Clasificación de fonemas mediante representaciones corticales auditivas

Hugo L. Rufiner[1], César E. Martínez[1] y John Goddard[2]
[1]*Facultad de Ingeniería - Bioingeniería - UNER, lrufiner@bioingenieria.edu.ar*
[2]*Dpto. de Ingeniería Eléctrica - UAM-Iztapalapa, México, jgc@xanum.uam.mx*

***Resumen*— El empleo de métodos de procesamiento de señales *biológicamente inspirados* ha permitido mejorar el desempeño de los sistemas artificiales que tratan de emular algunos aspectos de la comunicación humana. A partir de técnicas recientes como el análisis de componentes independientes o las representaciones ralas es posible lograr un análisis de la señal de voz con características muy similares a las obtenidas experimentalmente a nivel de la corteza auditiva primaria. En este trabajo se presenta una primera aproximación al problema de clasificación de fonemas empleando este tipo de representaciones. Los resultados respecto a los datos de entrenamiento mejoran notablemente a los obtenidos a partir del enfoque clásico basado en los coeficientes cepstrales en escala de Mel.**

***Palabras clave*— Análisis de componentes independientes, representaciones ralas, representación auditiva cortical, reconocimiento del habla**

I. INTRODUCCIÓN

LA aparición de nuevos enfoques en el campo del análisis y la representación de señales promete superar algunas de las limitaciones de las técnicas clásicas en problemas reales con señales complejas como la voz humana. A partir de estas *representaciones no convencionales* se pueden plantear soluciones alternativas para problemas como el de limpieza de ruido o el de reconocimiento automático del habla. 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 [1].

En el proceso de comunicación humana, 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 [2]. A pesar del conocimiento que se tiene acerca de las representaciones auditivas tempranas, los principios que sustentan la representación de la señal de voz a niveles sensoriales más altos como en la *corteza auditiva primaria* (AI), son todavía objeto de estudio [3]. 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. Es posible entonces plantear un modelo para las representaciones corticales, lográndose correlaciones importantes con las características de las representaciones reales obtenidas experimentalmente [4].

Para obtener este modelo cortical se utilizan técnicas relacionadas con el *análisis de componentes independientes* (ICA) y las *representaciones ralas* (SR). Estas técnicas permiten emular el comportamiento de las neuronas corticales a partir de los *campos receptivos espectro-temporales* (STRF) [5]. Se puede decir que los STRF constituyen el estímulo óptimo requerido para que una neurona cortical auditiva responda con la mayor activación posible. Para su estimación mediante datos de la actividad neuronal se utilizan métodos como el de la *correlación inversa* [6]. A partir de las representaciones tiempo-frecuencia de los espectrogramas auditivos, se puede estimar un diccionario de átomos bidimensionales que son los que finalmente modelan a los STRF.

Este trabajo se organiza como se explica a continuación. La Sección II presenta el modelo utilizado para la representación de la voz. En la Sección III se explica como esta representación puede asimilarse a la existente a nivel de la corteza auditiva primaria. La Sección IV detalla los pasos seguidos para la obtención de los patrones de representación cortical, y el diseño de los experimentos de clasificación de fonemas en este nuevo espacio. La Sección V expone los resultados obtenidos en la experimentación, junto a una discusión sobre los mismos. Finalmente, la Sección VI resume los aportes de este trabajo y plantea los trabajos futuros en esta dirección.

II. REPRESENTACIÓN RALA Y FACTORIAL

En lo que sigue se plantea cuál es el modelo utilizado para representar las señales de voz en forma rala y factorial junto con la forma en la que se estiman los parámetros de esta representación.

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:

$$\mathbf{x} = \sum_{\gamma \in \Gamma} \boldsymbol{\phi}_\gamma a_\gamma + \boldsymbol{\varepsilon} = \mathbf{\Phi}\mathbf{a} + \boldsymbol{\varepsilon}, \qquad (1)$$

donde $\boldsymbol{\varepsilon} \in \mathbb{R}^N$ constituye un término de ruido aditivo y $M \geq N$. El diccionario $\mathbf{\Phi}$ resulta en una colección de formas de onda o funciones parametrizadas $(\boldsymbol{\phi}_\gamma)_{\gamma \in \Gamma}$, donde cada forma de onda $\boldsymbol{\phi}_\gamma$ constituye un átomo.

Aunque la apariencia de la ecuación (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 $\mathbf{x}$. 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 $\mathbf{\Phi}^{-1} = \mathbf{\Phi}^T$.

Por lo discutido hasta aquí un criterio de interés para seleccionar una representación, de entre todas las posibles, consiste en que ésta sea lo más rala posible y también la más "independiente".

Para obtener una representación rala puede suponerse una distribución con curtosis positivo para cada coeficiente a_i. Además para que los a_i sean estadísticamente independientes puede utilizarse una distribución a priori conjunta de la forma:

$$P(\mathbf{a}) = \prod_i P(a_i). \tag{2}$$

Es posible ver a (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:

$$P(\mathbf{a}|\mathbf{\Phi}, \mathbf{x}) = \frac{P(\mathbf{x}|\mathbf{\Phi}, \mathbf{a})P(\mathbf{a})}{P(\mathbf{x}|\mathbf{\Phi})}. \tag{3}$$

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

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} \left[\log P(\mathbf{x}|\mathbf{\Phi}, \mathbf{a}) + \log P(\mathbf{a})\right]. \tag{4}$$

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 éstos.

En [7] Lewicki y Olshausen proponen utilizar una distribución a priori de tipo laplaciana con paramétro β_i :

$$P(a_i) = \alpha \exp -\beta_i |a_i|, \tag{5}$$

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 $\mathbf{a}$:

$$\Delta\mathbf{a} = \mathbf{\Phi}^T \mathbf{\Lambda}_{\boldsymbol{\varepsilon}} \boldsymbol{\varepsilon} - \boldsymbol{\beta}^T |\mathbf{a}|. \tag{6}$$

Para estimar el valor de $\mathbf{\Phi}$, es posible maximizar la siguiente función objetivo:

$$\mathcal{L} = \mathcal{E}[\log P(\mathbf{x}|\mathbf{\Phi})]_{P(\mathbf{x})}, \tag{7}$$

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:

$$P(\mathbf{x}|\mathbf{\Phi}) = \int_{\mathbb{R}^M} P(\mathbf{x}|\mathbf{\Phi}, \mathbf{a})\, P(\mathbf{a})\, d\mathbf{a}, \tag{8}$$

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:

$$\frac{\partial \mathcal{L}}{\partial \phi_{ij}} = 0, \tag{9}$$

se obtiene una regla de actualización para la matriz $\mathbf{\Phi}$:

$$\Delta\mathbf{\Phi} = \eta \mathbf{\Lambda}_{\boldsymbol{\varepsilon}}\ \mathcal{E}\left[\boldsymbol{\varepsilon}\mathbf{a}^T\right]_{P(\mathbf{a}|\mathbf{\Phi},\mathbf{x})}, \tag{10}$$

donde $\eta \in \mathbb{R}$ es un coeficiente de aprendizaje (que varía entre 0 y 1).

III. Representación cortical

Las propiedades de los sistemas sensoriales deben coincidir con la estadística de los estímulos naturales que operan sobre ellos [8]. Si se supone un modelo sencillo para describir estos estímulos, como el planteado en la ecuación (1), es posible entonces estimar sus propiedades a partir del enfoque estadístico presentado en la sección anterior.

El sistema auditivo codifica aspectos importantes para la discriminación fonética en los espectrogramas auditivos. En esta representación, de nivel más alto que el temporal, 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 [9]. Por ello, esta representación constituye un buen punto de partida para lograr otras más elaboradas siguiendo el símil biológico.

La estimación de un diccionario de átomos bidimensionales $\mathbf{\Phi}$ correspondientes a características tiempo-frecuencia estimados a partir de datos $\mathbf{x}$ del espectrograma auditivo resulta equivalente a las STRF de un grupo de neuronas corticales. El nivel de activación de cada neurona puede asimilarse entonces con los coeficientes a_γ en (1).

Kording *et al* realizaron un análisis cualitativo de diccionarios obtenidos de manera similar y compararon sus propiedades favorablemente con las de los campos receptivos reales [4].

IV. Materiales y Métodos

De acuerdo con las consideraciones anteriores se procedió a diseñar un experimento de clasificación de fonemas que permitiera evaluar las capacidades de un sistema que emplee una representación cortical para esta tarea. Para ello se utilizaron los datos de habla correspondientes a la región DR1 de TIMIT para el conjunto de cinco fonemas altamente confundibles */b/*, */d/*, */jh/*, */eh/*, */ih/*. Para cada una de las emisiones se calculó el espectrograma auditivo correspondiente, a partir de un modelo auditivo temprano [10] mediante el paquete de rutinas NSL[1]. Posteriormente se redujo la resolución frecuencial de los datos para disminuir sus dimensiones. De esta manera se obtuvieron espectrogramas auditivos con un total de 32 ó 16 coeficientes frecuenciales 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 sirvieron de base para la estimación de los diccionarios. En la Figura IV se pueden apreciar algunos de los pasos de este proceso.

[1] `http://www.isr.umd.edu/CAAR/pubs.html`

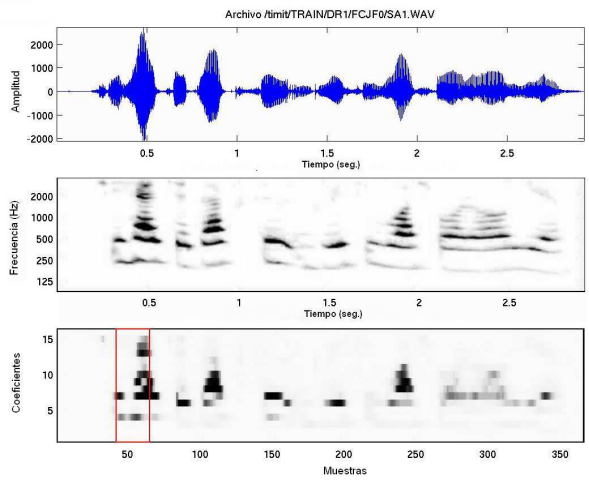

Fig. 1. Algunos pasos del proceso para generar los patrones espectro-temporales que sirven de base para estimar los STRF: Sonograma (arriba), espectrograma auditivo original (centro) y espectrograma de baja resolución (abajo). En este último se ha remarcado la sección correspondiente a la ventana deslizante a partir de la cual se genera cada uno de los patrones espectro-temporales.

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 de átomos bidimensionales utilizando la aproximación a la ecuación (10) seguida por Lewicki y Sejnowski que culmina en el conjunto de rutinas denominado NOCICA [11]. Se corrieron varias pruebas para los casos completos y sobrecompletos de cada configuración.

Un ejemplo de los diccionarios aprendidos se muestra en la Figura 2. Este caso corresponde al diccionario completo $\mathbf{\Phi} \in \mathbb{R}^{512x512}$, utilizando patrones de 32x16. Se pueden observar diversos comportamientos característicos útiles para la discriminación entre los diferentes fonemas que formaron parte del material empleado para el aprendizaje. 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). 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, y patrones bien localizados en tiempo o en frecuencia.

Una vez obtenidos los STRF se estimaron los coeficientes de activación a partir de los datos de los espectrogramas auditivos en forma iterativa mediante NOCICA utilizando (6). La clasificación se llevó a cabo mediante redes neuronales utilizando un *Perceptrón Multi-Capa* (PMC). La arquitectura empleada consistió en una capa oculta y una capa de salida de 5 neuronas, con variación en la cantidad de nodos de entrada según la dimensión de los patrones.

V. Resultados y discusión

Los resultados de los experimentos descriptos en la sección anterior se pueden observar en la Tabla I. Como se puede apreciar, los resultados de clasificación sobre los datos de entrenamiento para la representación cortical son mejores que los obtenidos al utilizar la representación auditiva directa. Más aún, estos resultados son mejores que los obtenidos con representaciones clásicas para esta tarea, como los coeficientes cepstrales en escala de Mel (ver tabla I). 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 incluso llegan a ser peores 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, generando detectores demasiado específicos de características presentes principalmente en los datos de entrenamiento.

Se evaluó la significancia estadística de estos resultados estimando la probabilidad de que un clasificador dado resulte mejor que el de referencia ($\Pr(\epsilon_{ref} > \epsilon)$). Para realizar esta estimación se supuso independencia estadística de los errores para cada tramo, y se aproximó la distribución binomial de los errores por medio de una distribución gaussiana. Ésto es posible debido a que se tiene una cantidad suficientemente grande de tramos (mayor para entrenamiento que para prueba). De esta forma se obtiene una $\Pr(\epsilon_{ref} > \epsilon) > 95\,\%$ en todos los casos.

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 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.

VI. Conclusiones

En este trabajo se presenta una aproximación a la clasificación de fonemas empleando una técnica no convencional para modelización del habla, que encuentra una representación auditiva temprana de la señal de voz a nivel cortical.

Basados en las analogías establecidas con los sistemas neurosensoriales, se estimaron diccionarios óptimos sobre esta representación. 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”.

Utilizando como patrones estas representaciones basadas en características bidimensionales, se entrenaron Perceptrones Multi-Capa como clasificadores de fonemas. Los resultados obtenidos sobre los datos de entrenamiento mejoran ampliamente aquéllos respecto a patrones MFCC estándar.

Harpur realizó experimentos sencillos de clasificación de fonemas utilizando códigos de baja entropía, con coeficientes sólo positivos generados a partir de bancos de filtros [12]. Sin embargo, no se han reportado experimentos mediante representaciones obtenidas a partir de modelos de campos receptivos auditivos más elaborados como los que aquí se presentan.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
H. L. Rufiner, C. E. Martínez & J. Goddard; "Clasificación de fonemas mediante representaciones corticales auditivas"
XV Congreso Argentino de Bioingeniería (SABI 2005), sep, 2005.

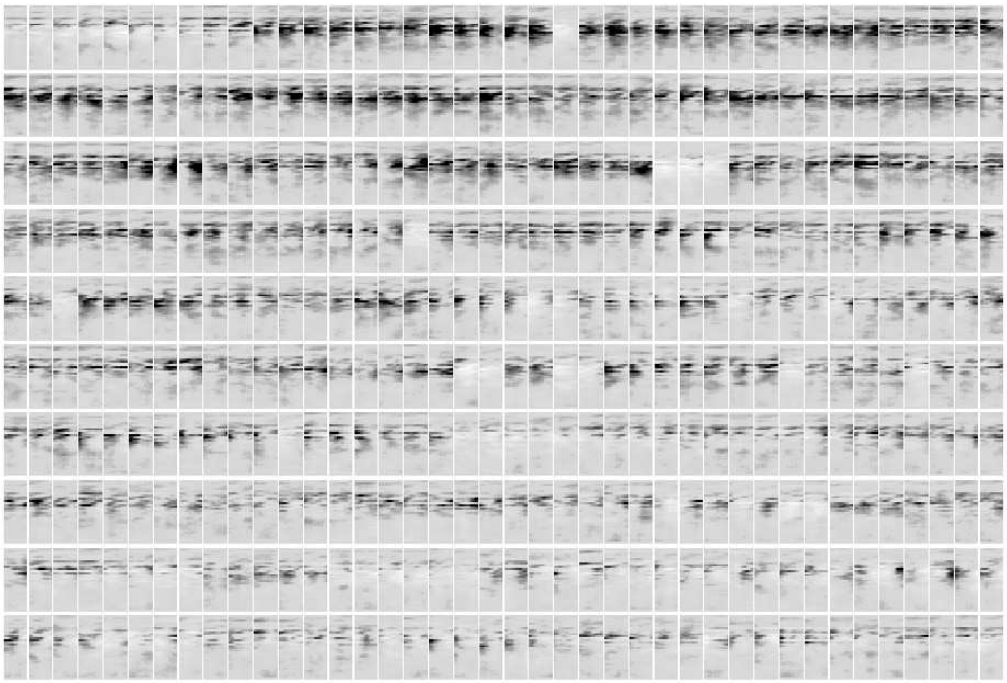

Fig. 2. Campos receptivos espectro-temporales estimados a partir de los patrones de 32x16 obtenidos de las representaciones auditivas tempranas de emisiones de de la región DR1 de TIMIT (TRAIN).

TABLA I

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.

Nº	EXPERIMENTO	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
12	Mel Cepstra+En. (128,14)	14+14/28/5	77.39	77.28	46.51	75.38	91.11	80.56	74.40

Aunque la capacidad de generalización no resultó completamente satisfactoria, estos resultados impulsan la investigación y experimentación con esta representación alternativa de la señal de voz, dada su similaridad biológica con las características del procesamiento a nivel cortical.

Agradecimiento Este trabajo fue soportado mediante proyecto PICT 11-12700A (ANPCyT-UNER).

REFERENCIAS

[1] S. Greenberg, "The ears have it: The auditory basis of speech perception," en *Proceedings of the International Congress of Phonetic Sciences*, vol. 3, pp. 34–41, 1995.

[2] B. Delgutte, "Physiological models for basic auditory percepts," en *Auditory Computation*, H. Hawkins, T. McMullen, A. Popper, y R. Fay, Eds. New York: Springer, 1996.

[3] S. A. Shamma, "Neural and functional models of the auditory cortex," en *Handbook of Brain Theory and Neural Networks*, ser. Bradford Books, M. Arbib, Ed. The MIT Press, 1995.

[4] K. P. Kording, P. Konig, y D. J. Klein, "Learning of sparse auditory receptive fields," en *Proc. of the International Joint Conference on Neural Networks (IJCNN '02)*, vol. 2, pp. 1103–1108, Honolulu, HI, United States, May 2002.

[5] F. Theunissen, K. Sen, y A. Doupe, "Spectro-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds," *J. Neuroscience*, vol. 20, pp. 2315–2331, 2000.

[6] R. deCharms, D. Blake, y M. Merzenich, "Optimizing sound features for cortical neurons," *Science*, vol. 280, pp. 1439–1443, 1998.

[7] M. Lewicki y B. 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.

[8] H. Barlow, "Redundancy reduction revisited," *Network: Computation in Neural Systems*, no. 12, pp. 241–253, 2001.

[9] O.-W. Kwon y T.-W. Lee, "Phoneme recognition using ica-based feature extraction and transformation," *Signal Processing*, vol. 84, no. 6, pp. 1005–1019, 2004.

[10] X. Yang, K. Wang, y S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inf. Theory*, vol. 38, pp. 824–839, 1992, special Issue on Wavelet Transforms and Multiresolution Signal Analysis.

[11] M. Lewicki y T. Sejnowski, "Learing overcomplete representations," en *Advances in Neural Information Processing 10 (Proc. NIPS'97)*, pp. 556–562. MIT Press, 1998.

[12] G. F. Harpur, "Low entropy coding with unsupervised neural networks," Disertación doctoral, Department of Engineering, University of Cambridge, Queens' College, February 1997.