Anales Acad. Nac. de Cs. Ex., Fís. y Nat., tomo 58 (2017): 71-78.

EXTRACCIÓN DE CARACTERÍSTICAS BIOINSPIRADAS BASADAS EN UN MODELO CORTICAL AUDITIVO

Hugo L. Rufiner[1, 2], *César E. Martínez*[1, 2], *Diego H. Milone*[1], *John Goddard*[3]

[1] Laboratorio de Señales e Inteligencia Computacional (SINC), Depto. Informática
Facultad de Ingeniería y Cs. Hídricas - Universidad Nacional del Litoral,
CC 217, Ciudad Universitaria, Paraje El Pozo, S3000 Santa Fe, Argentina
Tel: +54 (342) 457-5233 x 148, lrufiner@fich.unl.edu.ar
[2] Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Argentina
[3] Depto. de Ingeniería Eléctrica - UAM-Iztapalapa, México

Resumen

El empleo de métodos de procesamiento de señales de representación inspirados ha permitido mejorar el desempeño de los sistemas de representación que tratan de emular algunos aspectos de la representación humana. A partir de técnicas recientes como el representante de componentes independientes o las representaciones ralas es posible lograr una representación de la señal de voz con características muy similares a las obtenidas de representaciones a nivel de la corteza auditiva primaria. En este trabajo se presenta una nueva representación para la señal de voz, basada en los campos receptivos espectro-temporales, y se aplica por primera vez a un problema de clasificación de fonemas. Los resultados obtenidos mejoran notablemente los de las representaciones auditivas tempranas, e incluso los del enfoque clásico basado en los coeficientes cepstrales en escala de Mel. Se analiza también estas representaciones corticales frente al ruido aditivo.

Palabras claves: representaciones ralas, representaciones de componentes independientes, representaciones corticales auditivas, reconocimiento del habla.

Abstract

Bioinspired feature extraction by means of auditory cortical model. The use of biologically inspired, feature extraction methods has improved the performance of artificial systems that try to emulate some aspect of human communication. Recent techniques, such as independent component analysis and sparse representations, have made it possible to undertake speech signal analysis using features similar to the ones found experimentally at the primary auditory cortex level. In this work, a new type of speech signal representation, based on the spectro-temporal receptive fields, is presented, and a problem of phoneme classification is tackled for the first time using this representation. The results obtained are compared, and found to greatly improve both an early auditory representation and the classical front-end based on Mel frequency cepstral coefficients. Additive noise robustness analysis is also performed.

Key words: Sparse representations, independent component analysis, auditory cortical representation, speech recognition.

1. Introducción

El desarrollo de nuevas técnicas para el análisis y la representación de señales promete superar algunas de las limitaciones de los métodos clásicos en problemas reales con señales complejas, tales como los relacionados con el reconocimiento de 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 [8, 19].

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 [4].

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 [20]. 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. Utilizando ambos principios es posible entonces plantear un modelo para las representaciones corticales, lográndose correlaciones importantes con las características de las representaciones reales obtenidas experimentalmente [11, 12].

Para obtener este modelo cortical se utilizan técnicas relacionadas con el *análisis de componentes independientes* (ICA) y las *representaciones ralas* (SR) [6, 16]. Estas técnicas permiten emular el comportamiento de las neuronas corticales mediante el concepto de *campo receptivo espectro-temporal* (STRF) [21]. Se puede definir un STRF (en términos de su descripción tiempo-frecuencia) como el estímulo óptimo requerido para que una neurona cortical auditiva responda con la mayor activación posible. Para su estimación a partir de datos reales de la actividad neuronal en mamíferos se utilizan diferentes métodos, como por ejemplo el de la *correlación inversa* [3].

En este trabajo se estima un diccionario óptimo de átomos bidimensionales a partir de las representaciones tiempo-frecuencia de los espectrogramas auditivos de señales de voz. Utilizando este diccionario de STRFs se emula la activación a nivel de la corteza auditiva a través del cálculo de una representación rala e independiente. La representación obtenida es luego utilizada en un experimento de clasificación de fonemas, diseñado para evaluar la conveniencia de esta representación.

Este trabajo se organiza como se explica a continuación. La Sección 2 presenta el modelo utilizado en el artículo para la representación de la señal de voz. En particular, en la Sección 2.3, se explica como esta representación puede asimilarse a la existente a nivel de la corteza auditiva primaria. La Sección 3 detalla los datos utilizados para los experimentos de clasificación, así como también los pasos seguidos para la obtención de los patrones de representación cortical. La Sección 4 expone los resultados obtenidos en la experimentación, junto a una discusión sobre los mismos. Finalmente, la Sección 5 resume los aportes de este trabajo y plantea los trabajos futuros en esta dirección.

2. Representaciones ralas y factoriales

2.1. *Representaciones basadas en diccionarios discretos*

Existen diferentes posibilidades para representar una señal mediante diccionarios discretos generales. Para el caso donde el diccionario constituye una transformación unitaria u ortogonal, las técnicas para encontrar la representación de una señal resultan particularmente sencillas. Esto se debe, entre otros aspectos, a que la representación es única. Sin embargo, para el caso general no ortogonal existen muchas representaciones posibles de una señal a partir de un único diccionario. En estos casos es posible seleccionar una representación adecuada en base a pautas o criterios adicionales. En el problema que nos ocupa, dos criterios útiles para lograr una representación de la señal con características "sensoriales" consisten en que la misma resulte rala e independiente [17]. Además, es posible también encontrar el diccionario óptimo en términos de estos criterios [18].

Un código ralo es aquel que representa la información en términos de un número pequeño de descriptores tomados de un conjunto grande [18]. Esto quiere decir que sólo una pequeña fracción de los elementos del código son utilizados activamente para representar un patrón típico. En términos numéricos, esto significa que la mayoría de los elementos son cero, o "casi" nulos, la mayor parte del tiempo [9, 10].

Es posible definir varias medidas o normas que permitan cuantificar cuan rala es una representación, por ejemplo a través de la norma ℓ_0 o la norma ℓ_1. 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 la curtosis positivo grande. Esto 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. En el contexto estadístico resulta relativamente sencillo incluir aspectos relacionados con la independencia de los coeficientes, lo que conecta este enfoque con ICA [16].

A continuación se presentará una descripción formal de un método estadístico que permite estimar un diccionario óptimo y la representación correspondiente[1].

2.2. *Representaciones ralas y factoriales óptimas*

Sea $\vec{x} \in R^N$ una señal a representar en términos de un *diccionario* $\vec{\Phi}$, de tamaño $N \times M$, y un conjunto de coeficientes $\vec{a} \in R^M$. De este modo, la señal puede describirse como:

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

donde $\vec{\varepsilon} \in R$ es un término de ruido aditivo y $M \geq N$. El diccionario $\vec{\Phi}$ está compuesto por una colección de ondas o funciones parametrizadas $\left(\vec{\phi}_\gamma\right)_{\gamma \in \Gamma}$, donde cada onda $\vec{\phi}_\gamma$ es un *átomo* de la representación.

A pesar de que (1) parece muy simple, el problema consiste en que para el caso más general $\vec{\Phi}$, $\vec{a}$ y $\vec{\varepsilon}$ son desconocidos, y por lo tanto puede haber un número infinito de posibles soluciones. Más aún, en el caso sin ruido (cuando $\vec{\varepsilon} = \vec{0}$) y para $\vec{\Phi}$ dado, si hay más átomos que la cantidad de muestras de $\vec{x}$, o si los átomos no forman una base, entonces existe más de una representación posible para la señal. Para recuperar la unicidad se requiere entonces un método para seleccionar sólo una de estas representaciones. En este caso –a pesar de que se trata de un sistema lineal– los coeficientes seleccionados para ser parte de la solución generalmente poseen una relación no lineal con los datos $\vec{x}$. Para el caso completo y sin ruido, la relación entre los datos y los coeficientes es lineal y está dada por $\vec{\Phi}^{-1}$. Para las transformaciones clásicas, como la transformada discreta de Fourier, el cálculo de esta inversa se simplifica porque $\vec{\Phi}^{-1} = \vec{\Phi}^*$ $\left(\text{con } \vec{\Phi} \in C^{N \times M}\right.$ y $\left.\Phi^*(i,j) = \overline{\Phi(j,i)}\right)$.

Cuando $\vec{\Phi}$ y $\vec{x}$ son conocidos, una forma interesante de elegir el conjunto de coeficientes $\vec{a}$ entre todas las posibles representaciones, consiste en encontrar aquellos a_i que hacen la representación tan rala e independiente como sea posible. Para que la representación obtenida sea rala, puede suponerse que cada coeficiente a_i posee una distribución de probabilidad con curtosis positiva. Además, suponiendo la independencia estadística de los a_i, la distribución *a priori* conjunta satisface:

$$P(\vec{a}) = \prod P(a_i) \qquad (2)$$

El sistema descripto por (1) puede ser visto también como un *modelo generativo*, o sea un modelo para generar los datos $\vec{x}$. Siguiendo la terminología acostumbrada en el campo de ICA, esto significa que la señal $\vec{x} \in R^N$ se genera a partir de un conjunto de fuentes a_i (en la forma de un vector de estado $\vec{x} \in R^M$) utilizando una matriz de mezcla $\vec{\Phi}$ (de tamaño $N \times M$, con $M \geq N$), e incluyendo un término de ruido aditivo $\vec{\varepsilon}$ (gaussiano, en la mayoría de los casos).

El vector de estado $\vec{a}$ puede ser estimado a partir de la distribución *posterior*:

$$P\left(\vec{a}\middle|\vec{\Phi},\vec{x}\right) = \frac{P\left(\vec{x}\middle|\vec{\Phi},\vec{a}\right)P(\vec{a})}{P\left(\vec{x}\middle|\vec{\Phi}\right)}. \qquad (3)$$

De esta forma, una estimación *maximum a posteriori* de $\vec{a}$ sería:

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

Cuando $P\left(\vec{a}\middle|\vec{\Phi},\vec{x}\right)$ es suficientemente suave, el máximo puede encontrarse por el método de gradiente ascendente. La solución depende de las formas funcionales asignadas a las distribuciones del ruido y de los coeficientes, originando diferentes métodos para encontrar los coeficientes. Lewicki y Olshausen [14] propusieron el uso de una distribución *a priori* laplaciana con parámetro β_i:

$$P(a_i) = \alpha \exp\left(-\beta_i |a_i|\right), \qquad (5)$$

donde α es una constante de normalización. Utilizando esta distribución, en conjunto con la suposición de ruido aditivo gaussiano $\vec{\varepsilon}$, se obtiene la siguiente regla de actualización para $\vec{a}$:

$$\Delta\vec{a} = \vec{\Phi}^T \vec{\Lambda}_{\vec{\varepsilon}} \vec{\varepsilon} - \vec{\beta}^T |\vec{a}|, \qquad (6)$$

donde $\vec{\Lambda}_{\vec{\varepsilon}}$ es la inversa de la matriz de covarianza del ruido $E\left\{\vec{\varepsilon}^T \vec{\varepsilon}\right\}$, con $E\{\cdot\}$ significando al valor esperado.

[1] Por razones de simplicidad se describe sólo el caso unidimensional, aunque en el artículo se utilizan patrones bidimensionales.

El valor de $\vec{\Phi}$ puede estimarse maximizando la siguiente función objetivo [14]:

$$\hat{\vec{\Phi}} = \arg\max_{\Phi}\left[L\left(\vec{x},\vec{\Phi}\right)\right], \tag{7}$$

donde $L = E\left\{\log P\left(\vec{x}\middle|\vec{\Phi}\right)\right\}_{P(\vec{x})}$ es la verosimilitud de los datos. Esta verosimilitud puede encontrarse marginalizando la siguiente distribución condicional de los datos, dado el diccionario y los coeficientes, junto con la distribución a priori de los coeficientes:

$$P\left(\vec{x}\middle|\vec{\Phi}\right) = \int_{R^M} O\left(\vec{x}\middle|\vec{\Phi},\vec{a}\right)P(\vec{a})d\vec{a}, \tag{8}$$

donde la integral está definida sobre el espacio de estados M -dimensional de $\vec{a}$.

La función objetivo en (7) puede maximizarse utilizando gradiente ascendente a partir de la siguiente regla de actualización para la matriz $\vec{\Phi}$ [1]:

$$\Delta\vec{\Phi} = \eta\vec{\Lambda}_{\varepsilon}E\left\{\vec{\varepsilon}\vec{a}^{T}\right\}_{P(\vec{a}|\vec{\Phi},\vec{x})}, \tag{9}$$

donde η, en el rango (0, 1), es la tasa de aprendizaje.

De esta forma iterativa, pueden obtenerse tanto el diccionario $\vec{\Phi}$ como los coeficientes $\vec{a}$.

2.3. *Representaciones corticales auditivas*

Resulta conocido el principio por el cual los sistemas sensoriales adaptan sus propiedades a la estadística de los estímulos naturales que operan sobre ellos (para realizar su función de manera óptima) [2]. 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 temprano codifica las pistas importantes para la discriminación fonética, como las que pueden encontrarse en los espectrogramas auditivos. En esta representación –de más alto nivel que el acústico– 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, por ejemplo, la fase relativa entre algunas ondas acústicas [13]. Por ello, siguiendo el símil biológico, esta representación constituye un buen punto de partida para lograr otras más elaboradas.

La estimación de un diccionario de átomos bidimensionales $\vec{\Phi}$, correspondientes a características tiempo-frecuencia obtenidas a partir de datos $\vec{x}$ del espectrograma auditivo, resulta equivalente a las STRF de un grupo de neuronas corticales. De esta forma, el nivel de activación de cada neurona puede asimilarse con el valor de los coeficientes a_γ en (1). La Figura 1 muestra un diagrama esquemático del método utilizado para estimar esta representación cortical. Kording *y cols.* realizaron un análisis cualitativo de diccionarios obtenidos de manera similar y compararon sus propiedades favorablemente con las de los campos receptivos reales [12].

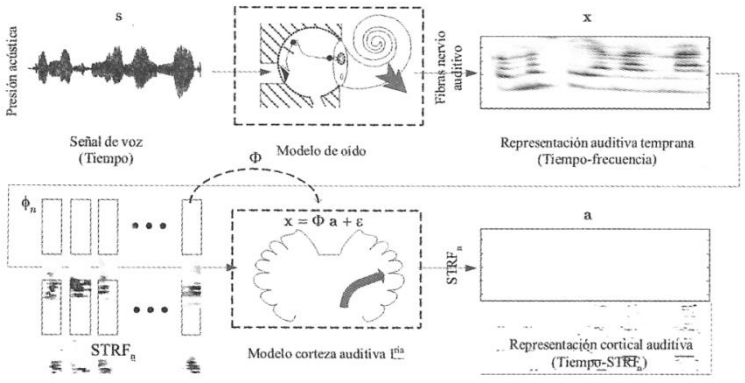

Fig. 1. Diagrama esquemático del método empleado para estimar la representación auditiva cortical.

3. Datos y experimentos

Se diseñó un experimento para clasificación de fonemas, de acuerdo a las consideraciones expuestas, para estimar el desempeño de un sistema que utilice una representación cortical para esta tarea. Para ello se utilizaron los datos de habla del conjunto de cinco fonemas altamente confundibles /*b*/, /*d*/, /*jh*/, /*eh*/, /*ih*/ de la región DR1 del corpus TIMIT [7] (Ver Tabla I).

Para cada una de las emisiones, muestreadas a 16 KHz, se calculó el correspondiente espectrograma auditivo utilizando el modelo auditivo temprano [22]. Luego se redujo la resolución frecuencial de los datos para disminuir la cantidad de dimensiones, obteniéndose espectrogramas auditivos de 64 coeficientes frecuenciales por unidad de tiempo. Finalmente, por medio de una ventana deslizante de 32 mseg desplazada a intervalos de 8 mseg, se obtuvo el conjunto de patrones espectro-temporales para la estimación de los diccionarios. En la Fig. 2 se muestran las principales etapas de este proceso junto con las señales correspondientes.

Tabla I. Distribución de los patrones por clase para los datos de entrenamiento (TRN) y prueba (TST).

Fonema	TRN		TST	
	#	(%)	#	(%)
/*b*/	211	(3.26)	66	(3.43)
/*d*/	417	(6.45)	108	(5.62)
/*jh*/	489	(7.56)	116	(6.04)
/*eh*/	2753	(42.58)	799	(41.63)
/*ih*/	2594	(40.13)	830	(43.25)
Total	6464	(100.00)	1919	(100.00)

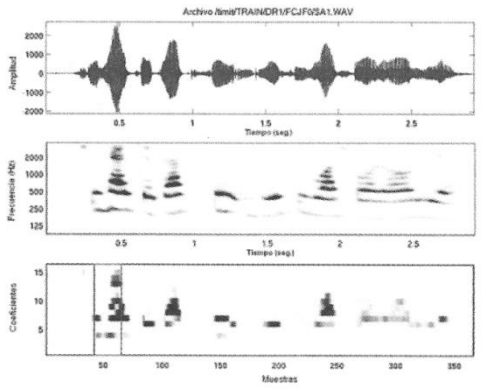

Fig. 2. Representaciones resultantes de las principales etapas en el proceso utilizado para generar los patrones espectro-temporales que sirven de base para obtener los STRFs: sonograma (arriba), espectrograma auditivo original (centro) y espectrograma de baja resolución (abajo). En esta última representación se ha marcado una sección correspondiente a la ventana deslizante, a partir de la cual se genera cada patrón tiempo-frecuencia.

A partir de estos patrones espectro-temporales se entrenaron diferentes diccionarios bidimensionales utilizando (9) [15]. Se realizaron pruebas para distintas configuraciones, para los casos completo y sobrecompleto.

Una vez estimados los STRF, se calcularon los coeficientes de activación de manera iterativa utilizando (6) a partir de los correspondientes espectrogramas auditivos de los fonemas. A los fines comparativos se calcularon también los *coeficientes cepstrales en escala de mel* (MFCC) con un coeficiente de energía (MFCC+E), de la manera usual para dos tramos consecutivos de la señal, resultando patrones en R^{28} [5].

En cada experimento, la clasificación se llevó a cabo utilizando una red neuronal artificial del tipo *perceptrón multicapa* (MLP). La arquitectura de la red consiste en una capa de entrada, donde el número de unidades de entrada depende de la dimensión de los patrones, una capa oculta, y una capa de salida con 5 unidades. El número de unidades en la capa oculta fue variando dependiendo del experimento.

Se realizaron también algunos experimentos preliminares a los fines de verificar la robustez de esta representación al ruido aditivo.

4. Resultados y discusión

En la Fig. 3 se muestra un ejemplo de algunas de las STRFs obtenidas. Este caso corresponde a un diccionario completo $\vec{\Phi} \in R^{256\times256}$, utilizando patrones de 64×4. Se pueden observar varios comportamientos típicos, que son útiles para discriminar entre diferentes fonemas. La posición relativa para cada elemento del diccionario está relacionada con su similitud con los otros elementos del diccionario (en términos de la norma ℓ_2 de sus diferencias). Es posible observar que las STRF parecen actuar como detectores de diferentes características fonéticas significativas para esta escala temporal, como por ejemplo frecuencias únicas, patrones estables de formantes, cambios en formantes, componentes sordas o ruidosas y patrones bien localizados en el tiempo o la frecuencia.

Los resultados de los experimentos descriptos en la sección anterior se detallan en la Tabla II. Como se puede observar en la tabla, los resultados de clasificación tanto para entrenamiento como para prueba son mejores que los

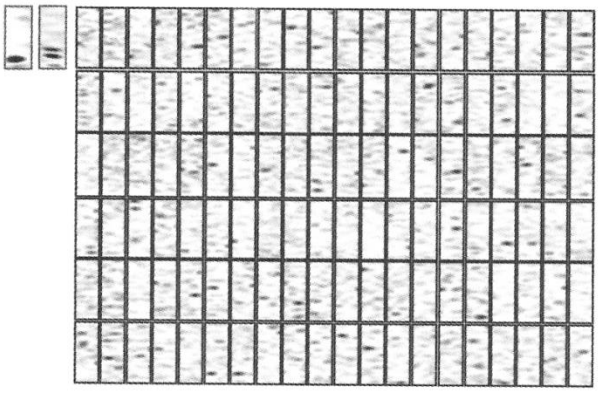

Fig. 3. Algunos campos receptivos espectro-temporales obtenidos a partir de patrones de las representaciones auditivas tempranas de 64 × 4 puntos. Las muestras de habla fueron tomadas de cinco fonemas del corpus de TIMIT, región DR1. Se muestran también, a modo de comparación, dos STRFs estimados a partir de datos reales de las neuronas de la corteza auditiva primaria en mamíferos (los casos más similares se hallan resaltados) [22]. Cada STRF tiene una altura de 4 KHz y un ancho de 32 mseg.

obtenidos al utilizar la representación auditiva temprana en forma directa. Para esta última representación, los resultados globales son aparentemente buenos; sin embargo, cuando se examinan las tasas de reconocimiento individuales para cada fonema (que aparecen en las columnas de la derecha de la tabla), sólo dos o tres fonemas están de hecho bien clasificados (ver experimentos Nº 1-8). Este problema aparece por una solución correspondiente a un mínimo local que la representación cortical logra evitar (ver la dispar distribución de los patrones en la Tabla I).

Más aún, los resultados para la representación cortical son mejores que aquéllos obtenidos utilizando los MFCC, que es la representación clásicamente utilizada para esta tarea en los sistemas del estado del arte (ver experimentos Nº 16 y 25 en la tabla). Otro aspecto importante es que, para las representaciones corticales, el desempeño del clasificador resulta satisfactorio inclusive para redes de arquitecturas relativamente pequeñas en relación con las dimensiones de los patrones. Este aspecto corrobora la hipótesis de que las clases son mejor separadas en este nuevo espacio de mayor cantidad de dimensiones, y de esta forma un clasificador más sencillo puede completar la tarea exitosamente.

Para evaluar la significancia estadística de estos resultados se consideró la probabilidad de que el error de clasificación para un clasificador dado e sea más pequeño que el error de otro tomado como referencia e_{ref}. Para realizar esta estimación, se supone la independencia estadística de los errores para cada tramo, y la distribución binomial de los errores se modela por medio de una distribución gaussiana (esto es posible porque se cuenta con un número suficientemente grande de tramos de prueba). De esta forma, si se compara el mejor resultado de la representación cortical con el de los MFCCs, se obtiene que $Pr(e_{ref} > e) > 92\%$.

Como una primera aproximación al análisis de la robustez de la representación lograda se realizaron una serie de experimentos en los cuales se agregó ruido blanco a los tramos de la señal de voz con diferentes relaciones señal/ruido (SNR). A continuación se evaluó el desempeño de la red entrenada con la versión sin ruido de los patrones a partir de la versión contaminada de los mismos. El resultado puede observarse en la Figura 4. Puede apreciarse que la representación cortical resulta más robusta para relaciones señal/ruido de entre ∞ y 10 dB.

Tabla II. Tasas de reconocimiento de tramos de fonemas para los experimentos de clasificación con MLP utilizando las representaciones obtenidas a partir de modelos auditivos tempranos, representaciones corticales obtenidas de la activación de las STRFs, y MFCCs (se han resaltado los mejores resultados).

Nº	Experimento		Red	TRN	TST	/b/	/d/	/jh/	/eh/	/ih/
1	Auditiva	64x4	256/4/5	45.84	44.76	0.00	0.00	6.90	100.00	6.27
2			256/8/5	44.35	43.25	0.00	0.00	4.31	100.00	3.13
3			256/16/5	64.28	65.03	0.00	0.00	9.48	94.99	57.59
4			256/32/5	68.92	69.67	0.00	0.00	100.00	95.87	54.82
5			256/64/5	70.70	72.69	0.00	0.00	83.62	72.34	86.75
6			256/128/5	70.50	72.17	4.55	0.00	62.93	84.73	76.14
7			256/256/5	72.15	73.74	0.00	0.00	97.41	85.23	74.82
8			256/512/5	69.21	71.76	0.00	0.00	100.00	94.49	60.96
9	Cortical	64x4	256/4/5	77.04	75.72	40.91	56.48	97.41	84.86	69.16
10			256/8/5	79.64	**77.64**	46.97	62.96	93.97	84.86	72.77
11			256/16/5	75.60	76.08	65.15	51.85	97.41	89.99	63.73
12			256/32/5	79.72	74.73	65.15	67.59	98.28	79.22	68.80
13			256/64/5	87.27	76.86	74.24	66.67	95.69	88.24	64.82
14			256/128/5	100.00	**78.37**	72.73	70.37	96.55	78.35	77.35
15			256/256/5	98.10	77.07	65.15	71.30	91.38	87.11	67.11
16			256/512/5	99.92	**79.16**	71.21	69.44	92.24	80.35	78.07
17	Cortical	64x4x2	512/4/5	78.65	73.79	48.48	59.26	86.21	85.61	64.58
18			512/8/5	80.62	75.51	63.64	59.26	98.28	85.36	65.90
19			512/16/5	78.65	74.26	54.55	53.70	99.14	82.98	66.63
20			512/32/5	82.58	75.66	62.12	66.67	95.69	85.11	66.02
21			512/64/5	87.27	75.87	54.55	65.74	98.28	83.48	68.43
22			512/128/5	84.72	75.98	65.15	56.48	95.69	84.23	68.67
23			512/256/5	81.37	76.55	65.15	62.96	95.69	86.86	66.63
24			512/512/5	82.64	76.32	65.15	61.11	97.41	77.97	74.70
25	MFCC+E	14+14	28/28/5	77.39	77.28	46.51	75.38	91.11	80.56	74.40

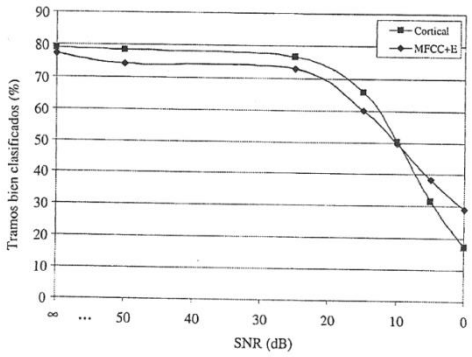

Fig. 4. Tasa de reconocimiento de tramos de fonemas bien clasificados por el MLP entrenado con diferentes representaciones (corticales y MFCCs) en función de la relación señal/ruido. La arquitectura empleada es la que obtuvo mejor desempeño en los experimentos con señales sin ruido (256/512/5).

Harpur [9] realizó algunos experimentos sencillos de clasificación de fonemas sin ruido utilizando códigos de baja entropía, con coeficientes sólo positivos generados a partir de un banco de filtros. Sin embargo, no se han reportado hasta ahora en la literatura experimentos utilizando modelos más complejos de los campos receptivos auditivos, como los mostrados en este trabajo.

5. Conclusiones

En este trabajo se ha propuesto una nueva aproximación para la extracción de características de la señal de voz, basada en una inspiración biológica, y ésta se ha aplicado exitosamente en una tarea de clasificación de fonemas. La aproximación empleada encuentra primero una representación auditiva temprana de la señal de voz a nivel del nervio auditivo. Luego, a partir de estos espectrogramas auditivos del habla, se estima un diccionario óptimo para describirlos utilizando analogías bien establecidas con el comportamiento estadístico de los sistemas neurosensoriales biológicos.

El método encuentra un conjunto de átomos que pueden ser relacionados con los campos receptivos espectro-temporales de la corteza auditiva, y que probaron ser capaces de funcionar como detectores de características fonéticas importantes. Es interesante mencionar, por ejemplo, la detección de eventos basados en característi-

cas espectro-temporales altamente localizadas, así como segmentos relativamente estacionarios, diferentes tipos de evolución de formantes y zonas no armónicas.

Se entrenó un perceptrón multicapa como clasificador de fonemas, utilizando como patrones de entrada los de la representación provista por este método. Los resultados obtenidos mejoran tanto los de la representación auditiva temprana como los de los MFCCs tradicionalmente utilizados. Otro aspecto importante de destacar es la robustez de este tipo de representaciones en la presencia de ruido blanco auditivo, sin ningún tipo de procesamiento adicional. Aunque claramente se requiere más experimentación aún, el objetivo ha sido demostrar la viabilidad del método propuesto.

Entre las líneas a explorar en el futuro se pueden mencionar: la estimación de STRF con más datos y para distintas duraciones de las ventanas (por ejemplo, a nivel de sílabas), un análisis más exhaustivo de las pistas acústicas detectadas, experimentos con otros tipos de clasificadores, diferentes medidas de la calidad de las representaciones obtenidas, métodos alternativos para estimar los diccionarios y las activaciones, y diferentes maneras de aprovechar la robustez de la representación.

6. Agradecimientos

Los autores desean agradecer a: la *Universidad Nacional del Litoral* (UNL-CAID 012-72), la *Agencia Nacional de Promoción Científica y Tecnológica* (ANPCyT-PICT 12700 & 25984) y el *Consejo Nacional de Investigaciones Científicas y Técnicas* (CONICET) de Argentina y al *Consejo Nacional de Ciencia y Tecnología* (CONACYT) y la *Secretaría de Educación Pública* (SEP) de México, por el apoyo a la investigación que permitió realizar este trabajo.

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] Barlow, H., *Network: Computation in Neural Systems*, 12, 241 (2001).

[3] deCharms, R.C., Blake, D.T. and Merzenich, M.M., *Science*, 280, 1439 (1998).

[4] Delgutte, B., Physiological models for basic auditory percepts. In H.H. Hawkins, T.A. McMullen, A.N Popper, and R.R. Fay (Eds.) *Auditory Computation*. Springer, New York, 1996.

[5] Deller, J., Proakis, J. and Hansen, J., *Discrete Time Processing of Speech Signals*. Macmillan Publishing, New York, 1993.

[6] Donoho, D.L., and Elad, M., *Proc.Natl. Acad. Sci.* 100, 2197 (2003).

[7] DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus Documentation. Technical Report, National Institute of Standards and Technology, February 1993.

[8] 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.

[9] Harpur, G.F, *Low Entropy Coding with Unsupervised Neural Networks*. PhD thesis, Department of Engineering, University of Cambridge, Queens' College, February 1997.

[10] Hyvärine, A., Sparse code shrinkage: Denoising of nongaussian data by maximum-likelihood estimation. Technical Report, Helsinki University of Technology, 1998.

[11] Klein, D.J., Konig, P. and Kording, K.P., *J. Appl. Signal Process.* 2003, 659 (2003).

[12] Kording, K.P., Konig, P. and Klein, D.J., 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.

[13] Oh-Wook Kwon and Te-Won Lee, *Signal Processing*, 84, 1005 (2004).

[14] Lewicki, M.S., and Olshausen, B.A., *J. Opt. Soc. Amer.* 16, 158 (1999).

[15] Lewicki, M.S. and Sejnowski, T.J., Learing overcomplete representations. In *Advances in Neural Information Processing 10 (Proc. NIPS'97)*, pages 556-562. MIT Press, 1998.

[16] Oja, E. and Hyvarinen, A. Independent Component Analysis: A Tutorial. *Helsinki University of Technology, Helsinki*, 2004.

[17] Olshausen, B.A., Sparse codes and spikes. In R. P. N. Rao, B. A. Olshausen, and M. S. Lewicki (Eds.), *Probabilistic Models of the Brain: Perception and Neural Function*, chapter 13. MIT Press, 2001.

[18] Olshausen, B.A. and Field, D.J., *Nature*, 381, 607 (1996).

[19] Rufiner, H.L., Goddard, J., Rocha, L.F. and Torres, M.E., *Physica A* 367, 231 (2006).

[20] Simon, J.Z., Didier, A., Depireux, A., Klein, D.J., Fritz, J.B. and Shamma, S.A., *Neur. Comput.* 19, 583 (2007).

[21] Theunissen, F.E., Sen, K. and Doupe, A.J., *J. Neurosc.* 20, 2315 (2000).

[22] Yang, X., Wang, K. and Shamma, S.A., *IEEE Trans. Inf. Theory*, 38, 824 (1992).

Manuscrito recibido el 10 de agosto de 2007.
Aceptado el 21 de setiembre de 2007.