ALGORITMO EBHA PARA EL ENTRENAMIENTO DE REDES NEURONALES

Anibal Cotrina Atencio, Fernando Peralta Reyes
acotrin@universia.edu.pe, f7peralta@hotmail.com

Facultad de Ingeniería Electrónica
Universidad Nacional Mayor de San Marcos

RESUMEN: En el siguiente artículo se propone un algoritmo de entrenamiento para una red neuronal artificial tipo Perceptron Multicapa cuya aplicación estuvo dirigida al reconocimiento de patrones de voz. El algoritmo ha sido denominado Entrenamiento por Bloques de Hablantes (EBHA) y se ha codificado en el estándar ANSI C sobre la versión 6 de Microsoft Visual C++.

ABSTRACT: We propose a training algorithm for an artificial neural network of Multil Layer Perceptron (MLP) type whose applications are adressed to speech recognition patterns. The algorithm has been denoted by Speaker Block Training (EBHA) and it has been coded in the standard ANSI C on the version 6 of Visual Microsoft C++.

Keywords: Señal de Voz, RNA, MLP, pesos, capas, nodos, Backprogation, palabras, Tasa de Acierto, hablantes y patrones.

I. INTRODUCCIÓN

Un sistema de reconocimiento de voz es utilizada para un gran número de aplicaciones, consta de varias etapas que se inicia cuando las palabras son convertidas a señales eléctricas a través de un micrófono y son digitalizadas a fin de ser interpretadas por el procesador del sistema. En la etapa de Clasificación, los algoritmos matemáticos extraen los patrones o parámetros característicos de la señal de voz, que luego tienen que ser identificadas en la etapa de Reconocimiento, tarea que es sumamente compleja y es donde se han optado varias soluciones, una de las cuales esta basada en la utilización de Redes Neuronales Artificiales (RNA). Sin embargo, optar por esta solución implica entrenar a la red de la forma más eficientemente posible. Es así, que el presente artículo propone un algoritmo denominado de Entrenamiento por Bloques de Hablantes (EBHA) para una RNA tipo Perceptron Multicapa (MLP) la cual fue implementada y probada por los autores en un Reconocedor de Voz [Peralta y Cotrina, 2002].

II. DESCRIPCIÓN DEL ALGORITMO EBHA

El Algoritmo EBHA es representado en la figura 1, los parámetros de entrada son representados por el número de hablantes H, el número de palabras que el sistema puede reconocer P y el conjunto de patrones de entrenamiento X_j (j=0,1,..P). La salida está representada por la matriz W, que contiene los pesos resultantes de conexión entre las neuronas.

Figura 1 - Esquema del algoritmo EBHA

2.1 Bloque de hablantes

Después de formar la base de datos de entrenamiento, llamada Corpus [Llamas y Cardeñoso, 1995], compuesta por los patrones de todas las palabras recolectadas, se agrupan aquellos que corresponden a cada hablante, se debe tomar en cuenta que cada

hablante debe haber pronunciado todas las palabras, la agrupación da como resultado a los denominados Bloques de Hablante que contienen las P palabras que el sistema reconoce.

2.2 Entrenamiento de un Hablante

Para describir este proceso de manera didáctica se analiza la figura 2, en la cual se puede apreciar el bloque que contiene los patrones de entrenamiento X_j (j=1,2,..,P). Al iniciarse el entrenamiento, el *switch* se encuentra en la posición 1, de esta manera se tiene como pesos iniciales valores aleatorios y acotados entre cero y uno [Freeman y Skapura, 1993] y contenidos en la matriz W_0. Teniendo como referencia los valores de W_0 se realiza el entrenamiento del primer patrón (X_1), mediante la regla de aprendizaje *Backpropagation* (BACKPR), que devuelve como resultado la matriz de pesos W_1, los mismos que serán utilizados como referencia en el entrenamiento del segundo patrón (X_2). Luego, se continúa con el tercero, y así sucesivamente hasta llegar a entrenar el último patrón X_P que dará como resultado los pesos de las neuronas contenidos en la matriz W_P.

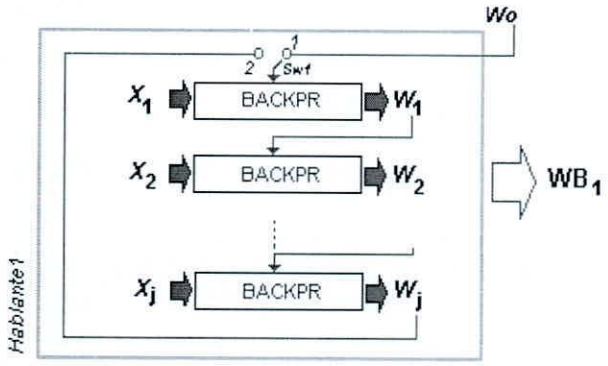

Figura 2 – Entrenamiento de un Hablante

Seguidamente se mide el grado de similitud entre todos los pesos obtenidos, en caso de que sean diferentes, se repite el procedimiento anterior, pero el *switch* pasa a la posición 2, es decir, el primer patrón X_1 vuelve a ser entrenado pero está vez teniendo como referencia los pesos de la matriz W_P. Este procedimiento se repite las veces que sea necesario, hasta que exista un alto grado de similitud entre todos los pesos obtenidos durante el desarrollo del proceso, es decir $W_1=W_2=W_3...=W_P$. Una vez que esto se logra los pesos obtenidos en el último entrenamiento se almacenan en la matriz de pesos WB_1 que es el resultado de entrenar el primer bloque, que corresponde al primer hablante.

2.3 Entrenamiento de Todos los Hablantes

Después de haber entrenado el primer hablante, se utiliza el resultado como referencia para entrenar un bloque siguiente, que corresponderá a un segundo hablante, en el cual se realizará un procedimiento análogo al anterior, y por consiguiente, tendrá como resultado una nueva matriz de pesos, WB_2, como se aprecia en la figura 3. El proceso continuará de manera análoga hasta que se termine de entrenar los H hablantes, es decir hasta que se obtenga WB_H.

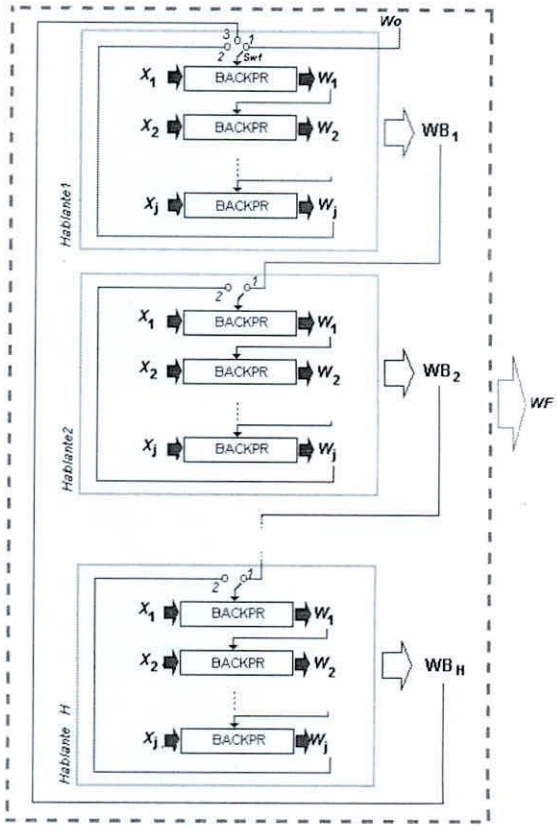

Figura 3 - Entrenamiento de Todo los Hablantes.

Después se efectúa una comparación entre los pesos resultantes del entrenamiento de cada hablante, con la finalidad de analizar la similitud entre ellos, si son diferentes, la secuencia de entrenamiento con cada uno de los bloques se repite, pero esta vez, con WB_H como referencia en el primer hablante (*switch* en posición 3).

Estas secuencias se repetirán hasta que el grado de similitud entre los pesos resultantes de cada bloque alcance un valor muy alto, es decir cuando $WB_1=WB_2=WB_3...=WB_H$. Cuando esto sucede el proceso de entrenamiento finaliza dando como resultado los pesos actualizados con el último

entrenamiento y se guardan en la matriz WF. Con los pesos finales el sistema es capaz de reconocer las palabras independientemente de quien sea el hablante que las pronuncie.

III. EVALUACIÓN DEL ENTRENAMIENTO

3.1 Descripción de la Red Neuronal Utilizada

Para ilustrar y mostrar la eficiencia del algoritmo se ha tomado una RNA de tipo Perceptron Multicapa de 3 niveles que pueda reconocer entre 10 palabras diferentes con las características mostradas en la tabla1.

Tabla 1 - Parámetros de la RNA

Capa	Entrada (1)	Oculta (2)	Salida (3)
Número de Nodos	100	32	10
Función de Transferencia	Lineal	Sigmoidal	Sigmoidal
Constante de la Sigmoidal (k)	...	0.09155	0.15695

El número de nodos en la capa de entrada es igual al número de elementos del vector de características, el cual está dado por 100 coeficientes MFCC (Mel Frecuency Cepstral Coefficients), normalizados entre cero y uno. El número de nodos de la capa de salida está en función del número de palabras que se desea reconocer, mientras que para cada palabra reconocida entregará como respuesta el vector Vi (i=1, 2,..,10), como se muestra en la figura 4, es decir, sólo la neurona asignada a esa palabra tendrá el valor de uno, las demás tendrán cero.

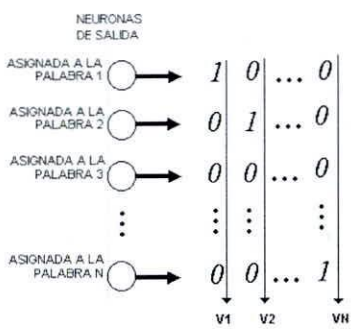

Figura 4 - Nodos de Salidas de la red

Mientras que para seleccionar el número de nodos en la capa oculta, al no existir una regla para definirlo, se optó por escoger el número 32, porque en este valor se tenía menor tiempo de convergencia en el entrenamiento.

Los nodos de las capas Oculta y de Salida, tienen como Función de Transferencia (FT) a la función Sigmoidal. Los valores de las constantes k, en cada una de las capas, han sido escogidas por ser las que produjeron menor tiempo de entrenamiento de la red.

La FT de los nodos de entrada está dada por la función lineal de pendiente unitaria. También se ha convenido asignar cero al valor del Umbral de cada neurona que conforma el Perceptron Multicapa con el propósito de disminuir la carga computacional de la red durante la etapa de entrenamiento y principalmente durante la etapa de reconocimiento, denominada Propagación. En la figura 5, se representa las matrices resultantes Wa entre las capas de Entrada y Oculta, y Wb entre las capas Oculta y de Salida.

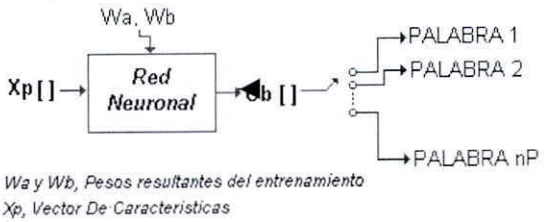

Figura 5 - Proceso de Reconocimiento

3.2 Recolección de Datos

Se recolectaron un conjunto de 56 hablantes, los cuales han sido clasificados en categorías, teniendo en cuenta el sexo y edad, tal como se muestra en la tabla 2.

Tabla 2 - Hablantes Recolectados

SEXO	M	F
Niño (8-11)	4	4
Adolescente (12-17)	6	6
Joven (18-30)	10	10
Adulto (30-60)	6	10
Total	26	30

Cada una de las personas pronunció diez palabras (los dígitos del Cero al Nueve). Se formó una base de datos formada por 560 patrones característicos (muestras), se efectuó un análisis de correlación de los patrones característicos con la finalidad de seleccionar a los hablantes que comparten más características en común dentro del grupo [Llamas y Cardeñoso, 1995]. Para el

análisis de correlación se utilizó la ecuación 4.1 [Chou, 1968], en la cual r_{xy} es el coeficiente de correlación entre las secuencias x e y.

$$r_{xy} = \frac{\sum x' y'}{\sqrt{\sum x'^2 \, y'^2}} \qquad (4.1)$$

En donde:

$$x' = x - \bar{x} \qquad (4.2)$$

$$y' = y - \bar{y} \qquad (4.3)$$

Se aplicó esta fórmula a todos los patrones recolectados (secuencias), tomados de dos en dos, y aquellos que no alcanzaron un coeficiente de correlación mayor a 0.75 fueron considerados como patrones dispersos y fueron separados.

Como consecuencia, se obtuvo un nuevo grupo conformado por 8 hablantes, los cuales se muestran en la tabla 3; con los que se procedió a entrenar la Red Neuronal.

Tabla 3 - Hablantes Seleccionados

SEXO	M	F
Niño (8-11)	1	0
Adolescente (12-17)	0	1
Joven (18-30)	3	1
Adulto (30-60)	1	1
Total	5	3

A la nueva base de datos generada se le denomina Corpus de Entrenamiento y está formado por 800 patrones correspondientes a 10 repeticiones por cada palabra. La base de datos formado por los patrones de cada uno de los 48 hablantes restantes que no fueron seleccionados, se le denomina Corpus de Evaluación [Llamas y Cardeñoso, 1995]. La cual sirve para efectuar una evaluación del sistema en modo *Off Line.*

3.3 Entrenamiento

El proceso de Entrenamiento de la Red Neuronal se realizó utilizando una Computadora Personal Pentium III de 750 MHz. Este proceso se llevó a cabo en 8h con 25 minutos. La tabla 4 muestra los demás parámetros de entrenamiento.

El algoritmo de entrenamiento desarrollado forma parte de un sistema de reconocimiento denominado Reconocedor y Analizador de Voz (RAV) [Peralta y Cotrina, 2002]

Tabla 4 Parámetros de Entrenamiento.

Numero de Patrones	800
Numero de Palabras (nP)	10
Numero de Hablantes	8
Hablantes Relativos (nH)	80
Error de patrón (ep)	0.025
Error de Hablante (mm)	0.0001
Error de grupo de hablantes (aa)	0.00001

3.4 Evaluación Off Line

La tabla 5, muestra los resultados obtenidos en la evaluación *Off Line, e*n ella se puede apreciar la tasa de acierto del algoritmo para cada palabra pronunciada. En la primera columna se indican la palabras que han sido pronunciadas, y en la primera fila se indican las palabras que se han reconocido. Como se puede observar, el sistema presentó una tasa de Acierto de 100% en casi toda las palabras, excepto en las palabras 'Tres' y 'Seis' por ser de naturaleza fricativa y además en la palabra 'Cuatro', debido a que esta tiene un silencio intermedio.

Tabla 5 - Evaluación Off line

	Cero	Uno	Dos	Tres	Cuatro	Cinco	Seis	Siete	Ocho	Nueve
Cero	48									
Uno		48								
Dos			48							
Tres			2	46						
Cuatro					45				3	
Cinco						48				
Seis							45	3		
Siete								48		
Ocho									48	
Nueve										48
t.a.	100	100	100	96	94	100	94	100	100	100

t.a. 98 %

La figura 6, muestra una gráfica de barras que expresa en términos de porcentaje los resultados obtenidos en la evaluación *Off Line.* La tasa de acierto global del sistema es de 98.125 %, lo cual indica que el resultado de la prueba fue exitoso.

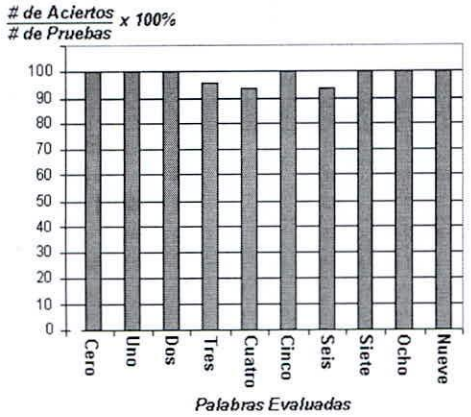

Figura 6 - Resultados de la evaluación *Off Line*

3.5 Evaluación On Line

La tabla 6, muestra la eficiencia del sistema en modo *On Line* sobre veinte personas distribuidas por género y edades.

Tabla 6 - Hablantes de prueba

SEXO	M	F
Niño (8-11)	2	2
Adolescente (12-17)	2	3
Joven (18-30)	3	3
Adulto (30-60)	2	3
Total	9	11

La prueba se realizó en dos etapas; en la primera se evaluó el sistema en un entorno que se considera aislado de ruido de fondo, aunque estaba presente el ruido eléctrico producido por los ventiladores de la computadora; se obtuvo como resultado una tasa de acierto de 91.65 %. La tabla 7 muestra los resultados obtenidos en esta prueba, en ella se detalla la eficiencia obtenida en cada categoría de hablante donde nH es el número de hablantes, nP-p es el número de palabras pronunciadas, nP-a es el número de palabras reconocidas con éxito y T.A. es la tasa de acierto.

Tabla 7 - Evaluación *On Line* sin ruido de fondo.

	nH	nP-p	nP-a	T.A.(%)
Niño (8-11)	4	400	346	86.5
Adolescente (12-17)	5	500	465	93
Joven (18-30)	6	600	561	93.5
Adulto (30-60)	5	500	461	92.2
Totales	20	2000	1833	91.65

En la segunda etapa se evaluó el sistema en un ambiente con ruido musical que producía una figura de ruido que oscila entre 20 y 30 dB; en estas condiciones, la tasa de acierto del sistema disminuyó a 87.4 %, tal como se muestra en la tabla 8.

Tabla 8 - Evaluación *On Line* con ruido de fondo.

	nH	nP-p	nP-a	T.A.(%)
Niño (8-11)	4	400	336	84
Adolescente (12-17)	5	500	445	89
Joven (18-30)	6	600	533	88.83
Adulto (30-60)	5	500	434	86.8
Totales	20	2000	1748	87.4

Las figuras 7 y 8 muestran la eficiencia alcanzada por cada palabra pronunciada en la evaluación *On Line* sin y con ruido respectivamente.

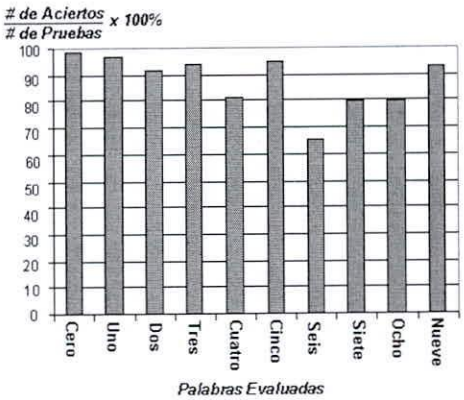

Figura 7 - Eficiencia del reconocimiento *On Line* sin ruido

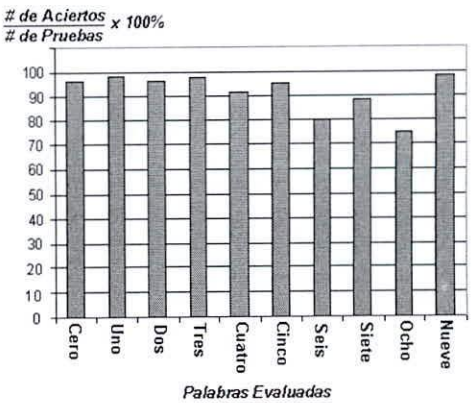

Figura 8 - Eficiencia del reconocimiento *On Line* con ruido

Finalmente, el gráfico de la figura 9 muestra una evaluación global de la eficiencia del algoritmo, contrastando los resultados del reconocimiento en los modos *Off Line* y *On Line*, este último, en ambientes con ruido y en ambientes sin ruido.

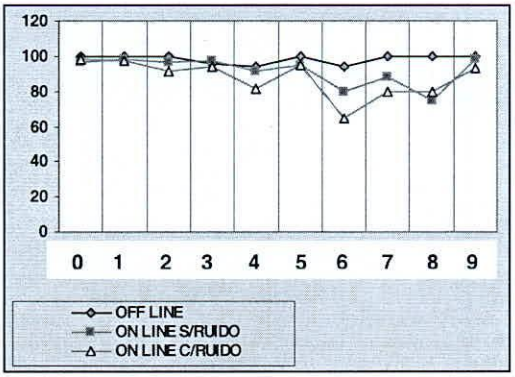

Figura 9 - Comparación de la Eficiencia del Algoritmo EBHA

IV. CONCLUSIONES

Se ha creado un algoritmo de entrenamiento para la Red Neuronal tipo Perceptron Multicapa basado en la regla de aprendizaje *Backpropagation*, la cual clasifica las palabras pronunciadas por distintos hablantes de acuerdo a la información lingüística y cuyos resultados muestran una tasa de acierto muy alta.

En la evaluación en modo *Off Line* se obtuvo una tasa de acierto de 98.125%, lo que demuestra que dicho algoritmo tiene una gran eficiencia para el establecimiento de clases. La tasa de acierto en el modo *On Line* en un ambiente sin ruido de fondo es de 91.65% que es menor que para el modo *Off Line*, como se puede apreciar en las tablas 5 y 7. Esta aparente contradicción se explica por el hecho de que en el modo *On Line* a pesar de no existir ruido de fondo, si existe ruido producido por la computadora especialmente de la fuente de poder y los ventiladores, y también los ocasionados de manera involuntaria por los hablantes evaluados; los cuales no ocurren para el modo *Off Line*.

Al evaluar el reconocimiento en entornos ruidosos con Figuras de Ruido entre 15 dB y 30 dB, la tasa de Acierto fue de 87.4%, lo que implica que la Red Neuronal puede trabajar en presencia de ruido de fondo intensos.

AGRADECIMIENTOS

Al Instituto de Investigación de la Facultad de Ingeniería Electrónica de la UNMSM, por dar las facilidades para realizar el trabajo.

REFERENCIAS

Bermúdez, J.B., J. Bobadilla y P. Gómez. (2000). Reconocimiento de Voz y Fonética Acústica. Ediciones Alfaomega.

Llamas, C., V. Cardeñoso (1995). Reconocimiento Automático del Habla. Teoría y Aplicaciones. Universidad de Valladolid.

Cater, J. (1984). Electronically Hearing: Computer Speech Recognition 1st Edition. Howard W. Sams y Co., Inc.

Kartalopoulos, S. (1996). Understanding Neuronal Networks and Fuzzy Logic. IEEE Press.

Hilera, J.R. y V.J. Martínez (1995). Redes Neuronales Artificiales. Fundamentos, Modelos y Aplicaciones. Editorial Addison-Wesley Iberoamericana.

Freeman, J.A. y D.M Skapura (1999). Redes Neuronales, Algoritmos, Aplicaciones y Técnicas de Programación, ADDISON-WESLEY.

Sphar C. (1999). Aprenda Microsoft Visual C++ 6.0 Ya. Mc. Graw-Hill.

Schildt ,H. Turbo C/C++. Manual de Referencia.

Hanselman, D. y B. Littlefield (1996). Matlab edición de estudiante. Guía de usuario Versión 4. Editorial Prentice Hall.

Peralta, F. y A. Cotrina (2002). Reconocedor y Analizador de Voz. Tesis Titulo Profesional. Universidad Nacional Mayor de San Marcos Lima Perú.

Crespo, C., C. de la Torre y J.C. Torrecilla. Detector de extremos para reconocimiento de voz. Telefónica Investigación y Desarrollo. Publicación de Telefónica I+D. S:A. Madrid España. http://www.tid.es/presencia/publicaciones/comsid/esp/home.html. Fecha de acceso: Enero de 2001.

Hernández, L., F. J. Caminero, C. de la Torre y L. Villarrubia. Estado del arte en Tecnología del Habla. Telefónica Investigación y Desarrollo. Publicación de Telefónica I+D. S.A. http://www.tid.es/presencia/publicaciones/co

msid/esp/home.html. Fecha de acceso: Enero de 2001.

Poza M. J., L. Villarrubia y J. A. Siles. Teoría y aplicaciones del reconocimiento automático del habla. Telefónica Investigación y Desarrollo. Publicación de Telefónica I+D. S.A. http://www.tid.es/presencia/publicaciones/comsid/esp/home.html. España. Fecha de acceso: Enero de 2001.

Esteban, A.C. y N.M. Carrascal. Reconocimiento de Voz mediante el Perceptron Multicapa. Laboratorio de Electroacústica II (SSR). http://www.intersaint.org/acid/rvpm0.htm. Fecha de acceso: Enero de 2001.