, pero extendiendo el rango en un 5 % con respecto al archivo de entrenamiento.

Se ensayaron diferentes configuraciones de la red neuronal y del algoritmo de aprendizaje (cantidad de nodos en la capa oculta, cantidad de retardos y otros parámetros) de las que se reportan aquí sólo las mejores. Durante los experimentos definitivos se intentó mantener constante el número total de pesos y umbrales, de manera de que la estructura no fuera un factor que pesara demasiado en la diferencia de desempeño entre las distintas alternativas de representación.

La mayoría de las técnicas implementadas generan patrones de 128 dimensiones o coeficientes. Esto produce, de acuerdo con el solapamiento de las ventanas de análisis, una tasa de 1 tramo o patrón cada 1 a 8 mseg. Como las redes utilizadas son dinámicas en general la información que procesan por unidad de tiempo corresponde a varios tramos (el actual y uno o dos anteriores). La cantidad de salidas corresponde a la cantidad de clases o fonemas a clasificar, que en este caso son 5 (*/b/*, */d/*, */jh/*, */ih/* y */eh/*).

Para estimar la capacidad de generalización del clasificador se empleó el método más sencillo, que consiste en separar los datos en un conjunto de entrenamiento y otro de prueba. Para ello se utilizó la partición de los datos que ya viene en la base TIMIT y que obedece a criterios que pretenden asegurar que los datos de prueba sean representativos de las clases implicadas, mientras que sean además distintos a los de entrenamiento.

Para todos los experimentos el aprendizaje se detuvo en el pico de generalización con respecto al conjunto de prueba [167]. Además se corrieron al menos dos pruebas con diferentes condiciones iniciales (de las que se reporta sólo la mejor). Esto puede hacer parecer que los resultados con respecto a entrenamiento no son tan buenos como los esperados. El error con respecto al archivo de prueba se calculó cada 2000 iteraciones del algoritmo (aproximadamente 4 veces por época del archivo de entrenamiento).

La forma de las salidas deseadas utilizadas para vocales y consonantes se puede apreciar en la Figura A.3. La razón de la diferencia entre ambas es que las vocales mantienen sus características relativamente constantes durante la emisión, por lo que bastan uno o dos tramos de la representación para identificar el fonema. Sin embargo, en el caso de las consonantes solo se puede emitir una clasificación segura después de que se ha seguido la evolución de la emisión durante algún tiempo. En la Figura A.4 se observa un ejemplo para la DFT en escala de mel calculada a partir de una emisión típica de TIMIT.

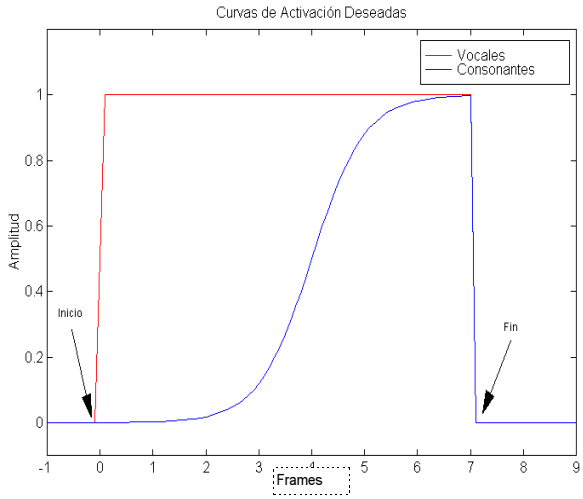

Figura A.3: *Curvas de activación deseada utilizadas en el clasificador para vocales y consonantes. Estas obedecen a las diferencias en la evolución dinámica de ambos tipos de fonemas.*

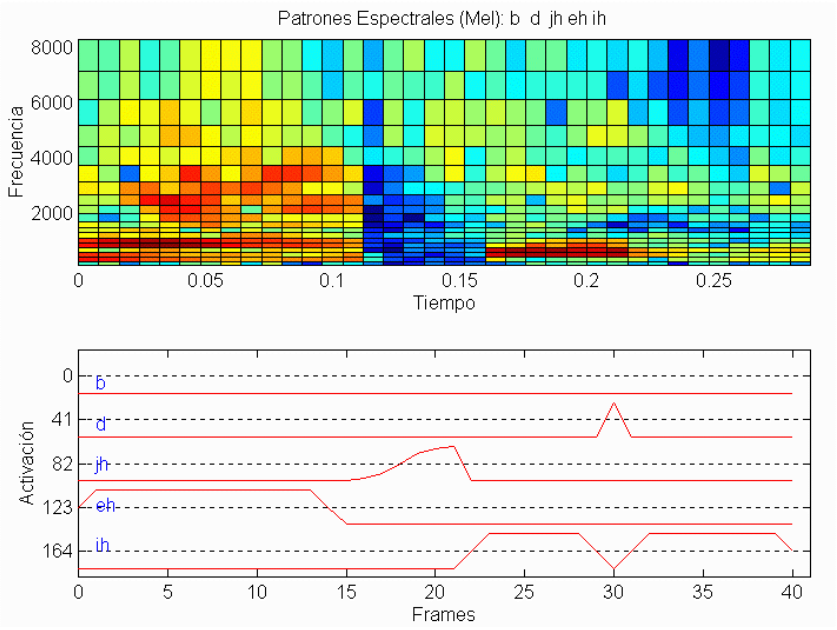

Figura A.4: *Ejemplo de patrones espectrales en escala de mel y las correspondientes curvas de activación deseada para un conjunto de fonemas extraídos de una emisión típica de TIMIT.*

A.5. Representación utilizada

Los patrones para entrenar el clasificador se generaron mediante algunas de las técnicas clásicas presentadas en la Sección 4.5. Se realizaron experimentos de clasificación de fonemas con Fourier y coeficientes cepstrales, en escala frecuencial lineal y psicoacústica de mel. Como ya se mencionó estos son algunos de los más utilizados en el campo del análisis del habla. Para la DFT los coeficientes de la representación se obtuvieron luego de calcular la magnitud en decibeles. Para el cálculo de cepstrum se utilizó la versión real. Para ambos casos se empleó una ventana de Hamming para cada tramo. En los experimentos se variaron los anchos de ventana para el análisis por tramos correspondiente, la cantidad de coeficientes generada, así como también la arquitectura de la red neuroanl utilizada. En todos los casos se utilizó una ventana de Hamming sobre cada tramo.

A.6. Detalles de implementación

Para los experimentos de clasificación basados en redes neuronales se empleó el programa comercial NeuroShell v2.4[2]. Para la generación de los patrones de entrenamiento, a partir de los datos de TIMIT, se confeccionaron una serie de programas en Matlab[3], con la ayuda de varios toolbox específicos. Las rutinas para adicionar ruido a los datos se implementaron en GNU C++ estándar [98].

A.7. Resultados de referencia

A continuación se presentan los resultados obtenidos con el sistema descripto y que serán tomados como referencia para los experimentos realizados con otras representaciones. Los resultados se presentan en términos del porcentaje de tramos bien clasificados por la red para el archivo de entrenamiento (TRN) y el de prueba (TST) (promedio sobre todo el archivo). Se presentan también los resultados individuales para cada clase que corresponden a la diagonal de la matriz de confusión. El cálculo del porcentaje de aciertos se realiza de acuerdo con la estrategía planteada en [164], donde no se toman en cuenta aquellas activaciones de la red cuyo valor es inferior a un umbral pequeño prefijado (es decir que se ignoran los denominados *errores de borrado*).

En la Tabla A.7 se muestran los porcentajes aciertos para los experimentos de clasificación de fonemas con redes neuronales mediante las representaciones generadas con: Fourier y Fourier en escala de Mel, Cepstra y Cepstra en escala de Mel para diferentes anchos de ventana y cantidad de coeficientes (indicados entre paréntesis para cada caso respectivamente). El último experimento de la tabla corresponde a la representación obtenida mediante 13 coeficientes cepstrales en escala de mel, más un coeficiente de energía, a partir de tramos de 128 muestras. Esta representación es bastante usual y se denomina como MFCC+E.

[2]NeuroShell 2 Release 3.0, Ward Systems Group, Inc. 1998. `http://www.wardsystems.com`
[3]Matlab Release 14, The MathWorks, Inc. 2004. `http://www.mathworks.com`

Tabla A.7: *Porcentajes de tramos bien clasificados globales y para cada fonema en función de la representación convencional utilizada para generar los patrones. Se indica también la estructura del clasificador correspondiente.*

Nº	Experimento	Estructura Red	TRN	TST	/b/	/d/	/jh/	/eh/	/ih/
1	Fourier (256, 128)	128+128/150/5	**79.67**	**77.53**	**52.60**	**63.60**	**97.20**	**83.60**	**71.70**
2	Cepstra (512,50)	50+50/88/5	78.31	70.75	49.21	67.78	66.18	79.95	66.45
3	Mel Fourier (256, 20)	20+20+20/135/5	**82.56**	**81.83**	**72.95**	**79.28**	**90.19**	**84.74**	**78.33**
4	Mel Cepstra (512,20)	20+20/73/5	76.84	73.91	53.97	65.56	79.41	68.98	78.73
5	Mel Cepstra (512,16)	16+16/32/5	77.78	76.39	58.73	67.78	86.76	79.72	75.88
6	Mel Cepstra (256,20)	20+20+20/135/5	81.89	79.10	31.58	45.45	97.33	80.71	77.87
7	Mel Cepstra (256,16)	16+16+16/32/5	**79.96**	**79.57**	**10.81**	**55.56**	**92.31**	**82.60**	**79.93**
8	Mel Cepstra (128,20)	20+20/60/5	75.93	74.57	69.44	60.17	93.03	71.22	77.11
9	Mel Cepstra + En. (128,14)	14+14/28/5	77.39	77.28	46.51	75.38	91.11	80.56	74.40

Tabla A.8: *Porcentajes de clasificación para MFCC+E a partir de señales de voz con diferentes tipos y cantidades de ruido aditivo.*

Tipo Ruido	SNR (dB)	Estructura Red	TRN	TST	/b/	/d/	/jh/	/eh/	/ih/
Limpio	∞	14+14/28/5	77.39	77.28	46.51	75.38	91.11	80.56	74.40
Blanco	50			74.13	60.00	81.94	77.46	71.85	76.23
Blanco	25			73.05	34.04	91.25	74.65	79.23	67.06
Blanco	15			60.07	3.77	97.80	64.47	77.10	40.58
Blanco	10			49.62	1.79	95.10	59.26	71.04	24.70
Blanco	5			38.41	0.00	86.11	65.52	59.87	12.31
Blanco	0			29.04	0.00	67.89	71.91	45.89	6.47
Limpio	∞	14+14/28/5	77.39	77.28	46.51	75.38	91.11	80.56	74.40
Murmullo	50			73.91	56.86	81.43	78.87	71.43	76.29
Murmullo	25			73.61	60.87	67.95	67.65	72.55	76.26
Murmullo	15			72.53	65.00	53.52	42.42	74.50	74.80
Murmullo	10			71.60	54.55	39.34	32.20	75.29	73.58
Murmullo	5			66.51	23.53	18.52	18.03	75.44	65.84
Murmullo	0			60.08	6.52	4.82	1.67	72.98	59.41

Para evaluar la robustez del sistema frente al ruido aditivo, en la Tabla A.8 se muestran los resultados de los experimentos de clasificación anteriores para el caso de la representación generada mediante MFCC+E. En este caso se ha entrenado la red con habla limpia y posteriormente se ha probado el desempeño de la misma con habla contaminada con diferentes tipos y cantidades de ruido. Es posible apreciar claramente en esta tabla la degradación en el desempeño a medida que disminuye la SNR, siendo más notorio para el caso del ruido blanco.

Apéndice B

Reconocimiento de habla continua

Contenido

B.1. Introducción

En este trabajo se requiere comparar diferentes representaciones de la señal de voz a través del desempeño de sistemas artificiales de clasificación (como el presentado en el Apéndice A) o reconocimiento automático. Por lo tanto se presentan en esta sección los detalles constructivos de un sistema de reconocimiento automático de habla continua en idioma español con independencia del hablante. Este sistema constituye la referencia (en inglés *base-line*) para la comparación de las representaciones convencionales con algunas de las alternativas desarrolladas en este trabajo. El mismo está basado en *modelos ocultos de Markov* (HMM, del inglés Hidden Markov Models), y constituye un sistema del "estado del arte" para el tamaño de vocabulario con que se realizaron las pruebas. El sistema implementado posee las siguientes características[1]:

- Preprocesamiento basado en 12 coeficientes cepstrales en escala de mel, con coeficientes delta y aceleración (MFCC+E+D), ventana de Hamming, preénfasis y eliminación de media temporal [37].

[1] Una versión inicial de este sistema se desarrollo y utilizó originalmente en [134]

- Para los fonemas del español y el silencio se utilizaron HMMs de 5 estados y para las pausas cortas (de los finales de palabra) se utilizaron 3 estados. Se realizaron pruebas con modelos continuos y semicontinuos (los resultados que aquí se presentan son con éstos últimos) [160].
- Modelado estadístico del lenguaje mediante bi-gramáticas suavizadas por el método de *backing-off* [91].
- Entrenamiento y prueba con la base de datos Albayzin (en español), corpus geográfico, subconjunto minigeo (vocabulario mediano) [17].
- Evaluación de la robustez del sistema mediante el agregado de ruido aditivo de la base de datos NOISEX [209].
- Validación cruzada por el método *leave-k-out* con 10 particiones [132, 66].

En lo que se sigue se describirán las características principales de los datos de Albayzin, así como su organización y tipos de archivos. Luego se describirá el subconjunto seleccionado para los experimentos junto con las condiciones utilizadas durante la grabación de los mismos y la separación de los datos para entrenamiento y prueba. Posteriormente se introducirán las características del sistema empleado basado en HMM, además de las correspondientes a las representación utilizada. Por último se describirán algunos detalles de implementación, mostrando finalmente los resultados del sistema que serán utilizados como referencia en experimentos con representaciones alternativas.

B.2. Descripción del corpus: Albayzin

El corpus de habla Albayzin ha sido desarrollado con el objetivo de contribuir al desarrollo y la evaluación de sistemas de reconocimiento y procesamiento del habla. El diseño fue realizado a principios de la década del 90 [17] aunque la producción completa se finalizó en 1998. El proyecto "Albayzin" fue llevado adelante por 5 Universidades de España:

- Universidad de Granada (UGR) Dpto. ETC
- Universidad Politécnica de Valencia (UPV) Dpto. SIS
- Universidad Politécnica de Madrid (UPM) Dpto. IE y Dpto. SSR
- Universidad Autónoma de Barcelona (UAB) Dpto. FE
- Universidad Politécnica de Catalunya (UPC) Dpto. TSC

El corpus se compone de 15600 elocuciones pronunciadas por 152 hombres y 152 mujeres de entre 18 y 55 años de edad. Los hablantes pertenecen a la variedad central del castellano, en su mayor parte de las comunidades de Castilla-La Mancha, Castilla-León, Cantabria y Madrid. El material que contiene el corpus es leído aunque para

el diseño se ha utilizado como punto de partida un estudio del habla espontánea. En promedio las frases poseen 4 seg. de duración y fueron muestreadas a 16 KHz con una resolución de 16 bits. Se pudo medir una relación señal a ruido promedio de 48 dB[2].

Las frases de la base de datos se encuentran distribuidas en 3 corpus bien diferenciados:

1. Corpus fonético: es un conjunto genérico de 6800 elocuciones equilibradas fonéticamente, sin restricciones sintáctico-semánticas, que brinda un marco de referencia de la lengua castellana [136]. Para el diseño de este corpus se han considerado tanto la proporción como la cobertura de las elocuciones de cada alófono en cada contexto. El corpus ha sido dividido en dos subconjuntos, uno de aprendizaje y otro de prueba. El subconjunto de aprendizaje consiste en la elocución de 200 frases diferentes por 4 locutores y 160 frases por otros 25 locutores (4800 elocuciones en total). El subconjunto de prueba consiste en 40 frases diferentes pronunciadas por 50 locutores.

2. Corpus geográfico: es un conjunto de 6800 elocuciones de frases dependientes de la aplicación, con restricciones semánticas y sintácticas relacionadas con la consulta de una base de datos de geografía española [39]. Las construcciones sintácticas reflejan la forma natural del habla en el lengua castellana. Para extraerlas se analizaron 14918 frases obtenidas mediante entrevistas a 408 personas que intentaban obtener información sobre geografía española. Todas las frases se clasificaron según criterios lingüísticos, semánticos y de complejidad estructural. El subconjunto de entrenamiento consta de 50 frases diferentes pronunciadas por 88 locutores y el subconjunto de prueba consta de otras 50 frases diferentes pronunciadas por 48 locutores.

3. Corpus "Lombard": se compone de 2000 elocuciones de los corpus anteriores, producidas en condiciones adversas. El efecto Lombard consiste en un conjunto de modificaciones de la voz que se producen cuando el locutor se encuentra sometido a un nivel alto de ruido. Este corpus consta de las elocuciones de 40 locutores que pronuncian 50 frases diferentes cada uno.

B.3. Datos elegidos para los experimentos

Los datos elegidos para los experimentos forman parte del subconjunto 1 del corpus geográfico(SC1). Éste subconjunto contiene 600 elocuciones y está diseñado con las pautas generales del corpus geográfico [38]. En la tabla B.1 se resumen las características más importantes de este subconjunto.

La base de datos final está constituida por las señales de voz convenientemente filtradas y muestreadas, organizadas de forma que sea sencillo acceder a diferentes subgrupos, atendiendo a características como el sexo, la variedad dialectal, etc. En las señales se

[2]A pesar de que las señales utilizadas tienen ya incluido algo de ruido se supondrán "limpias" para los experimentos (SNR=∞ dB)

Tabla B.1: *Características principales del subconjunto 1 del corpus geográfico de Albayzin.*

Característica	Valor
Total de elocuciones	600
Total de frases con texto diferente	200
Frases interrogativas	258
Duración promedio de las frases	3.55 seg.
Duración total	2442 seg.
Total de palabras	5678
Total de palabras diferentes	202
Perplejidad de la gramática	5.9
Hablantes femeninos	6
Hablantes masculinos	6
Edad de los hablantes	15 a 55 años
Origen de los hablantes	Área Central de España.
Registro de las frases	Estudio de grabación.
Relación señal ruido promedio	48 dB.
Formato de los archivos	UNIX binario.
Frecuencia de muestreo	16 KHz (reducida a 8 KHz).
Resolución digital	16 bits/muestra.

hallan consideradas las características relevantes de cada locutor, rasgos lingüísticos de la frase y entorno en el que se pronunció. Los fonemas presentan la distribución que se muestra en la Figura B.1.

Cada locutor se identifica por las dos primeras letras del nombre de archivo: "aa", "ac", "al", "an", "aq", "ar", "ma", "mg", "mj", "mk", "mm" y "mo". Las que comienzan con "a" corresponden a elocuciones de mujeres y las que comienzan por "m" a los hablantes masculinos. Los últimos tres números del archivo identifican la frase pronunciada y se detallan en la Tabla B.2.

Las 600 elocuciones fueron distribuidas al azar en 10 particiones entrenamiento/prueba (80 %/20 %) para realizar la validación cruzada. Luego de ser entrenado con habla limpia, el sistema fue probado con habla contaminada con diferentes tipos de ruido a distintas relaciones señal ruido (SNR), para determinar la robustez del mismo. Se utilizó ruido blanco y tipo conversación de la base NOISEX-92 [209]. Los datos de ruido blanco fueron digitalizados de un generador de ruido analógico de alta calidad (Wandel & Goltermann) a 19.98 KHz y 16 bits, con igual energía en todo el ancho de banda. El ruido de conversación se grabó mediante un dispositivo DAT (*digital audio tape*) equipado con un micrófono tipo condensador. La fuente del "murmullo" fueron 100 personas hablando en una cantina. El radio del cuarto fue de unos dos metros; por lo que, la voces individuales son ligeramente audibles. El nivel de sonido durante el proceso de grabación fue de 88 dB SPL.

El ruido se remuestreó a 8 KHz y se mezcló con las señales de habla de SC1 en distintas proporciones para lograr distintas SNR. Se debe aclarar que en todos los casos se ha procedido por simple adición de las señales, no utilizandose el corpus Lombard.

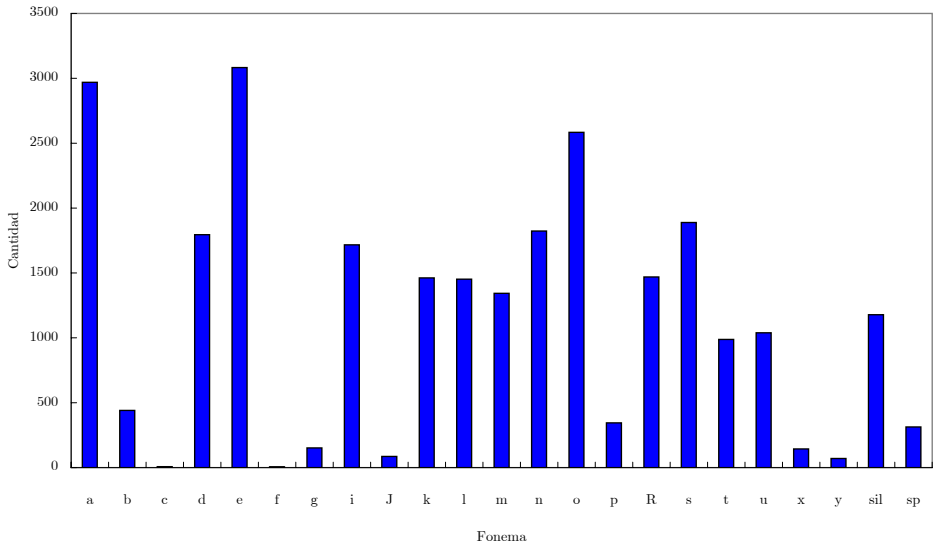

Figura B.1: *Distribución de los fonemas del subconjunto 1 del corpus geográfico de Albayzin (etiquetas de acuerdo al alfabeto fonético Worldbet).*

B.4. Modelos ocultos de Markov

Se construyeron en realidad 10 modelos, entrenados con cada una de los conjuntos de entrenamiento y probados con los correspondientes conjuntos de prueba.

Se puede decir que existen dos tipos de modelos que se entrenan independientemente, aunque finalmente se unen para formar un único gran modelo. En primer lugar están los denominados modelos acústicos, en los que se modela cada fonema (independiente del contexto) mediante un HMM de 3 estados [82, 216]. Por otro lado también deben estimarse las probabilidades para el modelo de lenguaje basado en bigramáticas [91].

La primera etapa consistió en el armado de los modelos iniciales de fonemas con el material de la base de Albayzin ya descripto. Se utilizaron modelos continuos de 3 estados con 1 fdp tipo gaussiana para las observaciones por cada uno de los estados. Se generaron los modelos iniciales de fonemas contexto independientes, basados en un único prototipo. Estos modelos se inicializaron mediante lo que se denomina un "modelo plano", en el cual se calculan las medias y varianzas sobre todo el conjunto sin importar el alineamiento. De esta manera todos los estados resultan equivalentes. Se construyó el modelo completo para todas las frases y se realizaron dos reestimaciones mediante el algoritmo de Baum-Welch. Posteriormente se agregaron los modelos de pausas cortas y se volvieron a reestimar los parámetros 4 veces más.

A continuación se aumentó la cantidad de gaussianas a 15 y realizó un enlace de parámetros (utilizando una población de 200 gaussianas para cada estado del modelo). Ésto disminuye la cantidad total efectiva de los mismos (de 855.000 a 26.200), mejorando la robustez de la estimación, lo que convierte al modelo en semi-continuo. Finalmente se procedió a realizar las reestimaciones restantes hasta completar un total de 16.

Para el modelo de lenguaje se utilizó una bigramática y la estimación de probabilidades y suavizado mediante el método de *backing-off*.

Tabla B.2: *Frases que integran el subconjunto 1 del corpus geográfico y su código correspondiente.*

001	¿A qué mar va a parar el río español de mayor longitud?
002	¿Cómo se llama el mar que baña Valencia?
003	¿Cuál es el caudal de todos los ríos de la Comunidad Valenciana?
004	¿Cuál es el caudal del Ebro?
005	¿Cuál es el caudal del río más largo que pasa por Andalucía?
006	¿Cuál es el caudal máximo de los ríos españoles?
007	¿Cuál es el caudal y longitud del Tajo?
008	¿Cuál es el mar en el que desembocan mayor número de ríos con una longitud mayor de 200 kilómetros?
009	¿Cuál es el mar que rodea las Canarias?
010	¿Cuál es el nombre del río más largo de la Península?
011	¿Cuál es el río de mayor longitud que desemboca en el mar Cantábrico?
012	¿Cuál es el río más caudaloso que pasa por Extremadura?
013	¿Cuál es el río más largo que atraviesa por lo menos 2 comunidades?
014	¿Cuál es la comunidad autónoma de mayor extensión por la que pasa el río Ebro?
015	¿Cuál es la extensión de la comunidad autónoma en la que nace el río Ebro?
016	¿Cuál es la longitud de todos los ríos?
017	¿Cuáles son las comunidades autónomas con una extensión superior a 20.000 kilómetros cuadrados?
018	¿Cuáles son las comunidades autónomas por las que pasan más ríos?
019	¿Cuáles son las comunidades que atraviesa el Tajo?
020	¿Cuáles son las comunidades que lindan con el mar?
021	¿Cuáles son los ríos catalanes más largos que 100 kilómetros?
022	¿Cuáles son los ríos cuya longitud es superior a 100 kilómetros?
023	¿Cuáles son los ríos que desembocan en el Cantábrico?
024	¿Cuáles son los ríos que pasan por Extremadura y otras 2 comunidades autónomas?
025	¿Cuáles son los ríos que pasan por la comunidad de Valencia?
026	¿Cuántas comunidades están bañadas por 2 mares?
027	¿Cuánto mide el Tajo?
028	¿Cuántos metros cúbicos por segundo lleva el Turia?
029	¿Cuántos mares reciben agua de un río?
030	¿Cuántos ríos con caudal mayor de 800 metros cúbicos por segundo pasan por la Comunidad Valenciana?
031	¿Cuántos ríos de Castilla y León tienen más de 100 kilómetros?
032	¿Cuántos ríos pasan por Aragón y Cataluña?
033	¿Cuántos ríos son más largos de 200 kilómetros?
034	¿Dónde desemboca el Guadiana?
035	¿Dónde nace el río Duero?
036	¿Dónde nace el río Ebro?
037	¿En qué comunidad autónoma está el río más caudaloso?
038	¿En qué comunidad autónoma hay más ríos?
039	¿En qué comunidad autónoma pasan nacen y desembocan más ríos?
040	¿En qué comunidad desemboca el río Ebro?
041	¿En qué comunidad nace y pasa el Pisuerga?
042	¿En qué comunidad nacen más ríos?
043	¿En qué mar desemboca el río más caudaloso de la comunidad andaluza?
044	¿En qué mar desembocan mayor número de ríos?
045	¿Es el Ebro más caudaloso que el Tajo?
046	¿Hay algún río cuyo caudal sea mayor que 100 metros cúbicos por segundo?
047	¿Me podría decir cuál es la comunidad donde está el nacimiento del Guadiana?
048	¿Pasa algún río por más de 4 comunidades?
049	¿Pasa el río Duero por la Comunidad de Madrid?
050	¿Por cuántas comunidades pasa el Ebro?

051	¿Por dónde pasa el río Duero?
052	¿Por dónde pasa el río con más caudal?
053	¿Por qué comunidad pasan más ríos?
054	¿Por qué mar está bañada Asturias?
055	¿Qué caudal tiene el Ebro?
056	¿Qué caudal tiene el Miño?
057	¿Qué comunidad autónoma es menos extensa?
058	¿Qué comunidad bañada por el Mediterráneo es la más extensa?
059	¿Qué comunidades no son bañadas por algún mar?
060	¿Qué comunidades son bañadas por el Tajo?
061	¿Qué comunidades tienen una extensión mayor de 1.000 kilómetros cuadrados?
062	¿Qué extensión tiene el País Vasco?
063	¿Qué longitud tiene el río más largo?
064	¿Qué mar baña Asturias?
065	¿Qué mar baña las costas de la Comunidad de Madrid?
066	¿Qué mar baña las costas del País Vasco?
067	¿Qué mar está junto a la Comunidad Valenciana?
068	¿Qué río cruza menos comunidades?
069	¿Qué río desemboca en el mar Mediterráneo y pasa por Murcia?
070	¿Qué río es más largo el Tajo o el Ebro?
071	¿Qué río tiene más caudal el Tajo o el Ebro?
072	¿Qué ríos desembocan en el mar Menor?
073	¿Qué ríos extremeños tienen una longitud superior a los 200 kilómetros?
074	¿Qué ríos hay en Asturias?
075	¿Qué ríos nacen en Cantabria?
076	¿Qué ríos pasan por Asturias y no nacen allí?
077	¿Qué ríos poseen un caudal superior a 800 metros cúbicos por segundo?
078	¿Qué ríos tienen más caudal que el río Duero?
079	¿Qué ríos tienen una longitud comprendida entre 500 y 1.000 kilómetros?
080	¿Seguro que el Segura pasa por la Comunidad de Valencia?
081	¿Tiene alguna comunidad más extensión que la comunidad andaluza?
082	¿Tienen la misma longitud y el mismo caudal el río Guadiana y el río Guadalquivir?
083	Caudal de los ríos con más de 100 kilómetros de longitud.
084	Caudal de los ríos que pasan por Castilla y León.
085	Caudal del río que pasa por la comunidad de Valencia.
086	Comunidad autónoma más grande.
087	Comunidades autónomas más grandes que Cataluña.
088	Comunidades con más de 5 ríos.
089	Comunidades por las que pasa el río Ebro.
090	Comunidades que baña el mar Mediterráneo.
091	Dígame el nombre del río más largo.
092	De los ríos del estado ¿cuántos desembocan en el Mediterráneo?
093	Deseo saber el caudal del río Miño.
094	Di el caudal del río menos caudaloso.
095	Di el río más caudaloso que desemboca en el Cantábrico.
096	Dime comunidades cuya superficie sea mayor a 1.000 kilómetros cuadrados.
097	Dime cuál es la comunidad autónoma de menor extensión.
098	Dime cuáles son las comunidades autónomas.
099	Dime cuántos ríos de la Comunidad Valenciana tienen más de 200 kilómetros de longitud.
100	Dime dónde desemboca el río Júcar.

101	Dime dónde muere el río Ebro.
102	Dime dónde nace el río Júcar.
103	Dime el caudal de los ríos de Cataluña.
104	Dime el caudal de todos los ríos que desembocan en el mar Mediterráneo.
105	Dime el caudal del río Cuervo.
106	Dime el caudal del río más pequeño que pasa por La Rioja.
107	Dime el caudal máximo de los ríos.
108	Dime el mar donde desemboca el río Turia.
109	Dime el mar en que desemboca el Miño.
110	Dime el número de ríos que desembocan en el Mediterráneo y que sean entre 1.000 y 200 kilómetros de largo.
111	Dime el nombre de las 3 comunidades autónomas más grandes.
112	Dime el nombre de las comunidades que linden con 2 mares.
113	Dime el nombre de los mares que bañan la comunidad de Andalucía.
114	Dime el nombre de los ríos que desembocan en el océano Atlántico.
115	Dime el nombre de los ríos que pasan por la Comunidad de Madrid.
116	Dime el nombre de los ríos que tienen menos de 100 kilómetros.
117	Dime el nombre de todas las comunidades que tienen mar.
118	Dime el río de mayor caudal que pase por la comunidad de Valencia.
119	Dime el río de menor longitud de Cataluña.
120	Dime en qué comunidad autónoma nace el Tajo.
121	Dime en qué comunidad nace el río Turia.
122	Dime la comunidad en la que desemboca el río Turia.
123	Dime la extensión de la comunidad asturiana.
124	Dime la extensión de las comunidades por donde pasa el Ebro.
125	Dime la longitud de los ríos que pasan por la Comunidad de Madrid.
126	Dime la longitud del río Guadalquivir.
127	Dime la longitud del río más largo.
128	Dime las comunidades autónomas con extensión superior a 1.000 kilómetros cuadrados.
129	Dime las comunidades autónomas.
130	Dime las comunidades que lindan con más de un mar.
131	Dime lo grande que es el Ebro.
132	Dime los mares que bañan Andalucía.
133	Dime los mares.
134	Dime los ríos con una longitud superior a 500 kilómetros.
135	Dime los ríos de la comunidad autónoma gallega.
136	Dime los ríos que desembocan en Andalucía.
137	Dime los ríos que desembocan en el Atlántico.
138	Dime los ríos que nacen en la Comunidad Foral de Navarra.
139	Dime los ríos que nacen y desembocan en la misma comunidad.
140	Dime los ríos que pasan por la Comunidad de Madrid.
141	Dime los ríos que tengan una longitud mayor que 500 kilómetros.
142	Dime qué longitud tiene el río Júcar.
143	Dime qué río tiene el caudal más grande.
144	Dime si por la comunidad de Valencia pasa o no más de un río.
145	Dime todos los mares que bañan Andalucía.
146	Dime todos los ríos que desembocan en el mar Cantábrico.
147	El río Ebro ¿pasa por la comunidad autónoma de Navarra?
148	El río Miño ¿por cuántas comunidades autónomas pasa?
149	Entre el río Ebro y el Júcar ¿cuál de ellos es más corto?
150	Enumera las comunidades autónomas por donde pasa el Ebro.

151	Enumera los ríos que tienen una longitud mayor de 100 kilómetros.
152	Enumerar los ríos que atraviesan la comunidad autónoma de Asturias.
153	Extensión de la comunidad autónoma por la cual pasa el río cuyo nombre es Guadalquivir.
154	Extensión del País Vasco.
155	La extensión de las comunidades autónomas que dan al mar Atlántico.
156	Lista de las comunidades por las que pase algún río de longitud mayor de 1.000 kilómetros.
157	Listado de todos los ríos con una longitud menor que la del Júcar.
158	Longitud de los ríos que desembocan en el mar Cantábrico.
159	Longitud del río Ebro.
160	Longitud del río que pasa por la Comunidad Valenciana.
161	Lugar donde desemboca el Júcar.
162	Mar en el que desembocan más ríos.
163	Mares en los que desembocan 5 o más ríos de longitud superior a 100 kilómetros.
164	Mares que bañan la comunidad gallega.
165	Nómbrame los ríos que pasan exactamente por 3 comunidades autónomas.
166	Número de mares del Estado Español.
167	Número de ríos que nacen y desembocan en la Comunidad Valenciana.
168	Nombra los ríos que pasan por las comunidades autónomas que no dan al mar.
169	Nombre de la comunidad autónoma en la que desemboquen mayor número de ríos.
170	Nombre de las 3 comunidades de menor extensión.
171	Nombre de las comunidades con extensión mayor que la Comunidad Valenciana.
172	Nombre de los mares que están en la Comunidad Valenciana.
173	Nombre de los ríos cuya longitud no supere los 1.000 kilómetros y no sea menor de 100 kilómetros.
174	Nombre de los ríos cuyo caudal es superior a 800 metros cúbicos por segundo.
175	Nombre de los ríos que desembocan en cada mar.
176	Nombre de los ríos que nacen en La Rioja y pasan por aquellas comunidades por las que sólo pasa ese río.
177	Nombre de los ríos que pasen por Castilla y León desembocan en el Atlántico y su caudal sea menor que el del río Tajo.
178	Nombre de todos los mares que bañan Andalucía.
179	Nombre del mar en el que desemboca un río que nace en Aragón.
180	Nombres de comunidades autónomas cuya extensión se encuentra entre 1.000 y 2.000 kilómetros cuadrados.
181	Obtener las comunidades autónomas por donde pasa el Ebro.
182	Quiero saber los nombres de los ríos más largos de 200 kilómetros.
183	Quisiera conocer cuántos ríos tienen un caudal de más de 200 metros cúbicos por segundo y son de menos de 1.000 kilómetros de largo.
184	Quisiera saber en qué mar desemboca el Segura.
185	Quisiera saber qué comunidades autónomas no tienen salida al mar.
186	Río más corto que desemboca en el Cantábrico.
187	Río más largo que nazca en Extremadura.
188	Ríos con caudal superior al del río Guadalquivir.
189	Ríos cuya longitud sea mayor de 1.000 kilómetros.
190	Ríos de Cantabria de más de 100 kilómetros de longitud.
191	Ríos de la comunidad autónoma gallega.
192	Ríos que atraviesen más de 3 autonomías.
193	Ríos que desembocan en el Cantábrico con una longitud mayor a 100 kilómetros.
194	Ríos que desemboquen en el Cantábrico.
195	Ríos que mueren en el Cantábrico.
196	Ríos que nacen en la Comunidad de Madrid.
197	Ríos que nacen en una comunidad bañada por el mar y desembocan en otra comunidad.
198	Ríos que pasan por la comunidad autónoma de Valencia.
199	Ríos que tengan un caudal superior a 800 metros cúbicos por segundo.
200	Todos los ríos.

El diccionario de pronunciaciones consta de 200 palabras. Para el reconocimiento a nivel fonemas se utilizó un diccionario en el que cada "palabra" es sólo un fonema y el lenguaje se modela mediante gramática plana (todas las transiciones con igual probabilidad).

B.5. Representación utilizada

Cada frase se normalizó en media, se preenfatizó y se separó en segmentos de 25 mseg de duración, multiplicados con ventanas de Hamming, cada 10 mseg. Cada segmento se parametrizó mediante un total de 28 coeficientes a partir de 13 coeficientes cepstrales, 1 coeficiente de energía y sus correspondientes derivadas temporales (MFCC+E+D), como se describió anteriormente [53].

B.6. Detalles de implementación

El reconocedor basado en HMM se implementó mediante "scripts" en lenguaje C-Shell y Perl, para el sistema HTK v3.0 [217]. Las rutinas para adicionar ruido a los datos se implementaron en GNU C++ estándar [98].

B.7. Resultados de referencia

La eficacia de reconocimiento se mide principalmente en términos de la comparación entre una secuencia de palabras deseada y la propuesta por el sistema. De esta forma el porcentaje de reconocimiento por palabras (RP %), la precisión en reconocimiento de palabras (PP %) y el porcentaje de frases reconocidas completas (RF %) se calculan mediante:

$$RP\,\% = \frac{Total - Borradas - Sustituidas}{Total} \times 100\,\%$$

$$PP\,\% = \frac{Total - Borradas - Sustituidas - Insertadas}{Total} \times 100\,\%$$

$$RF\,\% = \frac{Total - Sustituidas}{Total} \times 100\,\%$$

Como ya se mencionó, para obtener los resultados del sistema de referencia se realizaron pruebas de reconocimiento por el método de validación cruzada. Por ello los resultados finales corresponden a los valores medios sobre todas las particiones. En cada caso el conjunto de prueba tenía un 20 % de las frases que no habían sido utilizadas durante el entrenamiento. Todas las pruebas se repitieron para las frases con diferentes tipos y niveles de ruido.

Tabla B.3: *Resultados de reconocimiento del sistema completo con habla limpia para cada partición.*

Partición	RF (%)	RP (%)	PP (%)
1	54.62	91.62	90.95
2	55.46	92.11	90.79
3	57.14	91.94	90.55
4	67.80	94.01	93.73
5	63.03	94.16	93.38
6	58.47	91.71	90.21
7	62.18	91.73	91.20
8	63.87	92.02	90.76
9	66.39	94.04	93.05
10	68.07	91.30	90.08
Promedio	61.70	92.46	91.47

Tabla B.4: *Resultados de reconocimiento promedio del sistema completo con habla contaminada con ruido de conversación para cada partición.*

SNR (dB)	RF (%)	RP (%)	PP (%)
∞	61.70	92.46	91.47
50	66.74	93.43	93.06
25	49.64	92.07	89.57
15	3.11	82.16	70.23
10	1.17	69.21	55.70
5	0.25	47.53	26.99
0	0.00	29.16	6.21

B.7.1. Reconocimiento con el sistema completo

Los resultados de los experimentos de reconocimiento de palabras para el sistema completo, para habla limpia y con ruido se muestran en las Tablas B.3, B.4 y B.5.

B.7.2. Reconocimiento sin modelo de lenguaje

Para realizar estas pruebas se utilizó un modelo le lenguaje que asigna la misma probabilidad a todas las transiciones entre palabras. De esta forma, la información que posee el sistema es la correspondiente al diccionario de palabras y a los modelos acústicos de cada fonema. Los resultados de los experimentos de reconocimiento de palabras sin modelo de lenguaje, para habla limpia y con ruido se muestran en las Tablas B.6, B.7 y B.8.

Tabla B.5: *Resultados de reconocimiento promedio del sistema completo con habla contaminada con ruido blanco para cada partición.*

SNR (dB)	RF (%)	RP (%)	PP (%)
∞	61.70	92.46	91.47
50	66.66	93.62	93.22
25	47.14	89.96	87.52
15	4.20	77.71	67.10
10	1.26	55.93	46.92
5	0.16	28.86	17.96
0	0.08	15.01	6.67

Tabla B.6: *Resultados de reconocimiento de palabras sin modelo de lenguaje con habla limpia para cada partición.*

Partición	RF (%)	RP (%)	PP (%)
1	25.00	80.79	71.18
2	0.84	86.94	77.96
3	0.84	85.06	75.45
4	2.54	85.13	76.18
5	1.68	85.46	75.81
6	0.85	85.35	75.85
7	2.52	87.03	78.92
8	0.84	85.02	75.43
9	0.00	85.80	76.98
10	3.36	85.68	76.69
Promedio	3.85	85.23	76.05

Tabla B.7: *Resultados de reconocimiento de palabras promedio sin modelo de lenguaje con habla contaminada con ruido de conversación.*

SNR (dB)	RF (%)	RP (%)	PP (%)
∞	3.85	85.23	76.05
50	4.18	85.24	75.96
25	2.66	82.76	69.06
15	0.00	76.06	48.99
10	0.00	69.54	36.33
5	0.00	63.54	11.70
0	0.00	57.55	-6.13

Tabla B.8: *Resultados de reconocimiento de palabras promedio sin modelo de lenguaje con habla contaminada con ruido blanco.*

SNR (dB)	RF (%)	RP (%)	PP (%)
∞	3.85	85.23	76.05
50	1.34	85.20	76.11
25	3.84	81.49	69.27
15	0.00	72.18	49.21
10	0.00	63.14	41.08
5	0.00	54.20	16.51
0	0.00	47.31	6.53

Tabla B.9: *Resultados de reconocimiento de fonemas con habla limpia para cada partición.*

Partición	RF (%)	RP (%)	PP (%)
1	-	75.39	17.51
2	-	76.29	17.09
3	-	74.84	15.86
4	-	75.03	15.86
5	-	75.41	15.77
6	-	75.49	16.74
7	-	75.89	17.51
8	-	75.87	16.92
9	-	75.97	16.62
10	-	75.33	16.03
Promedio	-	75.55	16.59

B.7.3. Reconocimiento de fonemas

Estos resultados se obtuvieron con un sistema que no posee modelo de lenguaje y tampoco posee un diccionario de pronunciaciones. En este caso la única información es la del modelo acústico de los fonemas. Los resultados se muestran en las Tablas B.9, B.10 y B.11.

Tabla B.10: *Resultados de reconocimiento de fonemas promedio con habla contaminada con ruido de conversación.*

SNR (dB)	RF (%)	RP (%)	PP (%)
∞	-	75.55	16.59
50	-	75.52	16.60
25	-	72.78	-0.93
15	-	65.91	-43.22
10	-	62.26	-63.70
5	-	57.98	-77.84
0	-	54.10	-82.61

Tabla B.11: *Resultados de reconocimiento de fonemas promedio con habla contaminada con ruido blanco.*

SNR (dB)	RF (%)	RP (%)	PP (%)
∞	-	75.55	16.59
50	-	75.35	16.32
25	-	69.23	4.78
15	-	60.44	-16.72
10	-	55.73	-33.51
5	-	50.10	-58.54
0	-	45.27	-49.26

Referencias

[1] Abdallah, S. A.: *Towards music perception by redundancy reduction and unsupervised learning in probabilistic models.* PhD thesis, Department of Electronic Engineering, King's College London, 2002.

[2] Abdallah, S. A. and M. D. Plumbley: *Sparse coding of music signals.* `http://citeseer.ist.psu.edu/abdallah01sparse.html`, Preprint, 2001.

[3] Allen, J.B. and L.R. Rabiner: *A unified approach to short-time fourier analysis and synthesis.* Proc. IEEE, 65(11):1558–1564, 1977.

[4] Añino, M. M., M. E. Torres, and G. Schlottahauer: *Slight parameter changes detection in biological models: A multiresolution approach.* Physica A, 324(3–4):645–664, 2003.

[5] Aronson, L., H.L. Rufiner, H. Furmansky y P. Estienne: *Características Acústicas de las Vocales del Español Rioplatense.* Fonoaudiológica, 46(2):12–20, Julio–Septiembre 2000.

[6] Bachman, G. and L. Narici: *Functional Analysis.* Dover Publications, New York, 2000.

[7] Banbrook, M., S. McLaughlin, and I. Mann: *Speech characterization and synthesis by nonlinear methods.* IEEE Trans. on Speech and Audio Processing, 7(1):1–17, 1999.

[8] Barlow, H.: *Redundancy reduction revisited.* Network: Computation in Neural Systems, 12(3):241–253, 2001.

[9] Békésy, G. Von: *Experiments in Hearing.* McGraw-Hill, New York, 1960.

[10] Belin, P., R.J. Zatorre, R. Hoge, A.C. Evans, and B. Pike: *Event-related fMRI of the auditory cortex.* NeuroImage, 10:417–429, 1999.

[11] Bell, A. J. and T. J. Sejnowski: *An information maximization approach to blind separation and blind deconvolution.* Neural Computation, 7(6):1129–1159, 1995.

[12] Bell, A.J. and T.J. Sejnowski: *An information maximization approach to blind separation and blind deconvolution.* Neural Computation, 7:1129–1159, 1995.

[13] Boulard, H., H. Hermansky, and N. Morgan: *Towards increasing speech recognition error rates.* Speech Communication, 18(3):205–231, 1996.

[14] Cardoso, J.F.: *Supersymmetric decomposition of the fourth order cumulant tensor: blind identification of more sources than sensors.* In *Proceedings of ICASSP'91*, pages 3109–3112, 1991.

[15] Carmell, T., A. Cronk, E. Kaiser, R. Wesson, J. Wouters, and X. Wu: *Spectrogram reading.* Web page, CSLU-OGI, March 1997. `http://cslu.cse.ogi.edu/tutordemos/SpectrogramReading`.

[16] Carney, L. H. and C. D. Geisler: *A temporal analysis of auditory-nerve fiber responses to spoken stop consonant-vowel syllables.* Journal of the Acoustic Society of America, 79(6):1896–1914, June 1986.

[17] Casacuberta, F., R. García, J. Llisterri, C. Nadeu, J. M. Prado, and A. Rubio: *Development of a spanish corpora for the speech research.* In *Workshop on International Co-operation and Standardisation of Speech Databases and Speech I/O Assessment Methods*, pages 26–28, Chiavari, September 1991.

[18] Castañeda, N., J.M. Cornejo, and P. Granados: *Neuroanatomical representation of P1 component of the Long Latency Auditory Evoked Potential as a frequency function of a tone burst in normal hearing young children.* In *XVIII IERASG Biennial Symposium*, Puerto de la Cruz, Tenerife, Canary Islands, Spain, June 2003.

[19] Chechik, Gal: *An Information Theoretic Approach to the Study of Auditory Coding.* Ph.D. thesis, Hebrew University, July 2003.

[20] Chen, S.S.: *Basis Pursuit.* PhD thesis, Department of Statistics, Stanford University, 1995.

[21] Chen, S.S., D.L. Donoho, and M.A. Sanders: *Atomic decomposition by basis pursuit.* SIAM Journal on Scientific Computing, 20(1):33–61, 1999.

[22] Chennubhotla, Chakra and Allan Jepson: *Sparse coding in practice.* In Aldroubi, A., A.F. Laine, and M.A. Unser (editors): *Proc. 2nd Intĺ (IEEE) Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing, and Sampling (SCTV)*, Vancouver, Canada, July 2001.

[23] Cichocki, A. and A. K. Barros: *Robust batch algorithm for sequential blind extraction of noisy biomedical signals.* In *5th International Symposium on Signal Processing and its Applications (ISSPA'99)*, volume 1, pages 363–366, Brisbane, Queensland, Australia, August 1999.

[24] Cingolani, Horacio E. y Alberto B. Houssay: *Fisiología Humana*, volumen 4. El Ateneo, Buenos Aires, 6 edición, 1988.

[25] Coifman, R.R. et Y. Meyer: *Remarques sur l' analyse de Fourier a fenêtre.* C.R. Acad. Sci., pages 25–261, 1991.

[26] Coifman, R. and M.V. Wickerhauser: *Entropy-based algorithms for best basis selection.* IEEE Trans. Inform. Theory, 38(2):713–718, March 1992.

[27] Comon, P.: *Independent component analysis - a new concept?* Signal Processing, 36:287–314, 1994.

[28] Conant, R. C.: *Detecting subsystems of a complex system.* IEEE Transactions on Systems, Man and Cybernetics, 2:550–553, Sept 1972.

[29] Cooley, J. W. and J. W. Tukey: *An algorithm for machine calculation of complex Fourier series.* Mathematics of Computation, 19(90):297–301, April 1965.

[30] Cover, T. M. and J. A. Thomas: *Information Theory.* John Wiley and Sons, NY, 1991.

[31] Crouse, M., R. Nowak, and R. Baraniuk: *Wavelet-based statistical signal processing using hidden Markov models.* IEEE Transactions on Signal Processing, 46(4):886–902, 1998.

[32] Daróvczy, Z.: *Generalized information functions.* Inf. Control, 16:36–51, 1970.

[33] Daubechies, I.: *Time-frequency localization operators: a geomeric phase space aproach.* IEEE Trans. Info. Thry, 34(4):605–612, 1988.

[34] Daubechies, I.: *Ten Lectures on wavelets.* Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1992.

[35] deCharms, R.C., D.T. Blake, and M.M. Merzenich: *Optimizing sound features for cortical neurons.* Science, 280:1439–1443, 1998.

[36] Delgutte, B.: *Physiological models for basic auditory percepts.* In Hawkins, H.H., T.A. McMullen, A.N Popper, and R.R. Fay (editors): *Auditory Computation.* Springer, New York, 1996.

[37] Deller, J., J. Proakis, and J. Hansen: *Discrete Time Processing of Speech Signals*. Macmillan Publishing, New York, 1993.

[38] Diaz, J. E., A. M. Peinado, A. J. Rubio, E. Segarra, N. Prieto, and F. Casacuberta: *Albayzin: A task-oriented spanish speech corpus*. In *Proceedings of the 1st International Conference in Language Resources and Evaluation*, volume 1, pages 497–501, Granada, May 1998.

[39] Diaz, J. E., A. J. Rubio, A. M. Peinado, E. Segarra, N. Prieto, and F. Casacuberta: *Development of a task-oriented spanish speech corpora*. In *Proceedings of the 2th European Conference of Speech Communication and Technology*, Berlin, September 1993.

[40] Donoho, D.L.: *Sparse components of images and optimal atomic decomposition*. Technical report, Department of Statistics, Stanford University, December 1998.

[41] Donoho, D.L.: *Beyond wavelets*. Lectures given at University of Missouri, St. Louis, May 22–26 2000. National Science Foundation CBMS Lecture Series program.

[42] Donoho, D.L. and A. G. Flesia: *Can recent innovations in harmonic analysis "explain" key findings in natural image statistics?* Network: Comput. Neural Syst., 12(3):371–393, August 2001.

[43] Ducrot, Oswald y Tzvetan Todorov: *Diccionario enciclopédico de las ciencias del lenguaje*. Siglo Veintiuno, Mexico, 10 edición, 1984.

[44] Elad, Michael: *Sparse representations of signals - theory and applications*. In *Proc. of the IPAM MGA Workshop*, September 2004. Invited Talk.

[45] Elman, J.L.: *Finding structure in time*. Cognitive Science, 14:179–211, 1990.

[46] Farooq, O. and S. Datta: *Mel filter-like admissible wavelet packet structure for speech recognition*. IEEE Signal Processing Letters, 8(7):196–198, July 2001.

[47] Favero, Richard F.: *Comparison of perceptual scaling of wavelet for speech recognition*. In *Proceedings of the SST-94*, 1994.

[48] Fisher, W.M., G.R. Doddington, and K.M. Goudie-Marshall: *The DARPA Speech Recognition Research Database: Specifications and Status*. In *Proceedings of the DARPA Speech Recognition Workshop*, Palo Alto, February 1986. Report Nº SAIC-86/1546.

[49] Flandrin, P.: *Some aspects of non stationary signals processing with emphasis on time-frequency and time-scale methods*. In *Proc. of International Conference on Wavelets, Time-Frequency Methods and Phase Space*, pages 68–98, Marselle, France, 1989.

[50] Földiák, Peter: *Sparse coding in the primate cortex*. In *The Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, 2nd edition, 2002.

[51] Fletcher, H.: *Speech and Hearing in Communication*. Van Nostrand, New York, 1953.

[52] Fourier, Jean Baptiste Joseph: *Théorie analytique de la chaleur*. F. Didot, Paris, 1822.

[53] Furui, S. F.: *Cepstral analysis technique for automatic speaker verification*. IEEE Trans. ASSP, 29:254–272, April 1981.

[54] Gabor, D.: *Theory of communication*. J. IEE, 93:429–457, 1946.

[55] Gamero, L. y H.L. Rufiner: *Paquetes de onditas evolutivas para clasificación de señales*. En *Anales del Ier Congreso Latinoamericano de Ingeniería Biomédica*, volumen 1, páginas 784–787, Noviembre 1998.

[56] Garofolo, Lamel, Fisher, Fiscus, Pallett, and Dahlgren: *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus Documentation*. Technical report, National Institute of Standards and Technology, February 1993.

[57] Girosi, F.: *An equivalence between sparse approximation and support vector machines*. Neural Computation, 10(6):1455–1480, August 1998.

[58] Goddard, J., A.E. Martínez, F. M. Martínez, and H. L. Rufiner: *A comparison of string kernels and discrete hidden markov models on a spanish digit recognition task.* In *Proc. of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2962–2965, Cancun, México, September 2003.

[59] Goddard, J., F. M. Martínez, A.E. Martínez, and H. L. Rufiner: *Noisy speech recognition using string kernels.* In *Proceedings of the 9-th International Conference SPEECH AND COMPUTER (SPECOM'2004)*, St. Petersburg, Russia, September 2004.

[60] Goldberg, D. E.: *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, 1997.

[61] Goodwin, M.M. and M. Vetterli: *Matching pursuit and atomic signal models based on recursive filter banks.* IEEE Trans. Signal Proc., 47(7):1890–1902, 1999.

[62] Graham, N.: *The visual system does a crude Fourier analysis of patterns.* In *Proc. 13 Amer. Math. Soc. (SIAM-AMS)*, page 1–16, Philadelphia, 1980. Mathematical psychology and psychophysiology.

[63] Greenberg, S.: *The ears have it: The auditory basis of speech perception.* In *Proceedings of the International Congress of Phonetic Sciences*, volume 3, pages 34–41, 1995.

[64] Greenberg, S.: *Recognition in a new key - towards a science of spoken language.* In *ICASSP98, Internantional Conference on Acoustics, Speech and Signal Processing*, pages 1041–1045, Seattle, 1998.

[65] Greenberg, S.: *What are the essential cues for understanding spoken language.* In *Proc. 141st Meeting of the Acoustical Society of America*, Chicago, IL, June 2001.

[66] Guerin-Dugue, A. and et al.: *Elena ii: Enhanced learning for evolutive neural architecture.* Technical Report Number 6891, Research Program Report, June 1995.

[67] Gurlekian, J. A., M. Guirao, and H. E. Franco: *Acoustic characteristics and perception of spanish stop consonants.* Journal of the Acoustical Society of Japan, S85-36:271–278, 1985.

[68] Gurlekian, J., L. Colantoni, H. Torres, A. Rincon, A. Moreno, and J. Mariño: *Database for an automatic speech recognition system for argentine spanish.* In Bird, Buneman & Liberman (editor): *Proceedings of the IRCS Workshop on Linguistic Databases, Workshop on Linguistic Databases*, pages 92–98, Filadelfia, USA, December 2001. `http://www.ldc.upenn.edu/annotation/database/papers/Gurlekian_etal/36.2.gurlekian.pdf`.

[69] Guspí, F. y B. Introcaso: *Soluciones Ralas de Sistemas Lineales Indeterminados.* El Ingeniero en la Red, 1(VII):1–10, Mayo 2000. Revista Electrónica FCEIyA, UNR, Argentina.

[70] Hadamard, Jacques: *Sur les problèmes aux dérivées partielles et leur signification physique.* Princeton University Bulletin, pages 49–52, 1902.

[71] Harpur, George Francis: *Low Entropy Coding with Unsupervised Neural Networks.* PhD thesis, Department of Engineering, University of Cambridge, Queens' College, February 1997.

[72] Havrda, J. and F. Charvat: *Quantification method of classification process: Concept of structural α-entropy.* Kybernetica, 3:30–35, 1967.

[73] Helmholtz: *Die Lehre von den Tenepfindungen als physiologische Grundlage fur die Theorie der Musik.* Dover Publications, New York, 1863. Republished in 1954.

[74] Hermansky, H.: *Perceptual linear predictive (PLP) analysis of speech.* Journal Acoust. Soc. Am., 87(4):1738–1752, 1990.

[75] Hermansky, H.: *Should recognizers have ears?* In *Proc. ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 1–10, France, 1997.

[76] Hermansky, H. and N. Morgan: *RASTA processing of speech.* IEEE Trans. on Speech and Audio Processing, 2(4):578–589, 1994.

[77] Hillenbrand, J.M., M.J. Clark, and T.M. Nearey: *Effects of consonant environment on vowel formant patterns.* Journal of the Acoustic Society of America, 109(2):748–763, February 2001.

[78] Holland, J. H.: *Adaptation in Natural and Artificial System.* University of Michigan Press, Ann Arbor, 1975.

[79] Honda, Kiyoshi: *Study of speech communication mechanisms based on the understanding of human biological functions.* Journal of the Communications Research Laboratory, 48(3):45–48, 2001.

[80] Horn, R. A. and C. R. Johnson: *Matrix Analysis.* Cambridge University Press, 1985.

[81] Huang, L. S. and C. H. Yang: *A novel approach to robust speech endpoint detection in car environments.* In *Proceedings of the 2000 International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1751–1754, Istanbul, Turkey, June 2000. IEEE.

[82] Huang, X. D., Y. Ariki, and M. Jack: *Hidden Markov Models for Speech Recogniton.* Edinburgh University Press, 1992.

[83] Hyvärinen, Aapo, Patrik Hoyer, and Erkki Oja: *Sparse code shrinkage: Denoising by nonlinear maximum likelihood estimation.* In *Proc. Advances in Neural Information Processing Systems*, pages 473–479, 1999.

[84] Hyvärinen, A.: *Sparse code shrinkage: Denoising of nongaussian data by maximum-likelihood estimation.* Technical report, Helsinki University of Technology, 1998.

[85] Hyvärinen, Aapo, Juha Karhunen, and Erkki Oja: *Independent Component Analysis.* John Wiley & Sons, 2001.

[86] Hyvärinen, A.: *Survey on independent component analysis.* Neural Computing Surveys, 2:94–128, 1999.

[87] Hyvärinen, A.: *Complexity pursuit: Separating interesting components from time-series.* Neural Computation, 13(4):883–898, 2001.

[88] Hyvärinen, A., H.R. Cristescu, and E. Oja: *A fast algorithm for estimating overcomplete ICA bases for image windows.* In *Proc. Int. Joint Conf. on Neural Networks*, Washington, D.C., 1999.

[89] Hyvärinen, A., P.O. Hoyer, and M. Inki: *Topographic independent component analysis.* Neural Computation, 13(7):1525–1558, 2001.

[90] Hyvärinen, A. and E. Oja: *Independent component analysis: Algorithms and applications.* Neural Networks, 13(4-5):411–430, April 2000.

[91] Jelinek, F.: *Statistical Methods for Speech Recognition.* The MIT Press, Cambrige, Masachussets, 1999.

[92] Johnson, J.L., M.L. Padgett, and O. Omidvar: *Guest editorial: Overview of pulse coupled neural network (pcnn).* IEEE Transactions on Neural Networks, 10(3), May 1999. Special issue.

[93] Jung, T.-P., S. Makeig, T.-W. Lee, M.J. McKeown, G. Brown, A.J. Bell, and T.J. Sejnowski: *Independent component analysis of biomedical signals.* `http://www.cnl.salk.edu/~jung/ica.html`, 2001.

[94] Junqua, J. C.: *The lombard reflex and its role on human listeners and automatic speech recognizers.* Journal of the Acoustic Society of America, 1:637–642, 1993.

[95] Junqua, J.C.: *The lombard reflex and its role on human listeners and automatic speech recognizers.* In J.Acoust.Soc.Amer. (editor): *93*, volume 1, pages 637–642, 1993.

[96] Jutten, C. and J. Herault: *Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture*. Signal Processing, 24:1–10, 1991.

[97] Kailath, T. and H.V. Poor: *Detection of stochastic processes*. IEEE Transactions on Information Theory, 44(6):2230–2259, 1998.

[98] Kalev, Danny: *ANSI/ISO C++ Professional Programmer's Handbook*. Professional Series. Que Corporation, 1st edition, June 1999. ISBN: 0789720221.

[99] Kandel, E.R., J.H. Schwartz y T.M. Jessell: *Principios de Neurociencia*. McGraw-Hill, 2001.

[100] Khanna, S.M. and J. Tonndorf: *Tympanic membrane vibrations in cats studied by time-average holography*. Journal of the Acoustic Society of America, 51(6B):1904–1920, 1972.

[101] Kiang, Watenabe, and Thomas: *Discharge Patterns of Single Fibers in the Cats Auditory Nerve*. MIT Press, Cambridge, MA, 1965.

[102] Klein, D.J., D.A. Depireux, J.Z. Simon, and S.A. Shamma: *Robust spectrotemporal reverse correlation for the auditory system: Optimizing stimulus design*. Journal of Computational Neuroscience, 9:85–111, 1996.

[103] Kohonen, T.: *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).

[104] Kording, Konrad P., Peter Konig, and David J Klein: *Learning of sparse auditory receptive fields*. In *Proc. of the International Joint Conference on Neural Networks (IJCNN '02)*, volume 2, pages 1103–1108, Honolulu, HI, United States, May 2002.

[105] Kraus, N., A.R. Bradlow, M.A. Cheatham, J. Cunningham, C.D. King, D.B. Koch, T.G. Nicol, T.J. McGee, L.K. Stein, and B.A. Wright: *Consequences of neural asynchrony: A case of auditory neuropathy*. Journal of the Association for Research in Otolaryngology, May 2000.

[106] Kreutz-Delgado, K. and B.D. Rao: *Sparse basis selection, ICA, and majorization: Towards a unified perspective*. In *Proc. ICASSP*, volume 2, 1999. Paper no. 2411.

[107] Kwon, Oh-Wook and Te-Won Lee: *Phoneme recognition using ica-based feature extraction and transformation*. Signal Processing, 84(6):1005—1019, 2004.

[108] Lane, H. L. and B. Tranel: *The lombard sign and the role on hearing in speech*. Journal of Speech and Hearing Research, 14:677–709, 1971.

[109] Lappalainen, Harri: *A computationally efficient algorithm for finding sparse codes*. Master's thesis, Neural Networks Research Centre, Laboratory of Computer and Information Science, Helsinki University of Technology, 1996.

[110] Lee, J.H., H.Y. Jung, T.W Lee, and S.Y. Lee: *Speech feature extraction using independent component analysis*. In *Proc. ICASSP*, volume 3, pages 1631–1634, 2000.

[111] Lee, T.-W., M.S. Lewicki, M. Girolami, and T.J Sejnowski: *Blind source separation of more sources than mixtures using overcomplete representations*. IEEE Sig. Proc. Lett., 6(4):87–90, April 1999.

[112] Lee, T.-W., M.S. Lewicki, and T.J. Sejnowski: *Unsupervised classification with non-gaussian mixture models using ica*. Advances in Neural Information Processing Systems, 11, 1998.

[113] Lee, Te-Won and M.S. Lewicki: *Unsupervised image classification, segmentation, and enhancement using ICA mixture models*. IEEE Trans. on Image Proc., 11(3):270–279, March 2002.

[114] Lewicki, M.S.: *Efficient coding of natural sounds*. Nature Neuroscience, 5(4):356–363, 2002.

[115] Lewicki, M.S. and B.A. Olshausen: *A probabilistic framework for the adaptation and comparison of image codes*. Journal of the Optical Society of America, 16(7):1587–1601, 1999.

[116] Lewicki, M.S. and T.J. Sejnowski: *Learing overcomplete representations*. In *Advances in Neural Information Processing 10 (Proc. NIPS'97)*, pages 556–562. MIT Press, 1998.

[117] Lewicki, M.S. and T.J. Sejnowski: *Learning overcomplete representations.* Neural Computation, 12(2):337–365, 2000.

[118] Lewicki, M. and B.A. Olshausen: *Inferring sparse, overcomplete image codes using an efficient coding framework.* In *Advances in Neural Information Processing 10 (Proc. NIPS'97)*, pages 815–821. MIT Press, 1998.

[119] Lippmann, Richard P.: *Speech recognition by machines and humans.* Speech Communication, 19(22):1–15, 1997.

[120] Llorach, Emilio Alarcos: *Gramática de la Lengua Española.* Real Academia Española. Colección Nebrija y Bello. Editorial Espasa Calpe, Madrid, 1999.

[121] Lombard, E.: *Le signe de l'elevation de la voix.* Annales Maladies Oreilles, Larynx, Nez, Pharynx, 37 :101–119, 1911.

[122] Maass, W.: *Neural computation: a research topic for theoretical computer science? Some thoughts and pointers.* In Rozenberg, G., A. Salomaa, and G. Paun (editors): *Current Trends in Theoretical Computer Science, Entering the 21th Century.* World Scientific Publishing, 2001.

[123] Makeig, S., A.J. Bell, T-P. Jung, and T.J. Sejnowski: *Independent component analysis of electroencephalographic data.* Advances in Neural Information Processing Systems, 8:145–151, 1996.

[124] Makeig, S., T-P. Jung, A.J. Bell, and T.J. Sejnowski: *Blind separation of auditory event-related brain responses into independent components.* Proc. Nat. Acad. Sci. USA, 94:10979–10984, 1997.

[125] Mallat, S.G.: *Multiresolution approximation and wavelet orthonormal bases of l2(r).* Trans. Amer. Math. Soc., 315(1):69–88, 1989.

[126] Mallat, S.G: *A theory of multiresolution signal decomposition: the wavelet representation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(7), 1989.

[127] Mallat, S.G.: *A Wavelet Tour of Signal Processing.* Academic Press, 2nd edition, September 1999.

[128] Mallat, S.G. and Z. Zhang: *Matching pursuit with time-frequency dictionaries.* IEEE Trans. in Signal Proc., 41:3397–3415, December 1993.

[129] Manrique, Ana María Borzone: *Manual de Fonética Acústica.* Hachette, Buenos Aires, 1980.

[130] Meyer, Y.: *Ondelettes et Opérateurs, Tome I. Ondelettes.* Herrmann de., Paris, 1990.

[131] Michalewicz, Z.: *Genetic Algorithms + Data Structures = Evolution Programs.* Springer-Verlag, 1992.

[132] Michie, D., D. J. Spiegelhalter, and C. C. Taylor: *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, University College, London, 1994.

[133] Miller, M.I. and M.B. Sachs: *Representation of stop consonants in the discharge patterns of auditory-nerve fibers.* Journal of the Acoustic Society of America, 74(2):502–517, August 1983.

[134] Milone, D.H.: *Información acentual para el reconocimiento automático del habla.* Tesis de Doctorado, Facultad de Ciencias de la Universidad de Granada, Granada, España, 07 2001. Dir. Antonio J. Rubio Ayuso.

[135] Milone, D.H. y H.L. Rufiner (Eds.): *Introducción a las señales y los sistemas digitales.* Editorial UNER, octubre 2003. en prensa.

[136] Moreno, A., D. Poch, A. Bonafonte, E.Lleida, J.Llisterri, J.B.Marino, and C. Nadeu: *Albayzin speech data base: design of the phonetic corpus.* In *Proceedings of the 2th European Conference of Speech Communication and Technology*, pages 175–178, Berlin, September 1993.

[137] Morris, A.C. and J.M. Pardo: *Phoneme transition detection and broad classification using a simple model based on the function of onset detector cells found in the cochlear nucleus.* In *Proceedings of the Eurospeech'95*, pages 115–118, 1995.

[138] Newton, I.: *New theory about light and colors.* Philosophical Transactions of the Royal Society, 80(7):3075–3087, 1672.

[139] Nicolis, G. and I. Prigogine: *Exploring Complexity: An Introduction.* W.H. Freeman & Company, 1989.

[140] Oja, E.: *The nonlinear PCA learning rule in independent component analysis.* Neurocomputing, 17(1):25–46, 1997.

[141] Oja, E.: *Nonlinear PCA criterion and maximum likelihood in independent component analysis.* In *Proceedings of International Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 143–148, Aussois, France, 1999.

[142] Olshausen, B. A., P. Sallee, and M. S. Lewicki: *Learning sparse images codes using a wavelet pyramid architecture.* In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000.

[143] Olshausen, B.A.: *Sparse codes and spikes.* In Rao, R. P. N., B. A. Olshausen, and M. S. Lewicki (editors): *Probabilistic Models of the Brain: Perception and Neural Function*, chapter 13. MIT Press, 2001. In Press.

[144] Olshausen, B.A. and D.J. Field: *Natural image statistics and efficient coding.* In *Proc. of the Workshop on Information Theory and the Brain*, volume 7, pages 333–339, Scotland, September 4-5 1995. University of Stirling.

[145] Olshausen, B.A. and D.J. Field: *Emergence of simple cell receptive field properties by learning a sparse code for natural images.* Nature, 381:607–609, 1996.

[146] Olshausen, B.A. and D.J. Field: *Sparse coding with an overcomplete basis set: A strategy employed by v1?* Vision Research, 37(23):3311–3325, 1997.

[147] Olshausen, B.A. and D.J. Field: *Vision and the coding of natural images.* American Scientist, 88(3):238–245, 2000.

[148] Parks, T.W. and C.S. Burrus: *Digital Filter Design*, pages 226–228. New York: John Wiley & Sons Ltd., 1987.

[149] Parra, L.C. and C. Spence: *Convolutive blind source separation of non-stationary sources.* IEEE Trans. on Speech and Audio Processing, pages 320–327, May 2000.

[150] Pearlmutter, B.A. and L.C. Parra: *A context-sensitive generalization of ica.* In *International Conference on Neural Information Processing*, Hong Kong, September 1996. `citeseer.nj.nec.com/pearlmutter96contextsensitive.html`.

[151] Pearson, K.: *On lines and planes of closest fit to systems of points in space.* Philosophical Magazine, 2(6):559–572, 1901.

[152] Peterson and Barney: *Control methods used in a study of the vowels.* Journal of the Acoustic Society of America, 24(2):175–184, 1952.

[153] Poggio, Tomaso and Christian R. Shelton: *Machine learning, machine vision, and the brain.* Artificial Intelligence Magazine, Sept 1999.

[154] Portnoff, M.: *Time-frecuency representation of digital signals and systems based on short-time fourier analysis.* IEEE Trans. on Acoust., Speech and Signal Proc., 28:55–69, Feb. 1980.

[155] Purves, D., G. Augustine, D. Fitzpatrick, L. Katz, A. LaMantia y J. McNamara: *Invitación a la Neurociencia.* Editorial Médica Panamericana, 2001.

[156] Purwins, Hendrik, Benjamin Blankertz, and Klaus Obermayer: *Computing auditory perception.* In *Proceedings of the Cognition and Perception Issues in Computer Music*, 2000.

[157] Quian, Shie and Dapang Chen: *Joint Time-Frequency Analysis: Method and Applications.* Prentice Hall, 1996.

[158] Quilis, Antonio: *Tratado de Fonología y Fonética Españolas.* Biblioteca Románica Hispánica. Editorial Gredos, Madrid, 1993.

[159] Rabiner, L. R. and R. W. Schafer: *Digital Processing of Speech Signals.* Prentice Hall, NJ, 1978.

[160] Rabiner, L. and B. H. Juang: *Fundamentals of Speech Recognition.* Prentice-Hall, 1993.

[161] Rioul, O. and M. Vetterli: *Wavelets and signal processing.* IEEE Signal Processing Magazine, 8(4):14–38, Octubre 1991.

[162] Rissanen, J.: *Minimum description length principle.* In Kotz, S. and N. L. Johnson (editors): *Encyclopedia of Statistical Sciences*, volume 5, pages 523–527. New York: Wiley, 1985.

[163] Roberts, Stephen and Richard Everson (editors): *Independent Component Analysis: Principles and Practice.* Cambridge University Press, 2001.

[164] Robinson, A.J. and F. Fallside: *Static and dynamic error propagation networks with application to speech coding.* In Anderson, D.Z. (editor): *Neural Information Processing Systems*, pages 632–641, New York, NY, 1988. American Institute of Physics.

[165] Robles, L., M.A. Ruggero, and N. C. Rich: *Two-tone distortion in the basilar membrane in the cochlea.* Nature, 349:413–414, 1991.

[166] Rufiner, H.L: *Modelización Biológica, Redes Neuronales y HMM´s aplicados al Reconocimiento Automático del Habla.* Informe de Avance Beca de Investigación, CONICET, 1994.

[167] Rufiner, H.L.: *Comparación entre Análisis Onditas y Fourier aplicados al reconocimiento automático del habla.* Tesis de Maestría, Universidad Autónoma Metropolitana, December 1996.

[168] Rufiner, H.L. y D.Zapata: *Desarrollo de un Sistema de Reconocimiento Automático del Discurso Continuo, Independiente del Hablante y con Vocabulario Ampliable.* Tesis de grado, Facultad de Ingeniería - Bioingenieria, Mayo 1992.

[169] Rufiner, H.L. y J.C. Goddard: *Procesamiento y Clasificación de Fonemas.* En *Anales del XX Congreso Nacional de Ingeniería Biomédica*, Colima, México, Octubre 1997. Publicado en Revista Mexicana de Ingeniería Biomédica, Vol 18, Nro 2, 1197.

[170] Rufiner, H.L. y J. Goddard: *Procesamiento y Clasificación de Fonemas mediante Onditas y Redes con Retardos.* Revista Argentina de Bioingeniería, 4(1):11–20, Marzo 1998.

[171] Rufiner, H.L., J. Goddard, A.E. Martínez, and F.M. Martínez: *Basis pursuit applied to speech signals.* In *Proceedings 5th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2001) and the 7th International Conference on Information Systems Analysis and Synthesis (ISAS 2001)*, pages 517–520, Orlando, July 2001. IEEE. Paper No. H202.

[172] Rufiner, H.L., C. Martínez y H.M. Torres: *Clasificación de Fonemas mediante Paquetes de Onditas orientadas Perceptualmente y una Red Neuronal por Fonema.* En *Anales de las "VIII Jornadas de Jóvenes Investigadores Grupo Montevideo"*, Brasil, Septiembre 2000.

[173] Rufiner, H.L., L.F. Rocha, and J. Goddard: *Preserving acoustic cues in speech denoising.* In *Proc. of the 2nd Joint Meeting of the IEEE Engineering in Medicine and Biology Society and the Biomedical Engineering Society EMBS-BMES2002*, volume 1, pages 288–289, Houston, Texas, October 2002.

[174] Rufiner, H.L., L.F. Rocha, and J.Goddard: *Sparse and independent representations of speech signals based on parametric models.* In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pages 989–992, September 2002.

[175] Rufiner, H.L., M.E. Torres, L. Gamero, and D.H. Milone: *Introducing complexity measures in nonlinear physiological signals: application to robust speech recognition.* Physica A: Statistical Mechanics and its Applications, 332:496–508, February 2004.

[176] Ruggero, M.A.: *Responses to sound of basilar membrane of mammalian cochlea.* Current Opinion in Neurobiology, 2, May 1992.

[177] Saito, N.: *Local feature extraction and its applications using a library of bases.* PhD thesis, Department of Mathematics, Yale University, New Haven, CT 06520, USA, dec 1994.

[178] Saito, N.: *The least statistically-dependent basis and its applications.* In *Proc. 32nd Asilomar Conference on Signals, Systems, and Computers*, page 732–736. IEEE Press, 1998.

[179] Saito, N. and B. Benichou: *Sparsity vs. statistical independence in adaptive signal representations: A case study of the spike process.* In Welland, G. V. (editor): *Beyond Wavelets*, volume 10 of *Studies in Computational Mathematics*, chapter 9, pages 225–257. Academic Press, 2003.

[180] Saito, N., B. M. Larson, and B. Benichou: *Sparsity vs. statistical independence from a best-basis viewpoint.* In Aldroubi, A., A.F. Laine, and M.A. Unser (editors): *Wavelet Applications in Signal and Image Processing VIII*, Proc. SPIE 4119, pages 474–486, 2000.

[181] Saparin, P., A. Witt, J. Kurths, and B. Anishchenko: *The renormalized entropy – an appropriate complexity measure?* J. Chaos, Solitons and Fractals, 4(10):1907–1916, 1994.

[182] Sardy, S., A. Antoniadis, and P. Tseng: *Generalized basis pursuit.* Technical report, Swiss Federal Institute of Technology, October 2000.

[183] Schwartz, O. and E. P. Simoncelli: *Natural signal statistics and sensory gain control.* Nature Neuroscience, 4(8):819–825, august 2001.

[184] Secker-Walker and Searle: *Time-domain analysis of auditory-nerve-fiber firing rates.* Journal of the Acoustic Society of America, 88:637–642, 1990.

[185] Sejnowski, T.: *Open questions about computation in cerebral cortex.* In *Parallel Distributed Processing*, volume 2. The MIT Press, 1986.

[186] Shamma, S. A.: *Neural and functional models of the auditory cortex.* In Arbib, M. (editor): *Handbook of Brain Theory and Neural Networks*, Bradford Books. The MIT Press, 1995.

[187] Shamma, S.A.: *Auditory cortical representation of complex acoustic spectra as inferred from the ripple analysis method.* Comput. in Neural Syst., 7:439–476, 1996.

[188] Shannon, C. E.: *A mathematical theory of communication.* The Bell System Technical Journal, 27(3):379–423, 1948.

[189] Sigura, A.: *Sistema de Preprocesamiento Basado en Paquetes de Onditas y Algoritmos Genéticos para Reconocimiento Automático del Habla.* Tesis de Maestría, FIUNER-IPSJAE (Cuba), 2001. Dir. H.L. Rufiner.

[190] Silva, Adelino R. Ferreira da: *Bayesian wavelet denoising and evolutionary calibration.* Digital Signal Processing, 14(6):566–589, November 2004.

[191] Smaragdis, Paris: *Redundancy Reduction for Computational Audition, a Unifying Approach.* Ph.D. thesis, School of Architecture and Planning, Massachusetts Institute of Technology, June 2001.

[192] Steinberg, J.C. and N.R. French: *The portrayal of visible speech.* JASA, 18(1):4–18, 1946.

[193] Stevens, Kenneth N.: *Acoustic Phonetics.* MIT Press, 1998.

[194] Strang, G. and T. Nguyen: *Wavelets and Filter Banks.* Wellesley-Cambridge Press, 1996.

[195] Su, L.S., K.P. Li, and K.S. Fu: *Identification of speakers by use nasal coarticulation.* J. Acoust. Soc. Am., 56:1876–1882, 1974.

[196] Suga, N.: *What does single-unit analysis in the auditory cortex tell us about information processing in the auditory system?* In Rakic, P. and W. Singer (editors): *Neurobiology of the neocortex*. John Wiley & Sons, 1988.

[197] Tan, Beng T. and Minyue Fu: *A review of the performance of auditory processing front ends for an automatic speech recognizer in adverse environment*. Internal report, Dept. of Electrical and Computer Engineering, The University of Newcastle, Australia, May 1996.

[198] Tewfik, A. and M. Nafie: *An algebraic approach to the subset selection problem*. In *EUSIPCO*, Greece, September 1998.

[199] Theunissen, F.E., K. Sen, and A.J. Doupe: *Spectro-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds*. J. Neuroscience, 20:2315–2331, 2000.

[200] Torres, H. M. y H. L. Rufiner: *Identificación Automática del Hablante mediante Redes con Retardos*. En *Anales del XII Congreso Argentino de Bioingeniería*, volumen 1, páginas 319–322, Buenos Aires, Argentina, June 1999.

[201] Torres, H.M. and L.H. Rufiner: *Automatic speaker identification by means of mel cepstrum, wavelets and wavelets packets*. In *Proceedings of the Chicago 2000 World Congress IEEE EMBS*, July 2000. Paper No. TU–E201–02.

[202] Torres, H. y H. L. Rufiner: *Clasificación de fonemas mediante paquetes de onditas orientadas perceptualmente*. En *Anales del Ier Congreso Latinoamericano de Ingeniería Biomédica, Mazatlán 98*, volumen 1, páginas 163–166, México, Noviembre 1998.

[203] Torres, M. E.: *El procesamiento de señales ligadas a problemas no lineales.* Tesis de Doctorado, Universidad Nacional de Rosario - Argentine, 1999. (Math. D. Thesis).

[204] Torres, M. E., M. M. Añino, and G. Schlotthauer: *Slight parameter changes detection in complex signals: Multiresolution q–entropy automatic tool*. In *Proceedings of the 2001 Workshop on Nonlinear Signal and Image Processing (NSIP'2001)*, pages 1–5, Baltimore, Maryland, USA, June 2001. IEEE-EURASIP. Paper No. 1049.

[205] Torres, M. E. and L. G. Gamero: *Relative complexity changes in time series using information measures*. Physica A, 286(3-4):457–473, 2000.

[206] Trendelenburg, F.: *On the physics of speech sounds*. JASA, 7(1):142–147, 1935.

[207] Tsallis, C.: *Somme comments on Boltzmann-Gibbs statistical mechanics*. Chaos, Solitons and Fractals, 6:539, 1995. and references therein.

[208] Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, USA, 1995.

[209] Varga, A. and H. Steeneken: *Assessment for automatic speech recognition II NOISEX-92: A database and experiment to study the effect of additive noise on speech recognition systems*. Speech Communication, 12(3):247–251, 1993.

[210] Waibel, A., T. Hanazawa, G. Hinton, K. Shikano, and K. Lang: *Phoneme recognition using time–delay neural networks*. IEEE Transactions on Acoustics, Speech and Signal Processing, 37(3):328–339, 1989.

[211] Wang, K. and S. Shamma: *Self normalization and noise-robustness in early auditory representations*. IEEE Transactions on Speech and Audio, 2(3):421–435, 1994.

[212] Watson, A. B., H. B. Barlow, and J. G. Robson: *What does the eye see best?* Nature, 302:419–422, 1983.

[213] Wickerhauser, M.: *Lectures on Wavelets Packet Algorithms*. Washington University, Nov 1991.

[214] Wojtaszczyk, P.: *A Mathematical Introduction to Wavelets*. London Mathematical Society Student Texts. Cambridge University Press, 1997. No. 37.

[215] Young, E.D.: *What's the best sound.* Science, 280:1402–1403, 1998.

[216] Young, S.: *Large vocabulary continuous speech recognition: A review.* IEEE Workshop on Speech Recogniton, Diciembre 1995.

[217] Cambridge University: *HMM Toolkit*, htk v3.0 edition. `http://htk.eng.cam.ac.uk`.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
H. L. Rufiner; "Análisis y representación de la voz mediante técnicas no convencionales"
Universidad de Buenos Aires, Argentina, 2005.

Este documento se terminó de imprimir el 3 de octubre de 2005
Fue escrito en LaTeX, compilado con MiKTeX y editado en TeXnicCenter.