Artículo de Investigación. Revista Ciencia e Ingeniería. Vol. 35, No. 1, pp. 29-38, diciembre-marzo, 2014.
ISSN 1316-7081. ISSN Elect. 2244-8780 Universidad de los Andes (ULA)

Reconocimiento automático de fonemas en habla continúa venezolana por medio de sistemas híbridos basados en modelos ocultos de Márkov y redes neuronales artificiales

Automatic phoneme recognition in Venezuelan continuous speech based on hidden Markov models and artificial neural networks hybrid systems

Jabbour, Georges *[1] y Maldonado, José Luciano [2]
[1] Departamento de Investigación de Operaciones,
Facultad de Ingeniería, Universidad de Los Andes, Mérida
[2] Instituto de Estadística Aplicada y Computación, IEAC,
Facultad de Ciencias Económicas y Sociales, Universidad de Los Andes, Mérida
*jabbour@ula.ve

Resumen

En esta investigación se proponen y evalúan dos enfoques híbridos basados en modelos ocultos de Márkov (MOM) y redes neuronales artificiales (RNA) para el reconocimiento automático del habla. El desempeño de estos enfoques híbridos se compara con el de un reconocedor basado sólo en MOM (el reconocedor MOM). En el primero de los enfoques híbridos, una RNA cumple el papel de estimador de las probabilidades de las observaciones para los MOM, mientras que en el segundo enfoque se emplea una RNA como reconocedor de la señal de voz, en base a las probabilidades producidas por los MOM. Los tres reconocedores fueron programados a través de Matlab® y se entrenaron con señales de habla continua venezolana, pertenecientes a una base de datos que forma parte del proyecto europeo SpeechDat. La unidad de entrenamiento acústico que se utilizó fue el fonema. Los resultados obtenidos indican que mediante el primer enfoque híbrido, utilizando redes perceptrónicas multicapa, se logra un reconocimiento mejor que el del reconocedor MOM en un 2,3%, mientras que con el segundo enfoque híbrido, utilizando redes de funciones de base radial, se logra una mejoría del 4,7%.

Palabras clave: Reconocimiento automático del habla, modelos ocultos de Márkov, redes neuronales artificiales, modelo híbrido RNA/MOM.

Resumen

In this paper we propose and test two hybrid approaches based on hidden Markov models (HMMs) and artificial neural networks (ANNs) for automatic speech recognition. The performance of these hybrid approaches is compared to the performance of a recognizer based on HMMs only (the HMM recognizer). In the first hybrid approach, the probabilities of the observations for the HMMs are estimated through an ANN. In the second hybrid approach, we use an ANN as a recognizer of the speech signals, the entries of which are the probabilities produced by the HMMs. The three recognizers were programmed using Matlab® and trained with Venezuelan continuous speech signals, the speech signals of which form part of the SpeechDat European project. The phoneme was chosen as the acoustic training unit. The results obtained shows that the first hybrid approach, based on a feed-forward neural network, yields a better performance than that obtained with the HMM recognizer in 2.3%; whereas, with the second hybrid approach, based on a radial basis network, the results were 4.7% better than those of the HMM recognizer.

Key-words: Automatic speech recognition, hidden Márkov models, artificial neural networks, hybrid ANN/HMM Model.

1 Introducción

El reconocimiento automático del habla (RAH) ha evolucionado constante y progresivamente en las últimas cinco décadas. Sin embargo, los sistemas actuales de RAH aún no pueden competir con las capacidades del ser humano, a pesar de proporcionar resultados satisfactorios en muchos ca-

sos (Jurafsky y Martin, 2006; Maldonado, 2003). Esta realidad establece las bases para el desarrollo de nue-vas líneas de investigación que persiguen la obtención de sistemas de reconocimiento del habla cada vez más eficaces y robustos.

Los adelantos más significativos en el área del RAH han sido aportados por los modelos ocultos de Márkov (MOM). De hecho, hasta el momento sigue siendo la herramienta más utilizada, puesto que en el modelado y reconocimiento automático de voz aún no se ha encontrado otro método que supere sus resultados (Juang y Rabiner, 2005; Maldonado, 2003). Más aún, han surgido diversas propuestas que combinan los MOM con otras técnicas, como los enfoques híbridos entre los MOM, las redes neuronales artificiales (RNA) y las máquinas de vectores de soporte (MVS) (Liu y col., 2007; Gholampour y Nayebi, 1999), lo que constituye una nueva corriente de investigación en el RAH, cuyo propósito es mejorar los resultados de los MOM puros. A esta corriente pertenece la presente investigación

2 Justificación del uso de las redes neuronales artificiales en el reconocimiento automático del habla

Las RNA constituyen una poderosa herramienta que permite resolver una gran variedad de problemas que tratan, entre otros, del reconocimiento de patrones, de la predición, de la estimación y de la optimización (Bourlard y Morgan, 1993); entonces, siendo el RAH una tarea de clasi-ficación de patrones, resulta interesante explorar la integra-ción de dicha herramienta al esquema de modelos ocultos de Márkov, con el fin de mejorar la capacidad de este tipo de reconocedores.

3 Los datos y su preprocesamiento

3.1 Base de datos utilizada

En esta investigación se utilizaron señales de voz del español hablado en Venezuela obtenidas por telefonía fija, pertenecientes a la base de datos SpeechDat Venezolana (Maldonado, 2003). Esta es una base de datos construida por la Universidad Politécnica de Cataluña de España, con la participación de la Universidad de Los Andes de Venezuela. Dicha base de datos forma parte del proyecto europeo SpeechDat (Moreno y Mora, 1999). Se trabajó, específicamente, con señales de voz correspondientes a pronunciaciones de fechas.

En la base de datos SpeechDat Venezolana se tienen pro-nunciaciones de tres tipos de fechas por cada locutor:

1. Fecha espontánea. Por ejemplo, la fecha de nacimiento.

2. Fecha leída. Por ejemplo, “jueves tres de octubre de mil novecientos noventa y dos”.

3. Fecha relativa. Por ejemplo, “dentro de treinta días”.

Se seleccionaron 250 archivos de voz de hablantes de ambos sexos, de las diferentes zonas dialectales de Venezuela, como se muestra en la tabla 1. La escogencia de estas zonas (regiones) se realizó en base a la clasificación dialectal que se presenta en Maldonado (2003).

Tabla 1. Distribución de hablantes por región y género

Región	Masculinos	Femeninos	Total	(%)
Central	46	65	111	44.4
Zuliana	20	19	39	15.6
Llanos	7	8	15	6
Oriental	8	37	45	18
Los Andes	18	22	40	16
Total	99	151	250	100.0

En la tabla 2 se muestra la distribución de los archivos u-tilizados de acuerdo al tipo de fecha

Tabla 2. Tipos de pronunciaciones

Tipo de fecha	Cantidad	%
Espontánea	98	39.2
Leída	118	47.2
Relativa	34	13.6
Total	250	100.0

Se escogieron 200 señales (80%) como conjunto de entrenamiento y 50 (20%) como conjunto de validación, con réplicas en los dos conjuntos. Las características de los conjuntos de entrenamiento y validación se muestran en las tablas 3 y 4, respectivamente.

Tabla 3. Conjunto de entrenamiento

Región	Masculinos	Femeninos	Total	%
Central	32	37	69	34.5
Zuliana	20	18	38	19.0
Llanos	13	11	24	12.0
Oriental	15	20	35	17.5
Los Andes	17	17	34	17.0
Total	97	103	200	100.0

Tabla 4. Conjunto de validación

Región	Masculinos	Femeninos	Total	%
Central	7	6	13	26.0
Zuliana	7	5	12	24.0
Llanos	3	3	6	12.0
Oriental	4	5	9	18.0
Los Andes	6	4	10	20.0
Total	27	23	50	100.0

3.2 Transcripción ortográfica y fonética

Esta etapa consistió en escuchar los archivos de audio y realizar su transcripción ortográfica y fonética. La transcripción fonética se realizó con la notación Speech Assessment Methods Phonetic Alphabet, SAMPA (Gibbon y col., 1997; Wells, 1997). Específicamente, se utilizó el conjunto

de símbolos de la versión SAMPA, propuesta en Maldonado (2003), para el español venezolano.

En los archivos de las señales seleccionadas se encontró el siguiente conjunto de 26 sonidos, identificados tal como se hace en Maldonado (2003), al que se llamó Conjunto de Fonemas de Fechas Venezolanas: a, b, B, c, d, D, e, f, g, G, h, i, j, k, l, m, n, N, o, r, s, t, u, w, y y sil. El símbolo sil se utilizó para representar las zonas de silencio presentes en las pronunciaciones.

3.3 Etiquetado y segmentación

Se asoció manualmente cada símbolo de la transcripción fonética a un segmento de señal, a través de COLEA (Loizou, 2008). A partir del etiquetado y la segmentación se obtuvieron 9130 señales correspondientes a los fonemas mencionados, las cuales fueron analizadas para obtener las respectivas secuencias de vectores de parámetros cepstrales (Maldonado, 2003; Huang y col., 2001; Rabiner, 1989). Se decidió no considerar los fonemas f, G, g, e y, debido a que no se encontró un número considerable de realizaciones de los mismos (al menos 50), por lo que se descartaron las señales donde dichos fonemas estaban presentes, es decir, se trabajó con las realizaciones de los 22 fonemas restantes.

4 Diseño de los reconocedores

En esta investigación se consideran tres reconocedores de secuencias de fonemas de habla continua, independientes del hablante y constituidos por los modelos acústicos de 22 fonemas del español venezolano asociado a pronunciaciones de fechas: el reconocedor MOM, el reconocedor híbrido RNA1/MOM y el reconocedor híbrido RNA2/MOM.

4.1 El reconocedor MOM

El reconocedor MOM clásico, o reconocedor acústico clásico, consiste en el uso de un MOM puro para modelar cada fonema. Estos modelos fueron entrenados y validados a través de los algoritmos Baum-Welch y Viterbi (Huang y col., 2001). Una descripción más detallada de este enfoque se puede encontrar en Rabiner (1989).

4.2 Arquitecturas híbridas RNA/MOM

Los trabajos de investigación que combinan redes neuronales Artificiales y Modelos Ocultos de Márkov para RAH han producido, en general, dos tipos de arquitecturas (Milone, 2005): aquellas en las que las RNA se emplean como estimadores de las probabilidades de las observaciones en los MOM de observaciones continuas, en lugar de las mezclas de gaussianas y otras en las que las RNA se emplean como cuantificadores de los vectores de observaciones en MOM de observaciones discretas.

En este trabajo se presenta un enfoque de reconocimiento híbrido enmarcado dentro del primer tipo de arquitectura y otro enfoque en el que las RNA poseen una tarea diferente a las antes mencionadas, según se describe posteriormente.

4.3 El Reconocedor híbrido RNA1/MOM

Este reconocedor representa un enfoque en el cual las probabilidades de emisión de las observaciones son originadas por la combinación de las probabilidades producidas por las mezclas de gaussianas, particulares de cada estado, con las probabilidades producidas por una red neuronal artificial, comunes a todos los estados de todos los MOM.

Como cada MOM clásico es entrenado de manera independiente del resto, las mezclas de gaussianas no utilizan información sobre las observaciones asociadas al resto de los estados de los otros MOM, lo cual compromete la capacidad discriminativa de este tipo de reconocedores. Con la inclusión de una RNA entrenada para estimar las probabilidades de emisión de todos los estados y de todos los MOM, se busca mejorar tal capacidad discriminativa.

Por otro lado, en los modelos híbridos clásicos las mezclas de gaussianas son sustituidas por una RNA, mientras que en esta investigación se propuso un modelo que no sustituye completamente las probabilidades producidas por la mezcla de gaussianas. El modelo propuesto es similar al modelo de fusión de Liu y col., (2007), para un reconocedor basado en modelos ocultos de Márkov y máquinas de vectores soporte, con la diferencia de que en este caso se utiliza una RNA.

En la figura 1 se muestra la estructura general del reconocedor híbrido RNA1/MOM propuesto.

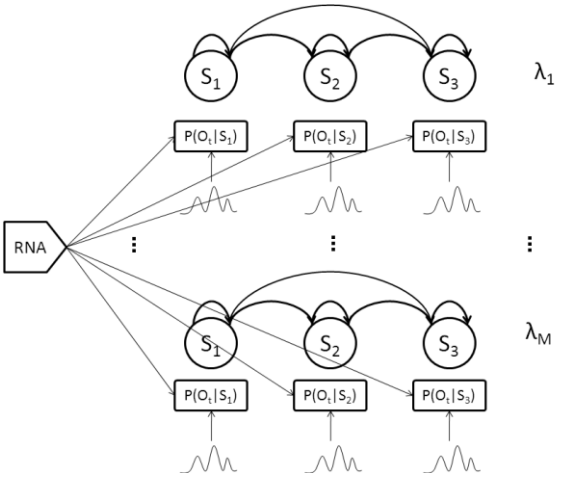

Fig. 1. El reconocedor híbrido RNA1/MOM

La figura 1 está basada en el caso particular de un reconocedor que consta de una RNA y *M* modelos ocultos de Márkov (λ_i; i=1,…,*M*) de 3 estados cada uno. Cada salida de la RNA está asociada con un estado de un MOM.

Bajo este enfoque, las probabilidades de emisión de las observaciones están determinadas por un componente discriminativo intra-estado (las mezclas de gaussianas) y por

un componente discriminativo inter-estado (la RNA).

Para evaluar la probabilidad $P(\mathbf{O}\lambda_i)$, se sustituye $b_j(\mathbf{O}_t)$ del MOM clásico por la ecuación:

$$bj(Ot) = \alpha PGMM(Ot|qj) + (1-\alpha)PRNA(qj|Ot) \quad (1)$$

donde q_j es el j-ésimo estado del modelo, $b_j(\mathbf{O}_t)$ es la probabilidad de emisión de símbolo para el dicho estado, **O** es la secuencia de observaciones, $\mathbf{O}_t$ es el vector de características u observaciones en el instante de tiempo *t*, $\alpha \epsilon$ [0, 1], $P_{GMM}(\mathbf{O}_t|q_j)$ y $P_{RNA}(q_j|\mathbf{O}_t)$ son la probabilidades arrojadas por la mezcla de gaussianas y la RNA, respectivamente.

La arquitectura del reconocedor híbrido RNA1/MOM es similar a la de un reconocedor acústico clásico, donde existe un MOM por cada unidad lingüística básica, en este caso, los fonemas; y donde para reconocer un fonema desconocido de entrada, se calcula la probabilidad de cada uno de los MOM de representar a tal fonema y se selecciona aquel con mayor probabilidad. La diferencia, en este caso, está en que además de los MOM, el reconocedor híbrido propuesto posee una RNA, cuyas salidas son utilizadas al mismo nivel que las probabilidades generadas por las mezclas de gaussianas.

Las entradas de la RNA son el conjunto de parámetros que constituyen una observación, mientras que el número de salidas es igual a la cantidad total de estados de todos los MOM del reconocedor, es decir, si *M* es el número total de MOM del reconocedor y *Q* el número de estados de cada MOM, la RNA tiene *M*x*Q* salidas.

4.4 Entrenamiento del reconocedor híbrido RNA1/MOM

El entrenamiento de este reconocedor se realiza en dos fases:

1. Entrenamiento de los MOM por cada fonema: Esta etapa es equivalente a la construcción de un reconocedor acústico clásico en el que cada MOM es entrenado para maximizar la probabilidad de modelado de la señal de un determinado fonema.

2. Entrenamiento de la RNA: Se entrena la RNA para estimar la probabilidad de ocurrencia de cada estado y de cada fonema, dada una secuencia de observaciones **O**.

Para el entrenamiento de la RNA se utiliza un conjunto de vectores de observaciones asociados a cada uno de los estados de cada MOM. Recuérdese que cada salida está asociada a uno de estos estados.

Para el entrenamiento de dicha RNA, se realizó previamente una decodificación Viterbi a cada secuencia de observaciones del corpus de entrenamiento de cada uno de los MOM clásicos (ver figura 2). De esta manera, se asoció un conjunto de vectores de observación a cada estado de los MOM.

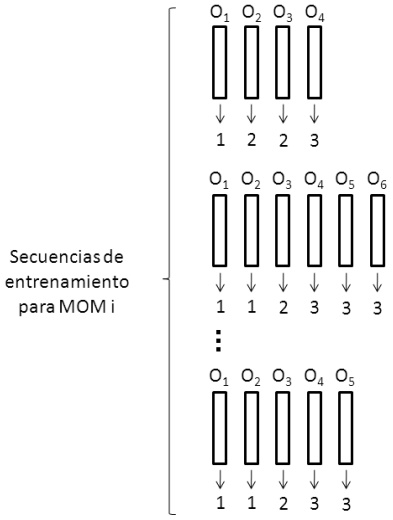

Fig. 2. Decodificación Viterbi de las secuencias asociadas al *i* ésimo MOM

El procedimiento de asociación consistió en fijar una etiqueta distinta a cada estado de los MOM, como se muestra en la figura 3.

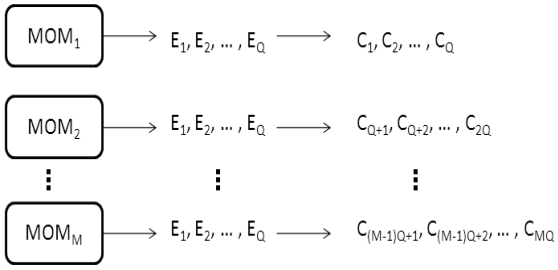

Fig. 3. Transformación de las etiquetas arrojadas por la decodificación Viterbi a clases globales

Se puede observar que los *Q* estados del MOM 1 se eti-quetan como C_1, C_2…, C_Q y C_{Q+1}, C_{Q+2}, …, C_{2Q} para el MOM 2; y así sucesivamente, de manera que la etiqueta *Q* del *M*-ésimo MOM corresponde a la clase *MQ* (ver figura 3).

Concretamente, cada MOM produce etiquetas E_1, E_2 ..., EQ, asociadas a sus estados *1,2,..,Q*. Se transforman esas etiquetas a MQ clases (C_1, C_2,..,C_{MQ}), de manera tal que las observaciones etiquetadas por un MOM como E_i, no se confundan con las observaciones etiquetadas como E_i, de ningún otro MOM. Por ejemplo, si se desea construir un reconocedor de voz para diferenciar dos palabras: "Sí" y "No", se tendría un MOM por cada palabra (*M*=2), y supóngase que cada MOM es de 3 estados (*Q* =3). Las 3 etiquetas asociadas al "Sí" (E_1, estado 1; E_2, estado 2; E_3, estado 3,) corresponden a las primeras 3 clases (C_1, C_2 *y* C_3), y las 3 etiquetas asociadas al "No" corresponderían a las clases C_4, C_5 *y* C_6.. De esta manera, cada vector correspondiente a una observación tendría una clase asociada que lo identifica de manera única con un estado de alguno de los

MOM.

A partir de estos patrones, la RNA es entrenada de manera que su *i*-ésima salida tienda a 1 para los patrones de entrenamiento de la clase *i* y tiendan a 0 el resto de sus salidas.

4.5 El reconocedor híbrido RNA2/MOM

Este segundo enfoque híbrido consiste en una interconexión entre las salidas de un reconocedor acústico clásico con una red neuronal artificial, cuya tarea es realizar una clasificación no lineal de la señal, en base a las probabilidades o puntuaciones producidas por los MOM. Este enfoque está basado en un modelo denominado modelo híbrido en cascada, desarrollado por Gholampour y Nayebi (1999), quienes utilizaron una red perceptrónica de dos capas, en conjunto con un reconocedor basado en MOM, para construir un reconocedor de dígitos aislados del idioma Farsi. En la figura 4 se muestra la estructura de este reconocedor.

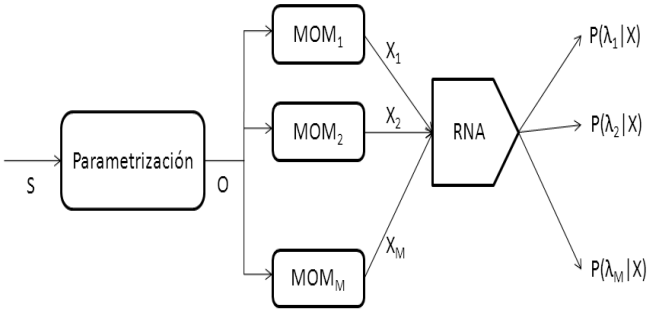

Fig. 4. El reconocedor híbrido RNA2/MOM

En este esquema, para una secuencia de observaciones **O** proveniente de una alocución desconocida S, se calculan las probabilidades $P(\mathbf{O}|\lambda_i)$ de cada uno de los MOM. Estas probabilidades se convierten en las entradas de la RNA, $\mathbf{X} = \{ X_1, X_2, ..., X_M \}$, la cual se encarga de determinar el modelo que mejor representa la secuencia de observaciones.

Bajo este enfoque, los MOM capturan las variaciones temporales de los sonidos y realizan una primera clasificación basada en el criterio de máxima verosimilitud. La RNA se encarga de realizar una segunda clasificación basándose en las probabilidades obtenidas en la etapa anterior, siguiendo algún criterio discriminativo.

La arquitectura del reconocedor híbrido RNA2/MOM comprende una estructura conexionista que utiliza en un caso una red perceptrónica multicapa y en otro caso una red de funciones de base radial, igual que en el primer enfoque híbrido presentado. La cantidad de entradas de la RNA es igual al número de MOM de fonemas, o equivalentemente, igual al número de fonemas.

La entrada X_i a la red corresponde a la probabilidad $P(\mathbf{O}|\lambda_i)$, es decir, a la salida arrojada por el *i*-ésimo MOM para la secuencia de observaciones **O**.

La RNA posee, también, tantas salidas como fonemas, y cada salida es interpretada como la probabilidad a posteriori de cada modelo dada la entrada **X**. Como salida del modelo híbrido se selecciona aquel modelo de fonema cuya proba-bilidad a posteriori sea mayor.

4.6 Entrenamiento del reconocedor híbrido RNA2/MOM

El entrenamiento de este reconocedor, igual que en el primer esquema híbrido presentado, consta de dos fases:

1. Entrenamiento de los MOM por cada fonema: Esta etapa es equivalente a la construcción de un reconocedor acústico clásico basado en MOM.

2. Entrenamiento de la red neuronal artificial: En esta etapa se entrena la red para la clasificación de las probabilidades o puntuaciones producidas por los MOM.

Para construir los patrones de entrenamiento de la RNA, cada patrón del corpus de entrenamiento de los MOM es evaluado en todos los modelos de fonemas y el conjunto de probabilidades producidas por éstos se convierten en las entradas de la red. Las salidas corresponden al fonema correcto en cada caso. La red es entrenada de manera tal que su i-ésima salida tiende a 1 para cualquier entrada correspondiente a una pronunciación del fonema i y 0 para el resto de las salidas.

4.7 Medidas de desempeño de los reconocedores

Para evaluar el desempeño de los reconocedores se definieron las siguientes medidas:

1. Porcentaje de reconocimiento global (%RG): Es el criterio básico y más comúnmente utilizado, que corresponde a la razón del número de realizaciones de fonemas identificados correctamente y el número total de realizaciones de todos los fonemas.

2. Porcentaje de reconocimiento promedio (%RP): Esta medida es el promedio aritmético de los porcentajes de re-conocimiento individuales para los fonemas.

3. Varianza (VAR): La varianza de los porcentajes de re-conocimiento individuales para los fonemas; aquellos que fueron considerados en el cálculo del %RP. Mientras menor sea la varianza, se considera mejor el reconocedor, ya que esto indica mayor homogeneidad en los resultados.

5 Resultados

5.1 Reconocedor MOM clásico

El entrenamiento de los MOM se realizó, en unos casos, con el algoritmo de Baum-Welch, y en otros casos, con el algoritmo de Viterbi (Maldonado, 2003; Huang y col., 2001; Rabiner, 1989). Se trabajó tanto con MOM tipo Bakis como ergódicos (Huang y col., 2001). En el primer caso, se utilizó el toolbox HMM de MATLAB (Loizou, 2008), mientras que en el segundo caso, las herramientas computacionales fueron implementadas como parte de este estudio.

En las tablas 5 y 6 se presentan los resultados de reconocimiento para el caso particular de MOM de 2 estados y diferente número de gaussianas (G) por estado. En estas tablas se muestran sólo los mejores resultados obtenidos según las medidas de desempeño antes definidas.

Tabla 5. Resultados de los reconocedores clásicos con MOM entrenados con Viterbi (2 estados)

	MOM Bakis			MOM ergódicos		
G	%RG	%RP	VAR	%RG	%RP	VAR
5	45.9	42.6	372.7	45.7	41.4	344.6
6	45.3	41.3	363.3	45.5	42.6	343.0
7	45.0	41.0	474.8	44.4	41.9	413.8
8	47.4	43.9	350.4	45.1	43.2	375.1
9	46.9	43.1	377.1	46.0	42.5	372.4
10	45.0	41.4	329.2	45.3	42.1	451.4
11	46.3	42.9	384.2	46.8	42.7	398.1
12	45.5	42.4	410.3	46.4	43.3	395.4
13	46.7	43.4	403.3	45.8	41.3	361.1
14	46.3	43.7	398.9	45.2	42.3	464.3
15	45.6	42.8	419.6	45.6	42.6	426.8

Tabla 6. Resultados de los reconocedores clásicos con MOM entrenados con Baum-Welch (2 estados)

	MOM Bakis			MOM ergódicos		
G	%RG	%RP	VAR	%RG	%RP	VAR
5	47.6	43.4	472.1	47.8	41.9	491.8
6	47.3	41.9	455.7	47.3	41.3	501.3
7	49.8	43.5	533.5	49.1	41.3	599.8
8	48.2	41.8	528.5	46.2	38.5	615.7
9	49.1	42.8	581.1	48.3	40.2	670.0
10	48.6	40.8	543.3	48.3	40.3	541.4
11	48.9	39.9	522.2	47.8	38.2	592.9
12	47.0	38.7	612.5	45.9	35.9	625.3
13	48.1	38.0	561.6	47.5	36.6	726.5
14	47.0	35.8	604.6	47.9	36.7	749.3
15	47.1	37.2	680.7	47.3	35.6	681.4

Como se puede observar, con el entrenamiento Baum-Welch se obtuvieron mejores porcentajes de reconocimiento global. Sin embargo, este entrenamiento arroja, en general, resultados menos homogéneos que el entrenamiento Viterbi dado que los %RP son menores y su varianza es más alta.

Con respecto al tipo de MOM, los mejores resultados se obtuvieron utilizando MOM Bakis.

Así, para el reconocedor clásico, se escogió como el mejor MOM aquel entrenado con Viterbi, de 2 estados, 8 gaussianas por estado y tipo Bakis. Se tomó esta decisión debido a que es el reconocedor que posee un porcentaje de reconocimiento global de 47.4% (el segundo más alto), un porcentaje de reconocimiento promedio por fonema de 43.9% (el más alto) y una varianza de 350.4 (la segunda más baja). Además, este modelo se utilizó como el reconocedor MOM de referencia para la comparación posterior con los reconocedores híbridos.

5.2 Pruebas de reconocimiento de los enfoques híbridos

Para estudiar el desempeño de los dos enfoques híbridos, se realizaron pruebas con dos arquitecturas conexionistas: redes perceptrónicas multicapa (RPM) y redes de funciones de base radial (RFBR). Esto originó 4 modelos híbridos:

- **Modelo híbrido 1:** El reconocedor híbrido RNA1/MOM basado en una RPM.
- **Modelo híbrido 2:** El reconocedor híbrido RNA1/MOM basado en una RFBR.
- **Modelo híbrido 3:** El reconocedor híbrido RNA2/MOM basado en una RPM.
- **Modelo híbrido 4:** El reconocedor híbrido RNA2/MOM basado en una RFBR.

Para el entrenamiento de los modelos híbridos se tomó como base el modelo clásico de referencia, por lo que para ambos enfoques híbridos ya se tenía cubierta la primera fase (entrenamiento de los MOM por fonemas), de modo que se procedió con la segunda fase (el entrenamiento de la RNA).

Con base en el teorema de aproximación universal (Haykin, 1999), para las RPM se realizaron pruebas con arquitecturas de dos capas.

Una vez entrenadas las RNA se evalúo el desempeño de los reconocedores híbridos resultantes bajo los mismos cri-terios utilizados en el caso de los reconocedores clásicos.

5.2.1 Primera prueba

Para todos los modelos híbridos se realizó una prueba exploratoria con el fin de determinar el número de neuronas de la primera capa como un parámetro alrededor del cual se encuentran los mejores resultados. El número de neuronas considerado para todos los modelos híbridos en esta prueba fue el siguiente: 50, 100, 200, 300, 400, 500, 600 y 750. Para los modelos híbridos 1 y 2 se consideraron 3 valores de α: 0.25, 0.50 y 0.75.

En cuanto a los tipos de funciones de activación, para las RPM se utilizaron funciones softmax o funciones logísticas múltiples en su capa de salida, mientras que para las RBF se utilizaron funciones gaussianas en la capa de base radial.

El entrenamiento de las RPM se realizó mediante el algoritmo backpropagation utilizando como método de optimización la conjugada del gradiente (Haykin, 1999).

Por otro lado, se realizaron 3 réplicas por cada arquitectura de red, variando los pesos y sesgos iniciales. Sus resultados fueron promediados para representar a la respectiva arquitectura.

Bajo las condiciones antes descritas, los resultados obtenidos fueron los siguientes:

Tanto para el modelo 1 como para el modelo híbrido 2 los mejores resultados se obtuvieron para el valor de $\alpha = 0,75$.

- Para el modelo híbrido 1 los mejores resultados se obtuvieron con RPM de 300 neuronas en la primera capa.

- Para el modelo híbrido 2 los mejores resultados se obtuvieron con RFBR de 400 neuronas de base radial.
- Para el modelo híbrido 3 los mejores resultados se obtuvieron con RPM de 200 neuronas en la primera capa.
- Para el modelo híbrido 4, a partir de 200 neuronas en la capa de base radial, los modelos se comportan de manera muy similar en una banda menor al 1% alrededor del 51% de reconocimiento global.

5.2.2 Segunda prueba

A partir de los resultados anteriores, se realizó una segunda prueba variando el número de neuronas de la primera capa en cada modelo híbrido alrededor del valor con el cual se obtuvieron los mejores resultados en la prueba exploratoria. Para esta prueba también se realizaron 3 réplicas por cada configuración de RNA. Así, los mejores mode-los fueron los siguientes:

- Modelo híbrido 1: El mejor modelo corresponde a 325 neuronas de la réplica 1, el cual posee un porcentaje de reconocimiento global del 49,7% (el más alto de todos), un porcentaje de reconocimiento promedio por fonema de 45,1% (el segundo más alto) y una varianza de 383,83 (la segunda más baja).
- Modelo híbrido 2: El mejor modelo corresponde a 400 neuronas de la réplica 2, el cual posee un porcentaje de reconocimiento global del 48,3% (el más alto de todos), un porcentaje de reconocimiento promedio por fonema de 44,5% (el más alto) y una varianza de 365,9 (una de las más bajas).
- Modelo híbrido 3: El mejor modelo corresponde a 200 neuronas de la réplica 2, el cual posee un porcentaje de reconocimiento global de 48,8% (el segundo más alto), un porcentaje de reconocimiento promedio por fonema de 43,7% (el más alto) y una varianza de 376,8 (la segunda más baja).
- Modelo híbrido 4: En este caso, dos modelos podrían escogerse como los mejores; el correspondiente a 300 neuronas de la réplica 1, el cual posee un porcentaje de reconocimiento global de 52,3% (el mayor de todos), un porcentaje de reconocimiento promedio por fonema de 47,7% (el segundo mejor) y una varianza de 468,9; y el correspondiente a 250 neuronas en la réplica 2, el cual posee un porcentaje de reconocimiento global de 52,1%, un porcentaje de reconocimiento por fonema de 47,8% y una varianza de 466,0. Se realizó una comparación a nivel de fonemas entre estos dos modelos y se encontró que presentaban muchas similitudes. De hecho, el porcentaje de re-conocimiento es el mismo en 12 de los 22 fonemas, y cada uno posee un porcentaje mayor en 5 de los 10 fonemas res-tantes. Por esta razón, y en base al criterio de parsimonia, se decide escoger la red de 250 neuronas.

En la tabla 7 se resumen estos resultados, y se comparan con los obtenidos con el modelo de referencia.

Tabla 7. Mejores modelos

MODELO	%RG	%RP	VAR
Referencia	47.4	43.9	350.4
Híbrido 1, 325 neuronas	49.7	45.1	383.3
Híbrido 2, 400 neuronas	48.3	44.5	365.9
Híbrido 3, 200 neuronas	48.8	43.7	376.8
Híbrido 4, 250 neuronas	52.1	47.8	466.0

En la figura 5 se presentan gráficamente los porcentajes de reconocimiento por fonema para los mejores modelos (H1: mejor modelo híbrido 1, H2: mejor modelo híbrido 2, H3: mejor modelo híbrido 3, H4: mejor modelo híbrido 4 y Ref.: el reconocedor de referencia).

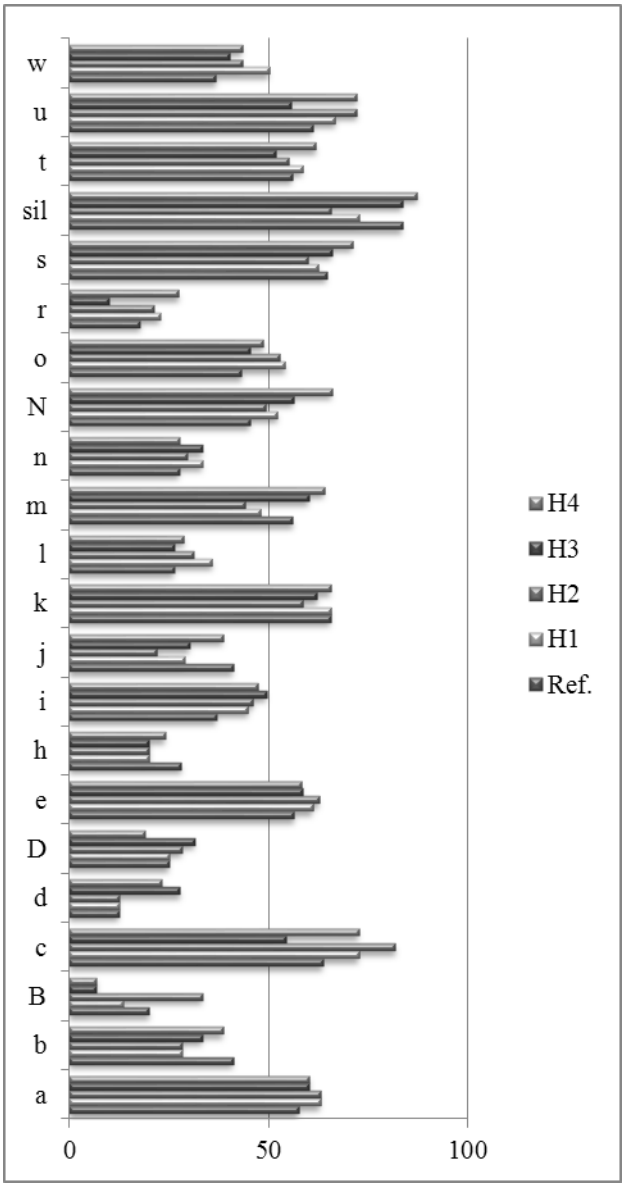

Fig. 5. Mejores modelos. Porcentaje de reconocimiento por fonema

En la figura 5 se observa que:

- Con todos los modelos híbridos se obtienen mayores porcentajes de reconocimiento en las vocales que con el reconocedor de referencia, con la excepción del modelo híbri-

do 3 en el caso de la vocal "u".

- El modelo híbrido 1 tiene el mayor porcentaje absoluto en 3 de los 22 fonemas (*l, o, w*) y tiene igual o mejor desempeño que el modelo de referencia en 15 fonemas.
- El modelo híbrido 2 tiene el mayor porcentaje absoluto en 3 de los 22 fonemas (*B, c, e*) y tiene igual o mejor desempeño que el modelo de referencia en 14 fonemas.
- El modelo híbrido 3 tiene el mayor porcentaje absoluto en 3 de los 22 fonemas (*d, D, i*) y tiene igual o mejor desempeño que el modelo de referencia en 12 fonemas.
- El modelo híbrido 4 tiene el mayor porcentaje absoluto en 6 de los 22 fonemas (*m, N, r, s, sil, t*) y tiene igual o mejor desempeño que el modelo de referencia en 17 de los 22 fo-nemas. Además, sólo en 1 de los 5 casos en que resultó ser peor que el modelo de referencia, la diferencia superó el 10% (el fono *B*).

En comparación con el modelo de referencia, el modelo híbrido 1 logra mejorar en un 2,3% el porcentaje de reconocimiento global y en 1,2% el porcentaje de reconocimiento promedio.

Con el modelo híbrido 2 se logran resultados, desde el punto de vista global, similares a los del modelo de referencia, ya que logra mejorar en menos de 1% los porcentajes de reconocimiento global y promedio de éste, teniendo una varianza ligeramente superior.

El modelo híbrido 3, a pesar de que logra mejorar en 1,4% el porcentaje de reconocimiento global respecto al modelo de referencia, posee un porcentaje de reconocimiento promedio por fonema 0,2% menor que la de éste y una mayor varianza.

Los mejores resultados se obtuvieron con el modelo híbrido 4, el cual logra mejorar en 4,7% el porcentaje de reconocimiento global respecto al modelo de referencia y en 3,9% el porcentaje de reconocimiento promedio a pesar de que su varianza también resultó ser mayor que la de éste. Más aún, fue el único modelo híbrido que superó en promedio al reconocedor de referencia tanto en las vocales como en las consonantes (en 6,26% y 3,22%, respectivamente).

6 Conclusiones

Luego de finalizado el trabajo experimental, se llegó a las siguientes conclusiones.

El algoritmo de entrenamiento Baum-Welch y el algoritmo de entrenamiento Viterbi, para modelos de fonemas del habla venezolana de fechas, arrojaron resultados similares en cuanto al porcentaje de reconocimiento global: el primero ligeramente mejor que el segundo. Sin embargo, los porcentajes de reconocimiento por fonemas para el reconocimiento Viterbi tienden a ser más homogéneos.

Los MOM tipo Bakis producen mejores resultados que los MOM ergódicos.

Para el caso del enfoque híbrido 1, las RPM resultaron tener un mejor desempeño que las RFBR. Con las RPM pudo lograrse hasta un 2,3% de mejora con respecto al sistema clásico de referencia en cuanto a porcentaje de reconocimiento global. Para los modelos basados en el enfoque híbrido 2, se obtienen mejores resultados utilizando RFBR (modelo híbrido 4), logrando mejorar casi en un 5% el porcentaje de reconocimiento global del reconocedor de referencia. A nivel de fonemas, el modelo híbrido 4 resultó ser igual o mejor que el modelo de referencia en 17 de los 22 fonemas.

Con los dos enfoques híbridos estudiados, se logró mejorar ligeramente las capacidades de reconocimiento del modelo de referencia, lo que demuestra la potencialidad de los métodos conexionistas en los sistemas de reconocimiento automático del habla. Sin embargo, el mayor beneficio de los 4 modelos híbridos se produjo en el caso de las vocales, pues sólo el modelo híbrido 4 sobrepasó al de referencia tanto en las vocales como en las consonantes, siendo por lo tanto el mejor de los reconocedores evaluados.

En general, los resultados obtenidos son comparables con los obtenidos por otros investigadores, como Bourlard y Morgan (1993), quienes realizaron pruebas de reconocimiento utilizando RPM para la estimación de las probabilidades de observación de un MOM y obtuvieron porcentajes entre 40% y 50% en reconocimiento de fonemas independientes del hablante. Fernández y Feijóo (2004), quienes desarrollaron un reconocedor fonético de habla continua utilizando información de segmentos adya-centes, obtuvieron un porcentaje de reconocimiento de 47,71%.

Referencias

Bourlard H y Morgan N, 1993, Connectionist Speech Recognition: A Hybrid Approach, Kluwer Academic Publishers, E.E.U.U.

Fernández S y Feijóo S, 2004, Reconocimiento Fonético en Habla Continua Usando Información de Segmentos Adyacentes, IV Congreso Ibero-americano de Acústica, Portugal.

Gholampour I y Nayebi K, 1999, The Cascade HMM/ANN Hybrid: A New Framework for Discriminative Training in Speech Recognition, Proceedings of NSIP¸ pp. 461-465.

Gibbon D, Moore R y Winski R, 1997, Handbook of Standards and Resources for Spoken Language Systems, Mouton de Gruyter, E.E.U.U.

Haykin S, 1999, Neural Networks: A Comprehensive Foundation, Prentice Hall, E.E.U.U.

Huang X, Acero A y Hon H, 2001, Spoken Language Processing. A guide to Theory, Algorithm and System Development, Prentice Hall, E.E.U.U.

Juang B y Rabiner L, 2005, Automatic Speech Recognition - a Brief History of the Technology. Elsevier Encyclopedia of Language and Linguistics. E.E.U.U.

Jurafsky D y Martin J, 2006, Speech And Language Processing, Prentice Hall, E.E.U.U.

Liu J, Wang Z y Xiao X, 2007, A Hybrid SVM/DDBHMM Decision Fusion Modeling for Robust Continuous Digital Speech Recognition, Pattern Recognition Letters, Vol. 28, No. 8, pp. 912–920.

Loizou P, 2008, Colea: A Matlab Software Tool for Speech

Analysis. Se encuentra en: http://www.utdallas.edu/loizou/speech/colea.htm. Fecha de consulta: 08 Junio 2011.
Maldonado J, 2003, Tratamiento y Reconocimiento Automático de Señales de la Voz Venezolana, Tesis Doctoral, Universidad de Los Andes, Mérida, Venezuela.
Milone D, 2005, Reconocimiento Automático del Habla con Redes Neuronales Artificiales, Ciencia, Docencia y Tecnología, Vol. 16, No. 31, pp. 261–322.
Moreno A y Mora E, 1999, Speech Spanish Venezuelan database for fixed telephone network, Universidad Politécnica de Cataluña (España) y Universidad de Los Andes (Venezuela).
Rabiner L, 1989, A Tutorial on Hidden Márkov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, Vol. 77, No. 2, pp. 257–286.
Rabiner L y Juang B, 1993, Fundamentals of Speech Recognition, Prentice Hall, E.E.U.U.
Wells, J, 1997, Sampa Computer Readable Phonetic Alphabet, Handbook of Standards and Resources for Spoken Language Systems, Mouton de Gruyter, E.E.U.U.

Recibido: *10 de mayo de 2012*

Revisado: *20 de julio de 2013*

Jabbour Chediak, Georges*. Venezolano. Ingeniero de Sistemas (ULA, 1998). MSc. en Estadística Aplicada (ULA, 2007). Profesor Agregado ULA (Ingresó: 2002).*

Maldonado, .José Luciano *Venezolano. Ingeniero de Sistemas (ULA, 1987). MSc. en Ingeniería de Sistemas de Control (ULA, 1994), Dr. en Ciencias Aplicadas (ULA, 2003). Profesor Titular ULA (Ingresó: 1993).*