Representación rala de la señal de voz

H. L. Rufiner[1*], J. Goddard[2], L.F. Rocha[3]
[1] *Laboratorio de Cibernética, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, CC47, Suc. 3 (CP 3100), Paraná, Entre Ríos, Argentina*
[2]*Depto Ing. Eléctrica, Univ. Autónoma Metropolitana, México*
[3]*Instituto de Ingeniería Biomédica, Facultad de Ingeniería, Univ. de Buenos Aires, Argentina*

Resumen

Recientemente se han encontrado conexiones interesantes entre el análisis de señales mediante bases sobrecompletas y la manera en la que el cerebro parece procesar algunas señales sensoriales. Este tipo de análisis puede dar representaciones que poseen muy pocos elementos activos o diferentes de cero, o sea sumamente *ralas*. Esta es una característica que comparten con los sistemas sensoriales biológicos, y hace que sean fácilmente codificables en términos de trenes de pulsos o espigas. Las representaciones ralas poseen también una robustez intrínseca al ruido aditivo, cuando este no pueda ser expresado fácilmente en términos de los elementos de la "base". Se puede decir también que describen a las señales que representan de una manera que permite una adecuada generalización, lo que es particularmente importante en el contexto de la clasificación de las mismas. Existen varios métodos que permiten encontrar una representación rala de una señal si se posee una base sobrecompleta adecuada. Así mismo se pueden diseñar los elementos de estas bases a la medida del tipo de pistas que se pretenden encontrar en la señal, o buscar automáticamente la base o diccionario que satisfaga algún criterio particular. Los métodos utilizados para este fin pueden plantearse a partir de encontrar la solución de un problema de optimización determinístico u estocástico con las restricciones adecuadas. Existen también varias formas para medir la eficacia de la representación lograda, es decir la "bondad" de la codificación de los datos en términos de los coeficientes de la representación. Entre ellas existen principalmente dos grupos: las relacionadas con la dispersión de los coeficientes con respecto a una norma, y las derivadas de la estadística de estos coeficientes. En este trabajo se presentan los conceptos fundamentales detrás de este tipo de representaciones, los principales métodos para generarlas y se analiza su aplicación particular al caso de las señales de voz.

1. Introducción

Las señales de voz están entre las señales naturales más estudiadas. En el campo de análisis y modelado del habla se han realizado avances muy importantes. Sin embargo existen varios problemas que aún no han logrado resolverse satisfactoriamente. El desempeño actual de las maquinas está muy lejos del de las personas en tareas similares de análisis y reconocimiento del habla [1, 2].

Las personas realizan complicados análisis de señales a través de los sistemas neurosensoriales y extraen información útil acerca de su entorno en forma prácticamente transparente para ellos. El sistema auditivo logra descifrar el mensaje "escondido" en los patrones de variación sonora producidos por el aparato fonador. Entendemos el mensaje codificado en ella de manera asombrosamente sencilla, en forma casi independiente de factores como la identidad del hablante o el ruido de fondo. Se ha demostrado recientemente que esto obedece a una adaptación para procesar los mismos de manera óptima [3]. Los principios detrás de esta codificación están todavía siendo estudiados, pero ya existen numerosas pistas que permiten guiar nuestra búsqueda en esta dirección. Uno de estos principios es el de lograr codificar cada una de las señales implicadas en términos de solo unas pocas características significativas. Esto es lo que se denomina una representación rala. Recientemente se ha demostrado su utilización como esquema de codificación eficiente a nivel de los sistemas sensoriales biológicos [4]. Se puede decir que el propio código neural — basado en *trenes de espigas* — también es ralo. Existen varios trabajos que estudian este modelo para las representaciones generadas en los campos receptivos de la corteza visual, y más recientemente también en la corteza auditiva [3].

Se puede decir que un código ralo es aquel que representa la información en términos de un número pequeño de descriptores tomados de un conjunto grande [4], es decir que, sólo una fracción de los elementos del código son usados activamente para representar un patrón típico. En términos numéricos, esto se adopta a menudo para significar que la mayoría de los elementos son cero, o prácticamente cero la mayoría del tiempo [5, 6]. La dificultad reside en definir que significa "prácticamente cero". La suposición implícita es que esos valores que son "cercanos a cero" pueden tratarse como si fueran exactamente cero, con una pequeña o ninguna pérdida de información útil.

En realidad la codificación rala es solo una de las posibilidades dentro de un espectro que va desde los códigos locales a los distribuidos [5]. 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 clasificación de los patrones, sin embargo poseen varias desventajas como la cantidad de dimensiones, la capacidad de generalización, y la

[*] leorufiner@ciudad.com.ar

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
H. L. Rufiner, J. Goddard & L. F. Rocha; "Representación rala de la señal de voz"
Memorias del XIV Congreso Argentino de Bioingeniería, oct, 2003.

dependencia entre las dimensiones. En el otro extremo están los códigos completamente distribuidos, donde todos los elementos son utilizados para representar cada patrón. Aquí tenemos en general menos dimensiones pero perdemos la facilidad de clasificación 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, y sin embargo son capaces de representar componentes separadas de los datos de manera directa [7].
Se han definido varias "normas" que permiten medir cuán ralo es una representación o vector [8]. 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 *kurtosis* positivo grande. Esto se traduce en que poseen un pico muy agudo en cero y colas largas a ambos lados.
Entre los métodos disponibles para lograr representaciones ralas, si se dispone de una base o diccionario adecuado, pueden citarse *Búsqueda de Bases* (BP) [9], *Búsqueda Ajustada* (MP) [10], o el de la *Mejor Base Ortogonal* (BOB) [10]. El *Método de los Marcos* (MOF) [11] se ha empleado a veces como comparación aunque estrictamente no resulta en una representación rala. Un primer acercamiento a su utilización para el análisis de señales de voz en el caso particular de diccionarios fijos de funciones *paquetes de onditas* (WPT) se encaró en [12] con resultados muy prometedores. También existen métodos derivados de un enfoque más estadístico para hallar los coeficientes, que pueden asimilarse como equivalentes a BP, destacando que como resultado adicional también encuentran una base óptima. En este camino existen varios trabajos orientados al análisis de imágenes "naturales" [13], y más recientemente algunos a señales sonoras de audio y música [14]. En este último caso la representación o codificación rala mediante *bases o diccionarios sobrecompletos* (OCD) también posee importantes relaciones con la técnica de *Análisis de Componentes Independientes* (ICA) [15]. Este procedimiento se ha comenzado a aplicar al campo de las señales biomédicas en general [16], a problemas como el de la *deconvolución ciega* [17], y más recientemente al campo del ASR [18].
En este trabajo se exploran algunas alternativas para la obtención de representaciones ralas aplicadas al análisis de señales de voz.

2. Métodos

2.1 El problema de la Representación rala

En esta sección plantearemos el problema de como encontrar una representación rala de una señal. Podemos decir que queremos representar a la señal $\mathbf{s}$ en términos de diccionario $\mathbf{\Phi}$ y un conjunto de coeficientes a_j. En general se agrega también un término correspondiente al ruido $\boldsymbol{\varepsilon}$. Un diccionario $\mathbf{\Phi}$ es solo una colección de formas de onda o funciones parametrizadas $(\boldsymbol{\phi}_\gamma)_{\gamma \in \Gamma}$. A cada forma de onda $\boldsymbol{\phi}_\gamma$ se la suele denominar *átomo*. Se prefiere el término diccionario frente al de base, debido a que esta colección generalmente no cumple con las propiedades de una base ortogonal tradicional.
De esta forma la expresión que describe mi señal es la siguiente:

$$\mathbf{s} = \sum_{\gamma \in \Gamma} a_\gamma \boldsymbol{\phi}_\gamma + \boldsymbol{\varepsilon} \tag{1}$$

donde $\mathbf{s} \in R^N$ es la señal, $\mathbf{a} \in R^M$ es el vector de coeficientes de la representación y $\mathbf{\Phi}$ el diccionario de tamaño $N \times M$, con $M \geq N$, y el término de ruido aditivo $\boldsymbol{\varepsilon} \in R^N$.
Aunque la apariencia de la ecuación (1) parece sencilla, el principal problema consiste en que para el caso general $\mathbf{\Phi}$, $\mathbf{a}$ y $\boldsymbol{\varepsilon}$ son desconocidos, existiendo infinitas soluciones. Aún en el caso sin ruido y conociendo $\mathbf{\Phi}$, cuando los átomos son más que la cantidad de muestras de $\boldsymbol{s}$ (OCD), o bien cuando los átomos no forman una base ortogonal, esto produce representaciones no únicas de las señales. Por lo tanto se debe encontrar un criterio que permita seleccionar alguna de ellas. En ese caso, a pesar de que la ecuación es lineal, los coeficientes que se eligen para formar parte de la solución resultan de una función no lineal de los datos $\boldsymbol{s}$. Diferentes métodos, como MOF, MP, BOB, BP, se han propuesto para obtener una descomposición de este tipo. En 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 por ejemplo la *Transformada Discreta de Fourier*, esta inversión se simplifica debido a que $\mathbf{\Phi}^{-1} = \mathbf{\Phi}^{*T}$.
Para su solución el problema completo podría dividirse en dos sub-problemas, a saber: ¿Cómo encuentro la menor cantidad de coeficientes que representan a mi señal original, eliminando también los efectos del ruido? , y ¿Cómo armo el diccionario que mejor describa el tipo de señales a analizar? A estos se los denomina como el problema de la *inferencia* y el del *aprendizaje* respectivamente. El primero es el que se identifica en el caso biológico con el fenómeno de *percepción* o *inferencia*, y de allí su nombre. El último también obtiene su denominación del símil biológico y suele ser el más complejo de ambos. Para solucionarlos necesito disminuir las soluciones posibles del problema mediante una serie de restricciones o criterios útiles en mi aplicación, y un método que me permita encontrar una solución adecuada en base a la formulación original y a estas nuevas

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
H. L. Rufiner, J. Goddard & L. F. Rocha; "Representación rala de la señal de voz"
Memorias del XIV Congreso Argentino de Bioingeniería, oct, 2003.

restricciones planteadas. Un criterio importante para escoger un método consiste en obtener una representación rala de la señal. Otra vez, esto significa que uno desearía que sólo unos pocos coeficientes, a_γ en (1), sean diferentes de cero.
Existen diferentes enfoques para obtener una solución de estos problemas. Desde el punto de vista determinístico las restricciones aparecen en forma de la minimización de ciertas medidas, distancias o costos 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 densidad probabilidad de los coeficientes y el ruido, y la solución se obtiene otra vez como un problema de optimización, maximizando alguna *verosimilitud* o *probabilidad posterior*. Para un análisis más detallado de esta conexión entre el enfoque probabilista y el determinista puede consultarse [19].
Existen varias formas para medir la eficacia de la representación lograda, es decir la "bondad" de la codificación de los datos **s** en los coeficientes **a**, mediante (1). Así mismo esto permite establecer las restricciones necesarias para plantear el problema de optimización. Entre ellas existen principalmente dos grupos: las relacionadas con la dispersión de los coeficientes con respecto a una norma, y las derivadas de la estadística de los coeficientes. En el primer grupo están aquellas que calculan la norma o medida de dispersión promedio de los coeficientes **a** para un conjunto de datos. A pesar de denominárselas como normas muchas de estas medidas no verifican estrictamente todas las propiedades de las normas, como por ejemplo la de homogeneidad o la desigualdad del triangulo [8]. En el segundo grupo podemos encontrar ejemplos como el kurtosis o la *entropía*, que pueden ser útiles en este contexto. También puede calcularse el mínimo número de bits necesario para codificar la información contenida en los coeficientes, con un enfoque más orientado a teoría de la información y compresión. Otras medidas útiles pueden servir para medir el grado de "ajuste" del modelo a los datos, como la denominada *evidencia*, o el *error cuadrático medio* (MSE).

2.2 Problema de Inferencia

En esta sección se plantean algunas soluciones posibles al problema de la inferencia, suponiendo el diccionario $\mathbf{\Phi}$ conocido, primero con $\boldsymbol{\varepsilon} = 0$ y luego el caso ruidoso.

2.2.1 Caso limpio, enfoque determinístico

Búsqueda de Bases: En [9], Chen y colaboradores proponen un método, denominado BP que se diseñó para producir este tipo de 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 la señal **s** tiene longitud N y hay M formas de onda en el diccionario, entonces el problema para resolver es:

$$\min \|\mathbf{a}\|_1 \text{ sujeto a } \mathbf{\Phi a} = \mathbf{s} \tag{2}$$

donde **a** un es un vector en 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.
Este problema puede convertirse en uno programación lineal tradicional (con coeficientes sólo positivos) y puede resolverse eficazmente y exactamente con los métodos de punto interior.
Chen y colaboradores dan varios ejemplos con señales artificiales que muestran los beneficios de su método, en términos de dispersión (en inglés *sparsity*) y super-resolución, comparados con las representaciones correspondientes encontradas por MOF, MP y BOB.
Una desventaja de BP es que en realidad ℓ_1 es solo una aproximación a la medida real de dispersión que es ℓ_0, es decir el número total de elementos iguales a cero. Sin embargo en este caso el problema de optimización se convierte en uno de programación entera mucho más difícil de resolver, y cuyo costo computacional resulta prohibitivo para muchas aplicaciones prácticas.
A continuación presentaremos brevemente otros métodos que permiten encontrar una representación de las señales a los fines de su comparación.

Búsqueda Ajustada: En 1993 Mallat y Zhang [10] presentaron un método general para aproximar la descomposición (1) que encara el tema de la dispersión directamente. Comenzando a partir de una aproximación inicial $\mathbf{s}^{(0)}=\mathbf{0}$ y un residuo $\mathbf{R}^{(0)}=\mathbf{s}$, construye una secuencia de aproximaciones ralas paso a paso. En la etapa k se identifica el átomo del diccionario que mejor correlaciona con el residuo y luego suma a la aproximación actual un múltiplo escalar de este átomo, de manera que:

$$\mathbf{s}^{(k)} = \mathbf{s}^{(k+1)} + \alpha_k \boldsymbol{\phi}_k \tag{3}$$

donde $\alpha_k = \langle \mathbf{R}^{(k-1)}, \boldsymbol{\phi}_{\gamma k} \rangle$, y $\mathbf{R}^{(k)} = \mathbf{s}^{(k)} - \mathbf{s}^{(k-1)}$. Luego de *m* pasos, se obtiene una representación de la forma (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* al problema planteado en (1) y es por ello que padece los problemas y ventajas de este tipo de métodos de optimización.

Mejor Base Ortogonal: Para algunos diccionarios, es posible desarrollar esquemas de descomposición específicos. Los diccionarios *Paquetes de Onditas* (WPT) y *Paquetes de Cosenos* (CPT) son ejemplos, ya que poseen propiedades muy particulares. Algunas subcolecciones especiales de elementos en estos diccionarios son bases ortogonales. De esta forma se obtiene un amplio rango de bases ortonormales. Coifman y Wickerhauser [10] propusieron un método para seleccionar una sola base ortogonal en forma adaptativa de entre todas estas bases ortogonales, que es la "mejor base" en el sentido de la entropía de los coeficientes **a**. En algunos casos este algoritmo da representaciones ralas cercanas al óptimo, sin embargo esto sólo es posible cuando la representación se puede hacer en términos de una base ortogonal. Otra desventaja del método es que esta atado a los diccionarios ya citados.

Método de los Marcos: MOF [11] selecciona entre todas las posibles soluciones de (1), una cuyos coeficientes posean norma ℓ^2 mínima:

$$\min \|\mathbf{a}\|_2 \text{ sujeto a } \boldsymbol{\Phi}\mathbf{a} = \mathbf{s} \tag{4}$$

La solución de este problema es única, y la llamaremos $\tilde{\mathbf{a}}$. Geométricamente, la colección de todas las soluciones de (1) es un subespacio afín en R^M; MOF selecciona el elemento de este subespacio más cercano al origen. Esto es llamado a veces una solución de longitud mínima. Existe una matriz, la inversa generalizada de $\boldsymbol{\Phi}$, la cual calcula la solución de longitud mínima al sistema de ecuaciones lineales:

$$\tilde{\mathbf{a}} = \tilde{\boldsymbol{\Phi}}\mathbf{s} = \boldsymbol{\Phi}^T\left(\boldsymbol{\Phi}\boldsymbol{\Phi}^T\right)^{-1}\mathbf{s} \tag{5}$$

Para los diccionarios del tipo denominado "*marco ajustado*" (en inglés *Tight Frame*), MOF se halla disponible en una forma cerrada. Un buen ejemplo es el caso del diccionario WPT usual.
Existen dos problemas claves con MOF. Primero 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. Otro problema es su limitación de resolución intrínseca.

2.2.2 Caso ruidoso

Enfoque Determinístico: Un aspecto importante en la solución del problema considerado es cuando incluimos el término referente al ruido. Esto permite encontrar formas de realizar una limpieza al mismo tiempo que encontramos los coeficientes. Cabe recalcar el hecho de que en muchos de los enfoques para solucionar problemas de habla ruidosa primero se procesa la señal para extraer la información relevante, y sobre esa transformación se utiliza alguna técnica de limpieza de ruido. Es decir que en general no se busca necesariamente que la representación posea algún tipo de robustez como en este caso.
Una forma de incluir un tratamiento explícito del ruido en la obtención de representaciones ralas de nuestra señal es mediante la RT [20]. En este caso el enfoque es determinista y la idea principal consiste en agregar un término de penalización a la minimización de (3). Este término involucra generalmente alguna medida de ajuste de los datos al modelo, suponiendo la existencia de ruido aditivo.
En general la expresión regularizada tiene la siguiente forma:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} \left\{ d(\mathbf{s}, \boldsymbol{\Phi}\mathbf{a}) + \lambda f(\mathbf{a}) \right\}, \tag{6}$$

donde $d(\mathbf{s},\boldsymbol{\Phi}\mathbf{a})$ es alguna función que mide la distancia entre la señal y el modelo, $f(\mathbf{a})$ es alguna función de los coeficientes, y λ es un factor de peso.
Esto complica un poco nuestro problema de minimización original (2) volviéndolo uno de programación cuadrática. Se podría decir que en realidad estamos intentando solucionar 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 con *limpieza de ruido BP* (BPD), y posteriormente fueron extendidas a casos con ruidos no gaussianos en el trabajo de Sardy sobre *BP Generalizada* (GBP) [21].

En el caso propuesto por Chen (suponiendo ruido gaussiano) tenemos:

$$d(\mathbf{s},\mathbf{\Phi a}) = \left\|\mathbf{\Phi a} - \mathbf{s}\right\|^2 \text{ y } f(\mathbf{a}) = |\,\mathbf{a}\,| \tag{7}$$

Este enfoque es paralelo al probabilista que veremos a continuación y arroja una serie de soluciones equivalentes a las de las ecuaciones de esta sección. Obsérvese que para el caso determinista se trata de minimizaciones (del error o distancia) y para el probabilista de maximizaciones (de la verosimilitud o probabilidad). Existen otros trabajos [5] donde se exploran otras posibilidades pero con un enfoque de las $f(\mathbf{a})$ como funciones de activación de una red neuronal que también tiene una interpretación probabilística.

Enfoque Estadístico: Como vimos el ruido complica un poco más el problema, pero permite también de manera natural su tratamiento desde el punto de vista probabilista o estadístico. Debido a que van a existir señales que sean más probables que otras para un diccionario y un conjunto de coeficientes dados, esto nos permite tomar decisiones si conocemos por ejemplo la distribución de probabilidad $P(\mathbf{s}|\mathbf{\Phi},\mathbf{a})$. Para obtener una representación rala podemos suponer una distribución con kurtosis positivo para cada coeficiente a_i. Además podemos asumir que los a_i sean estadísticamente independientes con una distribución a priori conjunta:

$$P(\mathbf{a}) = \prod_i P(a_i) \tag{8}$$

Por esta última propiedad a los códigos generados de esta forma se los suele llamar también *códigos factoriales* (FC) y además esto conecta los resultados con la técnicas basadas en ICA. Siguiendo la terminología utilizada en ICA, (1) es lo que se denomina el *modelo generativo*, para significar que genera la señal $\mathbf{s} \in R^N$ a partir de un conjunto de fuentes ocultas a_j, arregladas como un vector de estado $\mathbf{a} \in R^M$, utilizando una matriz de mezcla o diccionario $\mathbf{\Phi}$ de tamaño $N \times M$, con $M \geq N$, e incluyendo un término de ruido aditivo $\boldsymbol{\varepsilon}$ (generalmente gaussiano). Si se conoce $\mathbf{\Phi}$ y $\mathbf{s}$, podemos estimar $\mathbf{a}$ considerando la distribución a posteriori:

$$P(\mathbf{a}|\Phi,\mathbf{s}) = \frac{P(\mathbf{s}|\Phi,\mathbf{a})P(\mathbf{a})}{P(\mathbf{s}|\Phi)} \tag{9}$$

Una estimación de $\mathbf{a}$ de *probabilidad a posteriori máxima* (MAP) sería:

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmax}} \left[\log P(\mathbf{s}|\Phi,\mathbf{a}) + \log P(\mathbf{a})\right] \tag{10}$$

Si la posterior es suficientemente suave, podemos encontrar el máximo por gradiente ascendente. Como se puede notar la solución depende de la forma de la distribución asumida para el ruido y para los coeficientes (distribución a priori), dando lugar a diferentes métodos para el cálculo de los coeficientes que mencionaremos a continuación.
En [22] se utiliza para la a priori una función exponencial de la forma:

$$P(a_i) = \frac{1}{Z_s} e^{-f(a_i)}, \tag{11}$$

$$f(a_i) = \beta . \log(1 + (a_i/\sigma)^2), \tag{12}$$

$$Z_s = \int e^{-f(a_i)} da_i \tag{13}$$

Si asumimos ruido aditivo gaussiano $\boldsymbol{\varepsilon}$ con matriz de covarianza $\left\langle \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} \right\rangle = \mathbf{\Lambda}_\varepsilon^{-1}$, entonces la probabilidad de observar una $\mathbf{s}$ particular dada una base $\mathbf{\Phi}$ conocida y coeficientes $\mathbf{a}$ resulta ser:

$$P(\mathbf{s} \mid \Phi, \mathbf{a}) = N e^{-\frac{1}{2}\varepsilon^T \Lambda_\varepsilon \varepsilon} \tag{14}$$

y la solución MAP mediante gradiente ascendente se convierte en la siguiente regla de actualización para $\mathbf{a}$:

$$\Delta \mathbf{a} = \mathbf{\Phi}^T \mathbf{\Lambda}_\varepsilon \boldsymbol{\varepsilon} - f'(\mathbf{a}) \tag{15}$$

En [23] Lewicki y Olshausen proponen utilizar una distribución a priori de tipo laplaciano:

$$P(a_i) = N \exp(-\theta_i |a_i|), \tag{16}$$

en conjunción con ruido gaussiano esto nos lleva a la siguiente regla de actualización para **a**:

$$\Delta_{\mathbf{a}} = \mathbf{\Phi}^T \mathbf{\Lambda}_\varepsilon \varepsilon - \boldsymbol{\theta}^T |\mathbf{a}| \tag{17}$$

Esta regla es equivalente a la BPD propuesta por Chen previamente presentada [9].
La aproximación seguida por Plumbley [14] para lograr una distribución rala, consiste en una versión mixta formada por dos laplacianas de la siguiente forma:

$$P(a_i) = \begin{cases} N \exp(-|a_i|) & \text{si } |a_i| \geq \mu, \\ NC \exp(-K|a_i|) & \text{si } |a_i| < \mu, \end{cases} \tag{18}$$

donde N es una constante de normalización, y $C = \exp(\mu(K\text{-}1))$ para asegurar continuidad. Los parámetros μ y K controlan el ancho y la masa relativa del pico central. Este tipo de distribución lleva a un comportamiento tipo umbralamiento al calcular **â** [6] y la solución MAP mediante gradiente ascendente se convierte en la siguiente regla de actualización para **a**:

$$\Delta_{\mathbf{a}} = \mathbf{\Phi}^T \mathbf{\Lambda}_\varepsilon \boldsymbol{\varepsilon} - \gamma(\mathbf{a}) \tag{19}$$

donde

$$\gamma(\mathbf{a}) = \begin{cases} \text{signo}(\mathbf{a}) & \text{si } |\mathbf{a}| \geq \mu, \\ K.\text{signo}(\mathbf{a}) & \text{si } |\mathbf{a}| < \mu, \end{cases} \tag{20}$$

En [21] Sardy, desarrolla un método general para encontrar los coeficientes en el caso de distribución a priori laplaciana para distintas distribuciones convexas del ruido, no necesariamente gaussianas, mediante GBP. En el 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 [20] Girosi plantea otra alternativa que tendría que ver con otros tipos de ruido pero estableciendo una conexión entre RT y *máquinas de soporte vectorial* a través de las funciones núcleos [24].

2.3 El Problema del Aprendizaje

Ante el desarrollo anterior surge naturalmente la pregunta acerca de como elegir o generar un diccionario **Φ** adecuado para nuestra aplicación. En este sentido existen, en principio, dos enfoques posibles. Uno consiste en "armar" el diccionario a mano mediante la utilización del conocimiento a priori sobre las pistas o características que se quieren encontrar en la señal. La otra forma corresponde a la búsqueda automática imponiendo algunas restricciones a la solución de la ecuación (1), utilizando para ello datos reales de señales adecuadas. Como ya dijimos este problema es bastante más complejo y demandante que el de la inferencia.
Para el primer enfoque se puede hacer uso de los diccionarios basados en funciones paramétricas "tradicionales" (Gabor, Fourier, WT, WPT, CPT, Impulsos, Heaviside, etc.), eligiendo alguno que represente las características de los distintos tipos de señales adecuadamente, o tratando de buscar los conjuntos de funciones adecuados para cada tipo de señal y luego mezclando todos en un "superdiccionario".
En el caso automático podemos usar ideas similares a las expuestas para el caso de la inferencia, otra vez con enfoques determinísticos [19] o estocásticos [22]. Dada la extensión del presente articulo invitamos al lector interesado a consultar estas referencias para más detalles acerca de los métodos para solucionar este problema. En [25] los autores presentan un método que aprovecha las correlaciones temporales existentes en la señal de voz a fin de lograr una mejor representación de las mismas.

3. Resultados y Discusión

En el presente trabajo se aplican las técnicas mencionadas para lograr representaciones ralas de distintos fonemas del habla. Las señal de voz presenta comportamientos altamente transitorios junto con trozos estacionarios. Por tanto el uso del enfoque anterior es muy atractivo, permitiendo a las formas de onda en el diccionario adaptarse a la señal particular bajo consideración, extrayendo los rasgos pertinentes. La posibilidad de super-resolución y dispersión también presenta ventajas sobre las representaciones tradicionales, como las basadas en la transformada de Fourier. Finalmente resultan de particular interés en las aplicaciones la robustez intrínseca de estas representaciones. En esta sección se demostrarán estas propiedades para el caso de análisis y descomposición de distintos fonemas del habla castellana. Debido a la extensión del presente artículo solo se

incluirán aquí los resultados derivados de utilizar diccionarios fijos y los métodos determinísticos para encontrar los coeficientes.

Los experimentos reportados en este trabajo 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.

La selección se basó en elegir un conjunto de fonemas representativos de vocales, plosivas sonoras y 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. Esto resultó en señales con anchos de 1024 o 2048 muestras. Las señales conservaron su frecuencia de muestreo original de 16 KHz. Se utilizó el mismo diccionario sobrecompleto para todos los experimentos, que consistía 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 luego de algunas pruebas preliminares con distintos tipos y parámetros de onditas, tomando en cuenta el porcentaje de coeficientes diferentes de cero en las representaciones logradas.

Se aplicaron BP, MP, BOB y MOF a las señales extraidas de todos los fonemas utilizando el programa *Atomizer* desarrollado por Chen [9].

En la **Figura 1** se muestran los sonogramas de los fonemas /p/ (I) y /b/ (II) con sus correspondientes espectrogramas y con los gráficos de magnitud de los coeficientes de la representación. Los espectrogramas mostrados son de banda angosta y puede observarse claramente como los eventos temporales importantes, como por ejemplo el momento de la explosión del fonema /p/, se "diluyen" completamente. También puede apreciarse en los gráficos (c) que la representación lograda no es para nada rala.

En la **Figura 2** pueden observarse la representaciones tiempo-frecuencia (T-F) logradas por BP, MP, BOB y MOF para el caso del fonema /b/. Aquí puede observarse como BP, MP e inclusive BOB logran representaciones bastante ralas. Obsérvese por ejemplo la representación obtenida mediante BP (a), donde la porción inicial del fonema de naturaleza "cuasi-senoidal" ha sido detectada perfectamente a través del trazo horizontal correspondiente en el plano T-F. Así mismo los eventos temporales quedan también perfectamente descriptos sin pérdida de su localización. De esta manera podemos ver como los métodos estudiados logran comportarse como si fueran una "base adaptativa" que utiliza en la representación aquellos elementos que mejor describen a la señal. De esta forma, si bien no podemos evitar el principio de incertidumbre, la resolución tiempo-frecuencia en cada sector del plano se adapta en función de las características de la señal analizada. Se debe hacer notar que MP fue utilizado con una opción para seleccionar solo los primeros 1000 coeficientes y esto 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.

Para entender un poco mejor cuán rala es la representación lograda en la **Figura 3**, se ha procedido a reconstruir la señal del fonema /p/ (limpio (I) y con ruido blanco a 10 dB SNR (II)) utilizando solo los 15 átomos más importantes (de un total de 11264). En la parte inferior de la figura se pueden apreciar los átomos seleccionados por BP para realizar la síntesis en ambos casos. Nótese como los átomos empleados en ambos casos son muy similares variando 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 aprovecharon en [26] para proponer un método heurístico de limpieza de ruido que preserva las pistas acústicas de la señal de voz.

En la **Figura 4** (a) se muestran los resultados del *error cuadrático medio* (ECM) obtenido para cada fonema luego de aplicar cada método (en los casos limpio (I) y con ruido blanco a 10 dB SNR (II)) para seleccionar los 15 átomos más significativos (como en la figura anterior). En la **Figura 4** (b) se muestran los correspondientes porcentajes de los coeficientes que superan un umbral del 5% del máximo valor absoluto. Esto puede tomarse como una medida sencilla del grado de dispersión que estaría relacionada con ℓ_0. En los resultados del ECM mostrados en la **Figura 4** (I)(a) puede apreciarse como los errores más grandes se cometen en la vocales, esto en realidad se debe a que son los fonemas que poseen mayor energía relativa. Para el resto de los fonemas, los fricativos poseen el mayor error. En la **Figura 4** (II)(a) puede apreciarse como el ECM disminuye un poco en el caso ruidoso, esto se debe a que el error de reconstrucción se calcula sobre la señal ruidosa. El resto de las características son similares a las del caso limpio. En cuanto a la **Figura 4** (I)(b) puede verse como en todos los casos MOF es la que provee una representación menos rala. BP y MP proveen representaciones suficientemente ralas de los fonemas, con una adecuada localización de las pistas acústicas (ver **Figura 2**). 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. El caso con ruido (**Figura 4** (II)(b)) las representaciones logradas son un poco menos ralas debido a que se utilizan algunos átomos también para describir el ruido.

(I) (II)

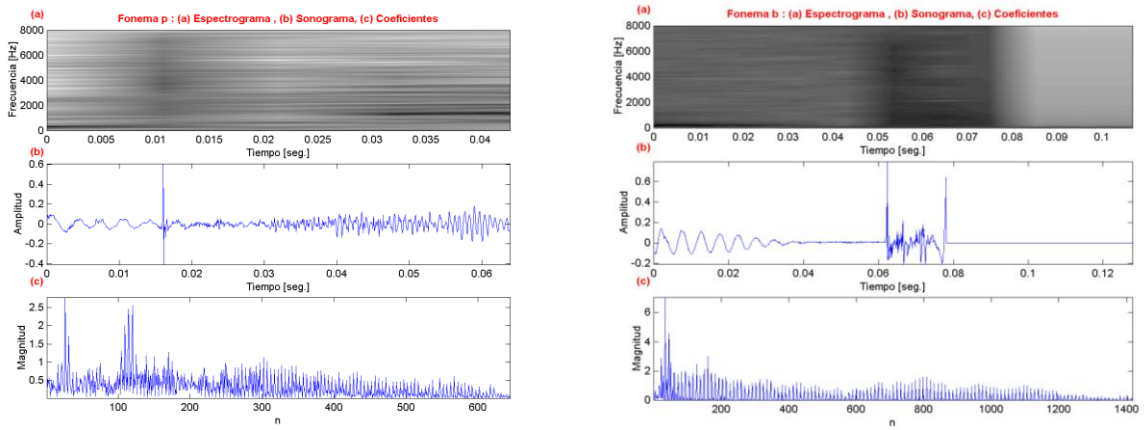

Figura 1: Espectrograma(a), sonograma (b) y magnitud de los coeficientes (c) para: (I) señal correspondiente al fonema /p/, y (II) señal correspondiente al fonema /b/

(I) (II)

(III) (IV)

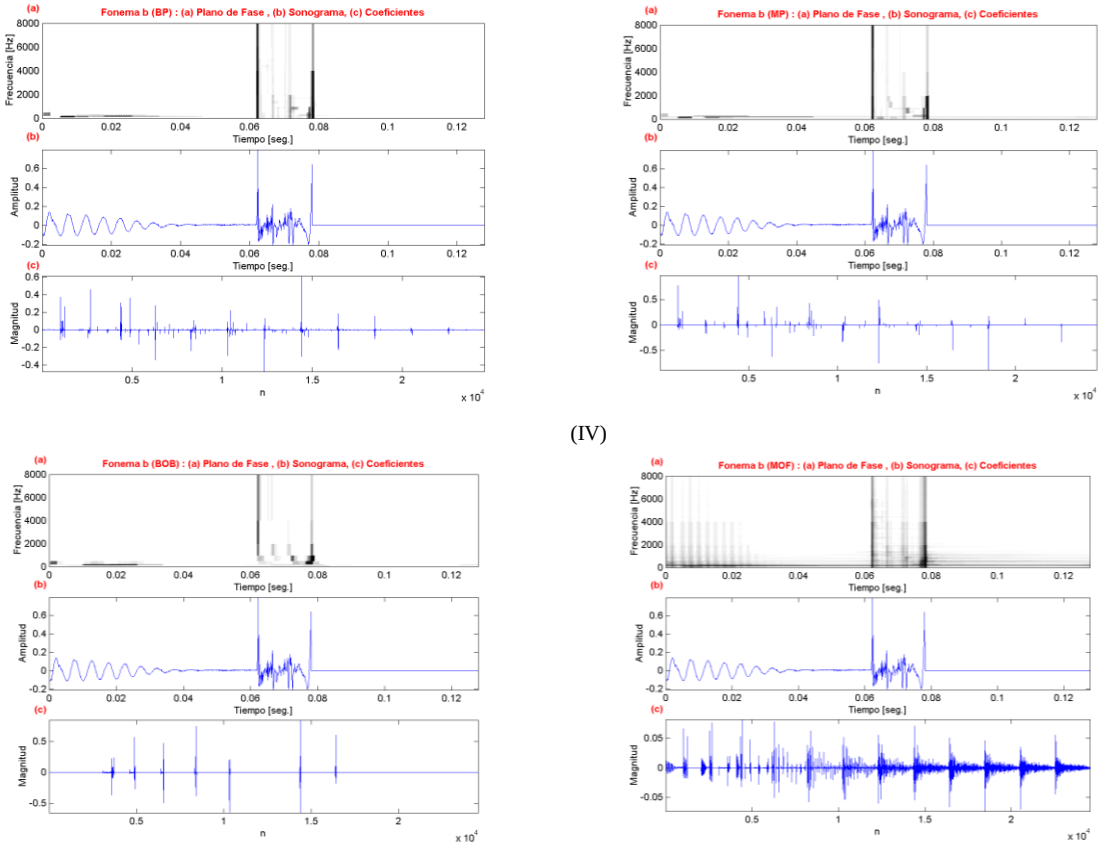

Figura 2: Plano Tiempo-Frecuencia (a), sonograma (b) y magnitud de los coeficientes correspondiente al fonema /b/ (c) para: (I) BP, (II) MP, (II) BOB, y (IV) MOF.

(I) (II)

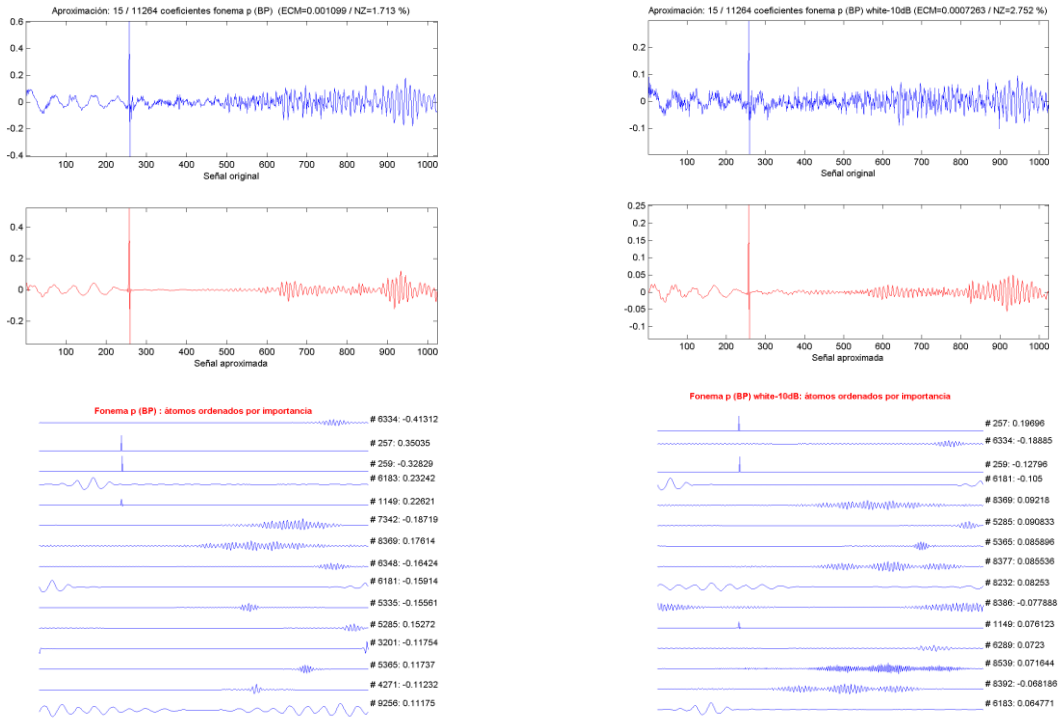

Figura 3: Reconstrucción por medio de los primeros BP átomos más importantes para el fonema /p/ ((I) limpio y (II) con ruido blanco a 10 dB SNR): (a) Señal Original (b) Aproximación (c) Átomos y coeficientes utilizados en la aproximación.

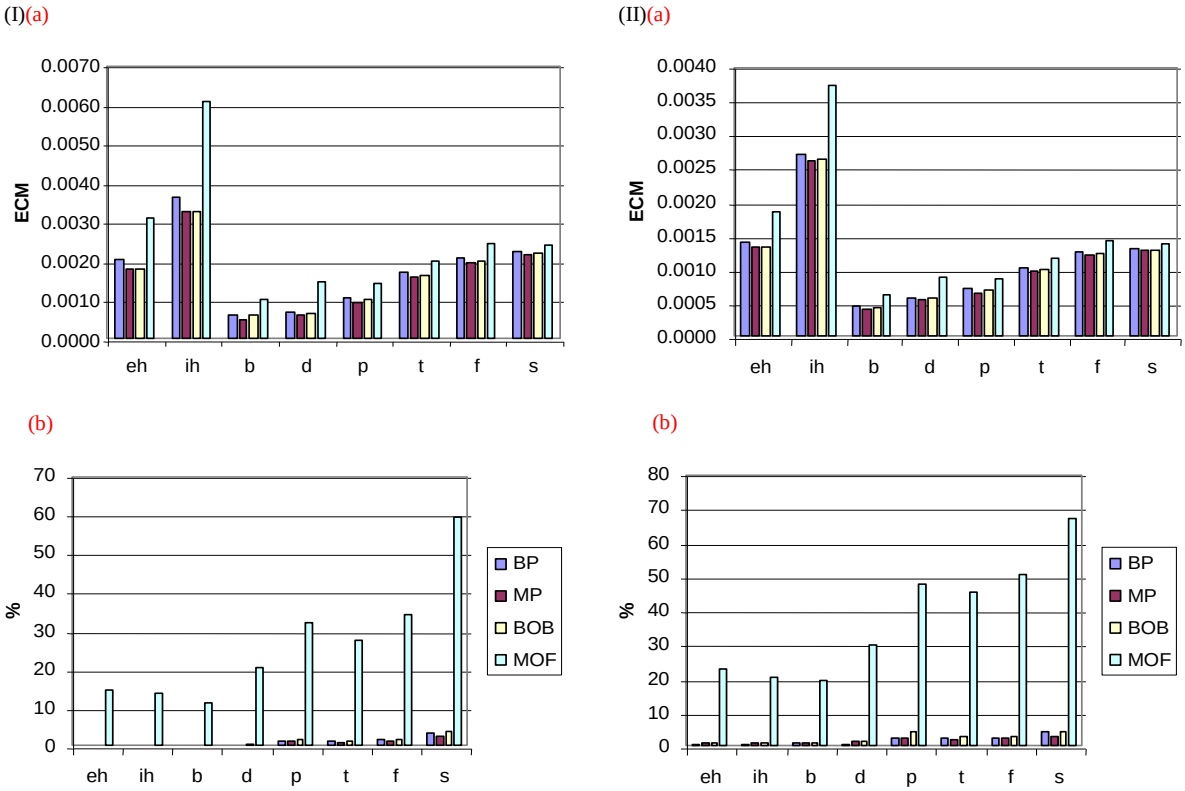

Figura 4: Espectrogramas(a), sonogramas (b) y magnitud de los coeficientes (c) para: (I) señal correspondiente al fonema /p/, y (II) señal correspondiente al fonema /p/

4. Conclusiones y trabajo futuro

En este trabajo se han revisado los fundamentos detrás de los métodos para lograr representaciones ralas de las señales. Mediante ejemplos de señales obtenidas de fonemas del español se han mostrado y discutido los resultados de aplicar BP, MP, BOB y MOF con diccionarios fijos a estas señales en condiciones originales y 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 solución a este problema, preservando las características importantes inclusive en presencia de ruido. Por supuesto esta mejora en las capacidades es a costa de incrementar el costo computacional de las técnicas empleadas en el análisis.

Un tema pendiente es el de estudiar con más detalle los efectos en el cambio del diccionario. También es importante continuar investigando acerca de los métodos de aprendizaje, ya que estos permiten encontrar los diccionarios que mejor representen a una clase particular de señales.

Referencias

[1] Richard P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 19, no. 22, pp. 1-15, 1997.

[2] H. Boulard, H. Hermansky, y N. Morgan, "Towards increasing speech recognition error rates," *Speech Communication*, vol. 18, no. 3, pp. 205-231, 1996.

[3] M. Lewicki, ``Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, pp. 356-363, 2002

[4] B.A. Olshausen y D.J. Field, "Emergence of simple cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607-609, 1996.

[5] George Francis Harpur, *Low Entropy Coding with Unsupervised Neural Networks*, Ph.D. thesis, Department of Engineering, University of Cambridge, Queens' College, February 1997.

[6] A. Hyvärinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum-likelihood estimation," Tech. Rep., Helsinki University of Technology, 1998.

[7] Harri Lappalainen, "A computationally efficient algorithm for finding sparse codes," M.S. thesis, Neural Networks Research Centre, Laboratory of Computer and Information Science, Helsinki University of Technology, 1996.

[8] F. Guspí y B. Introcaso, "Soluciones ralas de sistemas lineales indeterminados," *El Ingeniero en la Red*, vol. 1, no. VII, pp. 1-10, Mayo 2000, Revista Electrónica FCEIyA, UNR, Argentina.

[9] S.S. Chen, D.L. Donoho, y M.A. Sanders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33-61, 1999.

[10] R. Coifman y M.V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, March 1992.

[10] S.G. Mallat y Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. in Signal Proc.*, vol. 41, pp. 3397-3415, December 1993.

[11] I. Daubechies, "Time-frequency localization operators: a geomeric phase space aproach," *IEEE Trans. Info. Thry*, vol. 34, no. 4, pp. 605-612, 1988.

[12] H.L. Rufiner, J. Goddard, A.E. Martínez, y F.M. Martínez, "Basis pursuit applied to speech signals," en *5th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI)*, Orlando, Julio 2001, IEEE.

[13] B.A. Olshausen y D.J. Field, "Vision and the coding of natural images," *American Scientist*, vol. 88, no. 3, pp. 238-245, 2000.

[14] S. A. Abdallah y M. D. Plumbley, "Sparse coding of music signals," 2001.

[15] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94-128, 1999.

[16] T.-P. Jung, S. Makeig, T.-W. Lee, M.J. McKeown, G. Brown, A.J. Bell, y T.J. Sejnowski, "Independent component analysis of biomedical signals," http: //www.cnl.salk.edu/ jung/ica.html, 2001.

[17] T.-W. Lee, M.S. Lewicki, M. Girolami, y T.J Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Sig. Proc. Lett.*, 1998.

[18] J.H. Lee, H.Y. Jung, T.W Lee, y S.Y. Lee, "Speech feature extraction using independent component analysis," in *Proc. ICASSP*, 2000, vol. 3, pp. 1631-1634.

[19] K. Kreutz-Delgado y B.D. Rao, "Sparse basis selection, ICA, and majorization: Towards a unified perspective," in *Proc. ICASSP*, 1999, vol. 2, Paper no. 2411.

[20] F. Girosi, "An equivalence between sparse approximation and support vector machines," *Neural Computation*, vol. 10, no. 6, pp. 1455-1480, 1998.

[21] S. Sardy, A. Antoniadis, y P. Tseng, "Generalized basis pursuit," Tech. Rep., October 2000.

[22] B.A. Olshausen, *Sparse codes and spikes*, chapter 13, MIT Press, 2001, In Press.

[23] M.S. Lewicki y B.A. Olshausen, "A probabilistic framework for the adaptation and comparison of image codes," *Journal of the Optical Society of America*, vol. 16, no. 7, pp. 1587-1601, 1999.

[24] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, NY, USA, 1995.

[25] H. L. Rufiner, L. F. Rocha, y J. Goddard Close, "Sparse and independent representations of speech signals based on parametric models," en *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, pp. 989-992, Septiembre 2002.

[26] H. L. Rufiner, L. F. Rocha, y J. Goddard Close, "Preserving acoustic cues in speech denoising," en *Proc. of the 2nd Joint Meeting of the IEEE Engineering in Medicine and Biology Society And the Biomedical Engineering Society EMBS-BMES2002*, Octubre 2002.