UTILIZACIÓN DE FUNCIONES DISCRIMINANTES LINEALES GENERALIZADAS EN EL RECONOCIMIENTO DE PALABRAS AISLADAS CON CIERTOS DICCIONARIOS DIFÍCILES.

F.J. CORRELL ABAD, E. VIDAL RUÍZ, F. CASACUBERTA NOLLA, J.C. TABARÉS SEISDEDOS
UNIVERSIDAD POLITÉCNICA DE VALENCIA

Ciertos Clasificadores no paramétricos se han utilizado con bastante éxito en el diseño de sistemas de Reconocimiento de Palabras Aisladas. No obstante, cuando el diccionario está formado por palabras muy semejantes en las que sólo se diferencian unos pocos segmentos (zona discriminante), estos sistemas presentan una alta tasa de errores. En este trabajo se propone un método en el que se potencia el papel de las zonas discriminante y que está basado en una extensión del algoritmo de minimización del Criterio Perceptrón. En un estudio experimental se ha demostrado que la eficacia del método propuesto aumenta en más de un 50% con respecto a la aproximación clásica.

Keywords: Isolated Word Recognition; Difficult Dictionaries; Generalized Linear Discriminant Function.

1. INTRODUCCIÓN

En la actualidad existen diversos sistemas comercializados para el Reconocimiento de Palabras Aisladas (RPA) conectables a muchos computadores /1/. No obstante, presentan un funcionamiento defectuoso cuando el diccionario (conjunto de palabras reconocibles) está formado por palabras muy parecidas. Estos sistemas suelen ser Clasificadores de Distancia Mínima, los cuales comparan la representación, en un cierto espacio, de una palabra muestra, con todas las representaciones de las palabras de un diccionario (prototipos). En cada comparación se obtiene una cierta Medida de Disimilitud (MD) suministrada por algoritmos de Alineamiento Temporal No Lineal (ATNL) que siguen esquemas de Programación Dinámica /2/.

Los algoritmos de ATNL establecen una relación entre segmentos semejantes que dan lugar a un "camino de alineamiento óptimo" (CAO) en el plano de comparación entre las dos palabras. Dicho CAO se puede expresar como una secuencia discreta de "producciones", esto es de direcciones e incrementos permitidos en el plano. La MD se obtiene como un sumatorio de las distancias locales entre segmentos relacionados por el CAO y ponderado con una serie de pesos asociados a las producciones que solo dependen de éstas (sección 3).

• *F.J. Correll Abad, *E.Vidal Ruíz, *F. Casacuberta Nolla, **J.C. Tabarés Seisdedos.
*Dep. de Sistemas Informáticos y Computación - U.P.V.
**Centro de Informática de la U.P.C. - Camino de Vera, s/n Valencia.
• Article rebut el març de 1987.

En el cálculo de la MD entre dos palabras contribuyen de forma más o menos equitativa todos los segmentos de dichas palabras. Supongamos que se desean comparar las palabras /estalactita/ y /estalagmita/, en este caso las diferencias (/ct/ y /gm/ respectivamente) son de pequeña duración frente al resto de la palabra. Este hecho hace que las contribuciones por dichas diferencias a la MD puedan confundirse con el margen de variabilidad propia de la palabra.

Este problema ya ha sido tratado en la literatura /3-5/, aunque no se han encontrado soluciones plenamente satisfactorias.

En este trabajo se propone un nuevo método en el que se pretende que las zonas discriminantes de las palabras tengan más importancia que el resto. Este método consiste en sustituir los pesos de las producciones en el cálculo final de la MD por una serie de pesos variables asociados a cada prototipo. Dichos pesos variables dan lugar a un vector que define una función discriminante y que en este trabajo se obtienen en una etapa de aprendizaje mediante la aplicación del algoritmo de minimización del Criterio del Perceptrón /6/ que es descrito en la sección 2.

Los resultados experimentales (secciones 4 y 5) obtenidos con este método permiten una disminución de la tasa de error de más de un 50% con respecto a los métodos clásicos, en el mejor de los casos, con un aumento de la complejidad temporal y espacial poco significativo (sección 6).

2. FUNCIONES DISCRIMINANTES LINEALES: GENERALIZACION

2.1. Funciones discriminantes lineales.

En un sistema de Reconocimiento Geométrico de Formas (RGF) cada clase puede caracterizarse mediante una función del espacio de representación de los objetos en los reales, denominada discriminante. La propiedad fundamental de tales funciones, es que para una muestra dada, el valor de la función de la clase a la que pertenece es el mínimo de entre los valores de todas las funciones discriminantes para esa muestra.

Formalmente, sea E el espacio de representación donde están definidos los objetos, y sean las c clases o formas posibles:

$$(1)\ C_1\ ,\ C_2\ ,\ldots,\ C_c\ /\ C_i \in E \text{ con } C_i \cap C_j = \emptyset,\ i\neq j \text{ y } \forall i,j=1\ldots c$$

Definiremos $\forall \vec{x} \in E$ las funciones discriminantes $D_i(\vec{x})$, i=1...c /6/ asociadas a cada forma C_i i=1...c, como funciones del espacio E en $\mathbb{R}$ tal que si

$$(2)\ \vec{x} \in C_j \iff D_j(\vec{x}) > D_i(\vec{x}) \text{ con } i\neq j,\ \forall i,j=1\ldots c$$

eligiéndose arbitrariamente en caso de igualdad.

La frontera de decisión entre las regiones del espacio asociadas a las clases C_i y C_j viene determinada por la ecuación /6/

(3) $D_i(\vec{x}) - D_j(\vec{x}) = 0$

Supongamos que E es un espacio de N dimensiones, por lo tanto $\vec{x} \in E$ vendrá representado por un vector $x = (\vec{x}_1, \ldots, \vec{x}_N)$. En este contexto se define una función discriminante lineal (FDL) $g(\vec{x})$ como:

$$(4)\ g(\vec{x}) = \sum_{k=1}^{N} W_k\ x_k + W_o = \vec{W}^t\ \vec{x} + W_o$$

donde $\vec{W}$ es el "vector de peso", que define la función discriminante, y W_o es el "peso umbral".

En el caso de c clases tenemos c funciones discriminantes lineales; de forma que asignaremos $\vec{x}$ a la clase C_i si:

$$(5)\ g_i(\vec{x}) > g_j(\vec{x}) \qquad \forall j,\ i \neq j,\ i,j=1\ldots c$$

Añadiendo términos cuadráticos, cúbicos, etc. a la definición de FDL obtenemos el concepto más generalizado de funciones discriminantes polinómicas. Dichas funciones pueden verse, a su vez, como un caso particular del concepto de Función Discriminante Lineal Generalizada (FDLG), definida como /6/:

$$(6)\ g_i(\vec{x}) = \vec{a}_i^t\ \vec{y}(\vec{x}) = \sum_{k=1}^{\hat{e}} a_{i_k}\ \vec{y}_k(\vec{x}) \qquad i=1\ldots c$$

donde $\vec{a}_i$ es un vector de peso de dimensión $\hat{e}$, y la función $\vec{y}(\vec{x})$ es función arbitraria de $\vec{x}$.

Como se observa, las FDLG son lineales en $\vec{y}$ pero no en $\vec{x}$.

Utilizando estos vectores peso $\vec{a}_i$, cabe definir igualmente unas fronteras de decisión que dividen al espacio de representación en c clases. Así, el problema del reconocimiento, queda sustituido por el de evaluar la clase de pertenencia dentro de dicho espacio de representación.

2.2. *Aprendizaje*

El aprendizaje en el contexto de este trabajo, consiste en la adquisición de los vectores de pesos que definen las funciones discriminantes y que deben realizar una adecuada partición del espacio de representación.

Existen diversos métodos iterativos para la obtención de estos vectores de peso. En el caso de dos clases, basta definir una función discriminante, que será positiva para una clase y negativa para la otra, y por tanto un vector de peso $\vec{a}$. El método general del descenso por gradiente consiste en definir una función escalar $J(\vec{a})$ que sea mínima para el "vector de peso solución" $\vec{a}$.

La búsqueda de los vectores peso solución que hacen mínimas J_i $(\vec{a}_i)$ se realiza mediante el algoritmo siguiente /6/.

$$(7)\quad \begin{array}{l} \vec{a}^1 \text{ arbitrario} \\ \vec{a}^{k+1} = \vec{a}^k - \sigma \nabla J(\vec{a}^k) \end{array}$$

Siendo σ factor de escala real que intensifica o decrementa las correcciones que se van realizando en el proceso de descenso por gradiente.

Si tomamos σ constante para todas las correcciones realizadas se hablará de algoritmo de descenso por gradiente con incremento constante. Si tomamos σ variable en cada corrección se hablará de algoritmo de descenso por gradiente con incremento variable, y se introducirá entonces un superíndice en el factor de escala σ.

Una definición usual de las funciones escalares $J(\vec{a})$ es el Criterio Perceptrón, que definimos como:

$$(8)\quad J(\vec{a}) = \sum_{\vec{y} \in Y} (-\vec{a}^t \cdot \vec{y})$$

donde Y es el conjunto de muestras mal clasificadas por el conjunto de vectores peso $\vec{a}$.

En este caso el algoritmo de descenso por gradiente (7) será:

$$(9)\quad \begin{array}{l} \vec{a}^1 \text{ arbitrario} \\ \vec{a}^{k+1} = \vec{a}^k + \sigma \sum_{\vec{y} \in Y_k} \vec{y} \end{array}$$

donde Y_k es el conjunto de muestras mal clasificadas por $\vec{a}^k$.

Siguiendo este razonamiento, en el caso de c clases, el problema de la obtención de los c vectores de pesos puede ser resuelto mediante un algoritmo que parte de una máquina lineal L_1 (conjunto de vectores $\vec{a}_i^1$ i=1...c) arbitrarios y realiza la evolución de una máquina lineal L_k (conjunto de vectores $\vec{a}_i^k$ i=1...c), que nos produce cierta partición del espacio de representación, a una máquina lineal L_{k+1} que realiza una nueva partición del espacio de representación; siendo la máquina L_{k+1} una máquina más cercana a la máquina solución L_s (la que hace mínimas las funciones J_i i=1...c).

Dicho algoritmo puede presentarse de forma que la iteración "correctiva" se realice considerando las muestras una a una, repitiéndolas cíclicamente en caso necesario /6/. De esta forma en el caso de que una muestra $\vec{y}(x)$ haya sido mal clasificada (la clase de pertenencia de dicha mues-

tra es C_i, y la máquina lineal L_k la ha clasificado en la clase C_j), entonces el algoritmo resultante para la generación de una nueva máquina lineal L_{k+1} puede enunciarse como:

$$\underline{si}\ \vec{x} \in C_i \wedge g_j(\vec{x}) = \max_{\forall k:1\ldots c} (g_k(\vec{x}) \wedge i \neq j$$

$$(10)\quad \vec{a}_i^{k+1} = \vec{a}_i^{k} + \sigma^k \vec{y}\ \ \vec{a}_j^{k+1} = \vec{a}_j^{k} - \sigma^k \vec{y}\qquad \vec{a}_m^{k+1} = \vec{a}_m^{k}\quad m\neq i, m\neq j,\ m=1\ldots c$$

Es decir, que la máquina lineal L_{k+1} tendrá incrementados los vectores peso de la clase de pertenencia de $\vec{y}(\vec{x})$, y decrementados los de la clase en que $\vec{y}(\vec{x})$ se había clasificado incorrectamente, dejando los demás sin cambios. Puede demostrarse que el algoritmo definido por (10) converge a una máquina solución en un número finito de iteraciones /6/ si el conjunto de muestras $\vec{y}(\vec{x})$ es linealmente separable.

2.3. Generalización propuesta.

Las FDLG propuestas hasta ahora, requieren que las muestras sean representadas en un único espacio de dimensión constante. Sin embargo, hay algunos casos de RPA en los cuales esta condición no puede ser satisfecha, ante lo cual proponemos una nueva generalización de las FDLG. Para el caso de multicategoría (c clases), las definiremos como:

$$(11)\quad g_i(x) = \sum_{j=1}^{\hat{e}_k} a_{ij}\ y_{ij}(x) = \vec{a}_i\, \vec{y}_i(x) \qquad \text{con } i=1\ldots c$$

donde ahora las muestras son de "dimensión variable"; es decir donde un objeto x perteneciente a cierto conjunto E tendrá una representación vectorial $\vec{y}_j(x)$, que será distinta en cada clase j..

$$(12)\quad x \longrightarrow \vec{y}_j = (y_{j1}, \ldots, y_{jL}) \qquad\qquad j=1\ldots c$$

donde j indica la clase en cuyo espacio se representa x, y L la dimensión en dicha clase.

Esta generalización induce un nuevo algoritmo de aprendizaje que ahora pasará a ser:

$$(13)\quad \begin{aligned} \vec{a}_i^{k+1} &= \vec{a}_i^{k} + \vec{\sigma}^k \vec{y}_i \\ \vec{a}_j^{k+1} &= \vec{a}_j^{k} - \sigma^k \vec{y}_j \\ \vec{a}_m^{k+1} &= \vec{a}_m^{k} \qquad m\neq i,\ m\neq j,\ m=1\ldots c \end{aligned}$$

La demostración de convergencia de este algoritmo (que se omite en aras a la brevedad) se basa en la construcción de Kesler /6/, siguiendo una linea similar a la utilizada para demostrar la convergencia de (10).

Por último, para mejorar el comportamiento del método en las regiones fronterizas del conjunto E de definición de los objetos x (aquellas en las que dos o más funciones discriminantes toman valores próximos), se in-

troduce el concepto de margen b, de forma que una muestra x estará clasificada dentro de la clase C_j si:

$$(14)\quad g_j(x) - g_i(x) > b \qquad b \in \mathbb{R}, \quad \forall i,\ i \neq j,\ i,j=1...c$$

Y consideraremos la "frontera de decisión" de una clase C_j con otra clase C_i como aquella zona de E en la que se cumple:

$$| g_j(x) - g_i(x) | \leq b \qquad \forall i,\ i \neq j,\ i,j=1...c$$

En este caso el algoritmo de aprendizaje resulta como sigue:

$$\underline{\text{si}}\ x \in C_i \ \wedge\ g_j(x) - g_i(x) > b$$

$$\vec{a}_i^{\,k+1} = \vec{a}_i^{\,k} + \sigma^k \vec{y}_i$$

$$(15)\qquad \vec{a}_j^{\,k+1} = \vec{a}_j^{\,k} - \sigma^k \vec{y}_j$$

$$\vec{a}_m^{\,k+1} = \vec{a}_m^{\,k} \qquad m \neq i,\ m \neq j,\ m=1...c$$

Donde se adopta $\sigma^k = \sigma$ (incremento constante) o $\sigma^k = \sigma/k$ (incremento variable) según se convenga, y donde σ es una constante real que denominaremos factor de corrección.

3. RECONOCIMIENTO DE PALABRAS AISLADAS

En la Aproximación Global al RPA, los objetos, (palabras) se representan mediante una secuencia de longitud variable de vectores de parámetros. Esta se obtiene en una etapa de Parametrización, que consiste en aplicar algunas técnicas de procesado de señal en intervalos temporales constantes y sucesivos de la señal vocal (ventanas) capturada por algún sensor (micrófono) /2/.

Las fuentes de conocimiento asociadas al tipo de RPA que nos ocupa están formadas por una (o varias) representaciones paramétricas de cada una de las palabras que forman el diccionario de prototipos. El proceso de interpretación de una muestra dada, también en forma paramétrica, consiste en compararla con todas las representaciones de las palabras del diccionario, y en aplicar algún tipo de Criterio de Decisión. Como resultado de dicha comparación se obtienen un conjunto de Medidas de Disimilitud (MD), a las que se aplica un Criterio de Decisión. Por ejemplo, se asocia a la muestra dada, el patrón que haya presentado menor MD (criterio de decisión de distancia mínima).

3.1. Función de Alineamiento Temporal (FAT).

Una característica del habla es el de su variabilidad, esto es, un mismo locutor nunca pronuncia dos veces la misma palabra de forma idéntica. Dicha variabilidad no es uniforme, sino que es más acentuada en las zonas vocálicas que las consonánticas. Por lo tanto este problema pone de mani-

fiesto la necesidad de métodos de normalización temporal no lineal en el proceso de comparación. Dichos métodos fueron introducidos por Matheus y Davis en 1946 /2/ basándose en técnicas de Programación Dinámica (PD).

Dados dos objetos a comparar X e Y, el procedimiento de normalización temporal usual se basa en hallar la "función de alineamiento" óptima entre secuencias discretas de vectores de parámetros.

Si $I_x = \{1,2,...I\}$, $I_y = \{1,2,...,J\}$ son los intervalos temporales discretos en que están definidos los objetos X e Y, (donde I y J son sus duraciones: número de vectores de X e Y, respectivamente), las secuencias de vectores discretas de X e Y vendrán pues representadas como:

(16) $X = (x(1),x(2),...,x(I))$
$Y = (y(1),y(2),...,y(J))$

entonces podemos definir la FAT en el plano discreto $I_x \times I_y$ como un "camino" /2/:

(17) $F(k) = (i(k),j(k)) \quad i(k) \in I_x \quad j(k) \in I_y \quad \forall k \in I_F$

donde $I_F = \{1,...,L\}$ y L es el número de puntos de F.

La FAT óptima será aquella que partiendo del origen (0,0) llegue al extremo (I,J) minimizando la suma de las distancias locales:

(18) $d(x(i),y(j)) \quad i=i(k), \quad j=j(k)$

definidas sobre él (distancia entre vectores de parámetros). A esta suma de distancias locales a través del "camino" F(k) dividida (para normalización) por la longitud de F se le denomina distancia acumulada, y a la mínima suma normalizada (obtenida a través del camino "óptimo") se le denomina distancia mínima.

Existen, no obstante, restricciones físicas sobre el "camino". Atendiéndonos a ellas, la FAT deberá ser "continua" y "monótona creciente". Otras restricciones se pueden referir a la "elasticidad" tolerable para las deformaciones relativas de los objetos a comparar. Esto hace que se impongan restricciones a los valores máximos y mínimos de la FAT.

Todas estas restricciones, se pueden traducir en la elección de un conjunto de "producciones" (segmentos rectilineos orientados en el plano $I_x \times I_y$) /5/,/2/ que modelizan las direcciones locales así como los incrementos temporales permitidos para "moverse" en el plano discreto de tiempos $I_x \times I_y$.

Cada una de las producciones (a,b) de este conjunto lleva asociado un peso w(a,b) que afectará a la distancia local correspondiente. Estos pesos pueden considerarse como una definición de la "métrica temporal" en el plano de comparación, y usualmente se definen como w(a,b)=a+b. Esta definición permite realizar la minimización arriba citada mediante técnicas de PD /2/.

Ejemplo 1:

Veamos un ejemplo de obtención de la FAT y distancia mínima entre 2 objetos (cadenas alfanuméricas en este caso) x= "eeeme" y p = "emme".

Un posible cálculo de la FAT (camino mínimo) podría ser el de la figura 1.

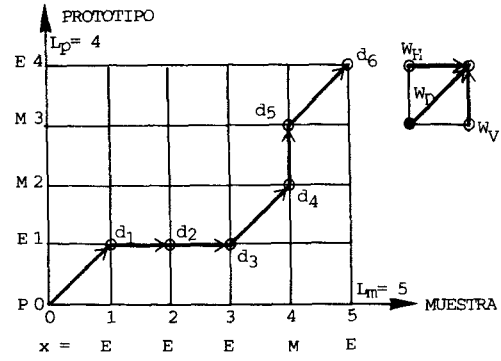


FIGURA 1: Ejemplo de FAT (camino mínimo) para una muestra y un prototipo concretos, con $w_H = 1$ $w_D = 2$ $w_V = 1$.

Así la mínima, calculada mediante PD clásica sería:

$$(19)\quad D_{CL}(x,p) = (1/(L_x + L_p))(w_D d_1 + w_H d_2 + w_H d_3 + w_D d_4 + w_V d_5 + w_D d_6) = (1/(L_x + L_p))(w_D (d_1 + d_4 + d_6) + w_H (d_2 + d_3) + w_V d_5)$$

siendo w_D , w_H , w_V los pesos (1,2,1) correspondientes a las producciones utilizadas ((1,0)(1,1)(0,1) en nuestro caso respectivamente) y siendo d_i las distancias locales.

Actualmente los algoritmos de PD son usados en RPA con muy buenos resultados. Estos sistemas tienen algunas ventajas:

a) Elevado porcentaje de reconocimiento.
b) Los algoritmos son bastante independientes de las particularidades del lenguaje.

Pero también tienen sus inconvenientes:

a) Su vocabulario es bastante limitado. Y esto es debido a la memoria necesaria para almacenar los diccionarios, así como el tiempo de reconocimiento que es del orden de: O(nl) siendo n el número de prototipos e l el tamaño medio de las palabras.
b) Su dependencia del locutor o locutores que hayan introducido el diccionario.

c) Rápida degradación cuando la similitud entre las palabras del diccionario aumentan.

3.2. El problema de los diccionarios difíciles: solución propuesta.

Supongamos que se desea comparar dos palabras que sólo se diferencian en un pequeño segmento de señal vocal, por ejemplo /estalactita/ y /estalagmita/. En el cálculo de la distancia mínima, que hemos comentado anteriormente, se observa que la contribución de las zonas discernientes /ct/ y /gm/ son una pequeña parte en comparación con las zonas comunes. Este hecho trae consigo el que en ocasiones dicha contribución pueda ser confundida con la variabilidad propia de las palabras. Cuando un diccionario contiene palabras muy parecidas, entonces se dice que el diccionario es "difícil".

Nuestro intento de mejora consiste en una aplicación de los métodos introducidos en la sección anterior considerando como representación vectorial de una muestra a un vector construido a partir de las distancias locales calculadas a través del camino "óptimo" y como "vectores peso" a los vectores construidos a partir de los pesos de las producciones utilizadas. El proceso de aprendizaje se inicia considerando como máquina lineal inicial L_1 , la obtenida asignando los valores constantes clásicos de los pesos (w_V , w_D , w_H) a todas las componentes correspondientes a los vectores de peso de cada clase. A continuación se utiliza un conjunto de muestras de aprendizaje (palabras pronunciadas) para corregir estos pesos hasta llegar a una máquina lineal L_S , que realiza la mejor discriminación posible de un conjunto de las muestras de aprendizaje.

La representación vectorial de una muestra x, dentro de una clase concreta, consiste en la construcción de un vector a partir de la composición de tres vectores. Cada uno de estos vectores está asociado a una producción, y la componente i de cada uno de ellos es la distancia local asociada al punto i del patrón al ser aplicada la producción asociada. Veamos esto con el ejemplo 1:

El punto 1 (figura 1) del patrón interviene en el cálculo de d_1 con una producción (1,1) y en la d_2 y d_3 con una producción (1,0). El punto 2 interviene en el cálculo de d_4 con una producción (1,1) de la misma forma el punto 3 en el de d_5 con una producción (0,1) y el punto 4 en el de d_6 con una producción (1,1). Si para cada punto del patrón hacemos intervenir las tres producciones, (con una distancia local 0 si el punto en cuestión no interviene en el cálculo de alguna distancia), podríamos reescribir (19) como:

$$D(x,p) = (1/(L_x + L_p))(w_D\ d_1 + w_H\ (d_2 + d_3) + w_V\ 0 + w_D\ d_4 + w_H\ 0 + w_V\ 0 + w_D\ 0 + w_H\ 0 + w_V\ d_5 + w_D\ d_6 + w_H\ 0 + w_V\ 0$$

o en la forma vectorial como:

$$D(x,p) = (1/(L_x + L_p))\vec{a}^t\ \vec{y}(x)$$

donde:

$$\vec{a} = (w_D, w_D, w_D, w_D, w_H, w_H, w_H, w_H, w_V, w_V, w_V, w_V)^t$$

$\vec{y}(x) = (d_1, d_4, 0, d_6, (d_2 + d_3), 0, 0, 0, 0, 0, d_5, 0)^t$

Obsérvese que $y_i + L_p(k-1)$ con $k=1,2,3$, L_p = tamaño del patrón, e $i=1,2,\ldots,L_p$, representa la distancia acumulada en la que interviene el punto i del patrón mediante la producción k (k=1 es (1,1), k=2 es (1,0) y k=3 es (0,1)).

Si $w_D = 2$, $w_H = 1$ y $w_V = 1$ tenemos que D(x,p) es la distancia acumulada (19) obtenida de forma clásica.

Si ahora consideramos que los pesos del vector a pueden variar con la posición correspondiente (i) del patrón, es decir:

$$\vec{a} = (w_D^1, w_D^2, w_D^3, w_D^4, w_H^1, w_H^2, w_H^3, w_H^4, w_V^1, w_V^2, w_V^3, w_V^4)$$

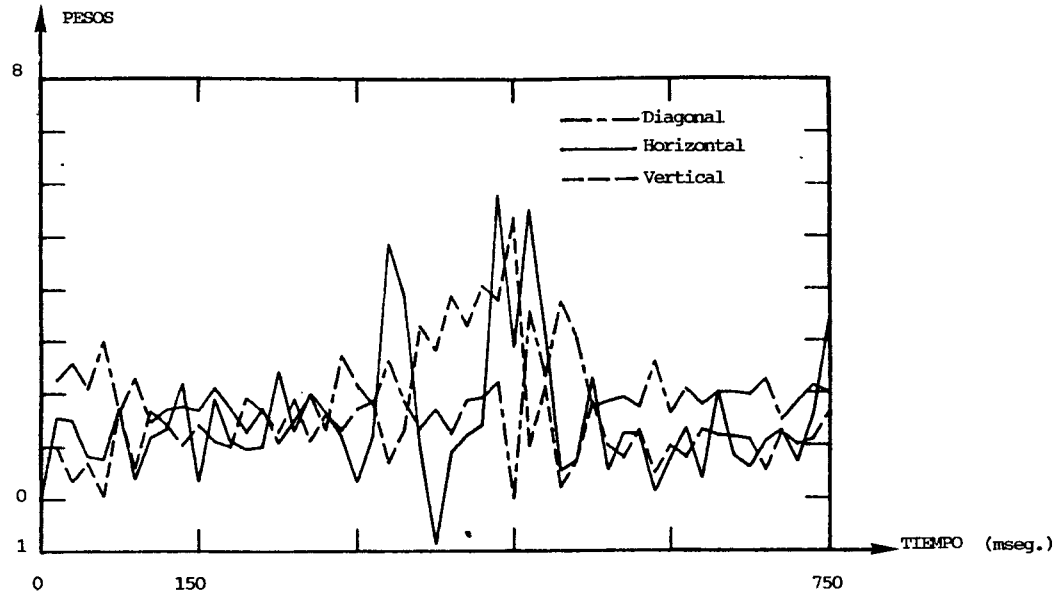
podemos utilizar el algoritmo (15) para aprender los valores de estos pesos que produzcan una clasificación óptima. Es de esperar, que la distribución de estos pesos en función de su posición respecto al prototipo correspondiente, reflejan las restricciones espectrales y temporales que determinan el carácter discriminante de los segmentos a ellos asociados.

Nótese que la representación de una muestra tal como x en el ejemplo anterior no es única; de hecho, habrá que tener en cuenta que al variar la clase (representada por su prototipo), la representación es totalmente distinta:

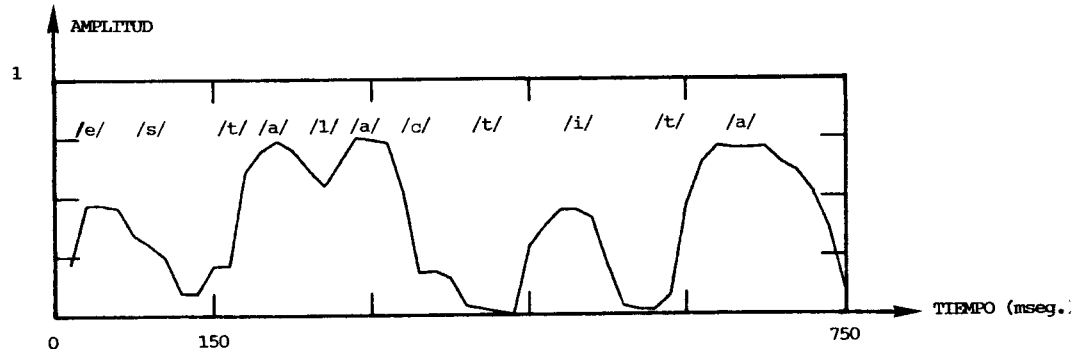
1) Varía el camino de PD, y las distancias locales, ya que cambiamos el prototipo.
2) Puede variar la longitud del prototipo, por lo que variaría la dimensión del vector $\vec{y}$ al pasar de una clase a otra.

Así implementado, siguiendo las generalizaciones del apartado 2.3, y utilizando un diccionario compuesto por: (/estalactita/, /estalagmita/), los vectores correspondientes a la máquina lineal obtenida tras un aprendizaje basado en 21 muestras se representan en las figuras 2 y 3.

Los pesos variables asociados a cada una de las 3 producciones se representan utilizando distinto tipo de trazo, en las figuras 2(a) y 3(a), mientras que en las figuras 2(b) y 3(b) se representa la amplitud de la señal vocal del prototipo correspondiente. Estas últimas permiten situar (al menos aproximadamente) los distintos elementos acústico-fonéticos de las palabras representadas, lo que permite juzgar la adecuación de los pesos obtenidos. Nótese como los valores de los pesos oscilan alrededor de los valores iniciales (clásicos 1,2,1) a lo largo de cada palabra excepto en las zonas discriminativas en donde presentan variaciones significativas.



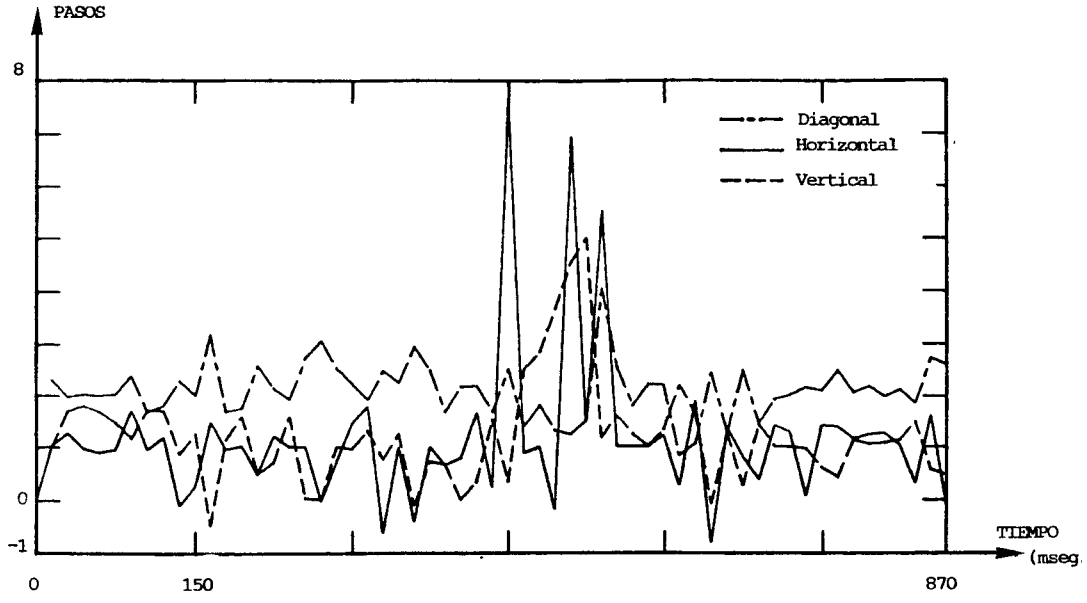
a)



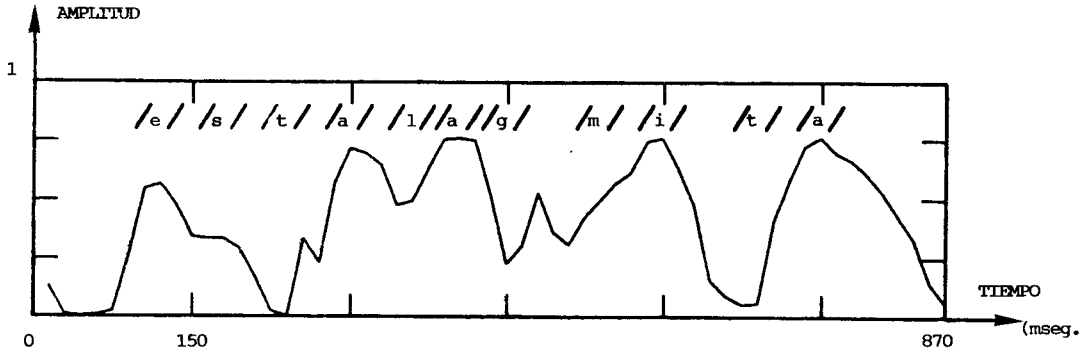
b)

FIGURA 2 a: Pesos modificados por Perceptrón de la palabra ESTALACTITA.

b. Amplitud de la palabra ESTALACTITA.



a)



b)

FIGURA 3: a: Pesos modificados por Perceptrón de la palabra ESTALAGMITA.

b: Amplitud de la palabra ESTALAGMITA.

4. SISTEMA EXPERIMENTAL

El esquema general del sistema experimental implementado es el de la figura (4), donde caben resaltar algunos puntos:

a) La representación numérica (RN) utilizada fue la de secuencia de vectores de 11 parámetros cepstrales /8/.
b) La implementaciíon del sistema experimental permite poder trabajar con dos tipos de prototipos-base:
   i) Prototipos $\emptyset$: utilizamos aquí arbitrariamente los prototipos correspondientes a la primera repetición de cada palabra en el aprendizaje.
   ii) Centroides: utilizamos aquí como prototipos a los centroides de las muestras de la clase i utilizadas en el aprendizaje:

$$(25) \quad p_i = \arg \min_{p \in C_i} \left\{ \sum_{q \in C_i} D_i(p \cdot q) \right\}$$

c) El módulo de representación vectorial utiliza el mismo algoritmos de alineamiento temporal por PD sin restricción de pendiente que el utilizado en la fig. 1; es decir, producciones = $\{(1,0)(1,1)(0,1)\}$; pesos = $\{1,2,1\}$, respectivamente.
d) Como introdujimos en el apartado 2.3, un objeto x, va a tener varias representaciones vectoriales $\vec{y}_j$ j=1...c y escogeremos una u otra dependiendo de la clase sobre la que se vaya a obtener la representación vectorial.

El módulo de aprendizaje corresponde a la implementación del algoritmo (15) a partir de las representaciones vectoriales de cada muestra devueltas por el módulo anterior.

El módulo de reconocimiento utiliza la máquina lineal resultante del proceso de aprendizaje y tiene implementado un clasificador (de distancia mínima) que clasifica una muestra desconocida a partir de sus representaciones vectoriales.

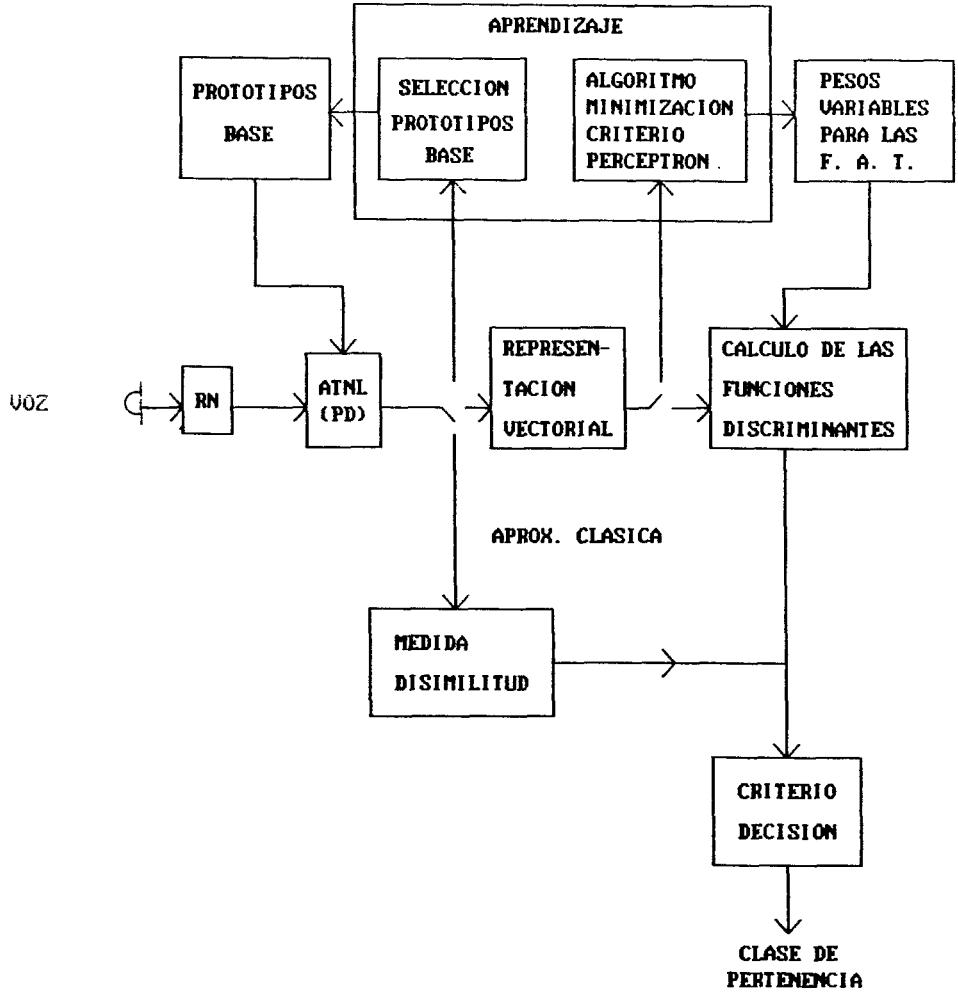


FIGURA 4: Diagrama de bloques del sistema experimental utilizado.

5. RESULTADOS EXPERIMENTALES

Para comprobar experimentalmente el sistema de aprendizaje-reconocimiento propuesto en este trabajo, se ha utilizado un diccionario, de los clasificados como difíciles, compuesto con los nombres de algunas consonantes: /efe/,/ele/,/eme/,/elle/,/ene/,/eñe/,/ere/,y /ese/.

El conjunto de muestras estuvo formado por 21 representaciones de cada palabra del diccionario pronunciadas por un único locutor y utilizando un micrófono de proximidad. Estas adquisiciones fueron parametrizadas según se ha expuesto en el apartado 4.

Con este corpus se llevaron a cabo 8 series de experimentos cuyas características comunes fueron:

1. Los valores del factor de corrección: 0.005, 0.01, 0.05.
2. Los márgenes utilizados: 0 y 3.

Las 8 series de experimentos se dividen en dos grupos. El primer grupo (Exp-1 a 4) se caracteriza por utilizar como prototipos base los correspondientes a la primera repetición de cada palabra:

Exp-1 Aprendizaje: la segunda a cuarta repeticiones de cada palabra.
Incremento: constante.
Prototipo base: el correspondiente a la primera repetición.
Reconocimiento: las 17 repeticiones restantes.
Resultados: tabla 5.1.a.

Exp-2 Aprendizaje: las doce primeras repeticiones de cada palabra.
Incremento: igual que exp-1.
Prototipo base: igual que exp-1.
Reconocimiento: las 8 repeticiones restantes.
Resultados: tabla 5.1.b.

Exp-3 Aprendizaje: igual que exp-1.
Incremento: variable.
Prototipo base: igual que exp-1.
Reconocimiento: igual que exp-1.
Resultados: tabla 5.2.a.

Exp-4 Aprendizaje: igual que exp-2.
Incremento: variable.
Prototipo base: igual que exp-1.
Reconocimiento: igual a exp-2.
Resultado: tabla 5.2.b.

La tasa de reconocimiento, mediante una aproximación clásica (pesos de las producciones fijos) y utilizando como prototipos los mismos que los de los experimentos 1 a 4, fue del 39%.

La serie de experimentos Exp-5 a Exp-8 son idénticos a los Exp-1 a Exp-4 respectivamente, salvo que los prototipos son en este caso los centroides.

Exp-5 Aprendizaje, Incremento y Reconocimiento: igual que en el exp-1
Prototipo base: centroide.
Resultados: tabla 5.3.a.

Exp-6 Aprendizaje, Incremento, Reconocimiento: como exp-2.
Prototipo base: centroide.
Resultados: tabla 5.3.b.

Exp-7 Aprendizaje, Incremento, Reconocimiento: como exp-3.
Prototipo base: centroide.
Resultados: tabla 5.4.a.

Exp-8 Aprendizaje, Incremento, Reconocimiento: como exp-4.
Prototipo base: centroide.
Resultados: tabla 5.4.b.

La tasa de reconocimiento mediante la aproximación clásica comentada anteriormente y utilizando como prototipos los centroides, fue del 14%.

6. CONCLUSIONES

En este trabajo se ha presentado un método que permite tratar el problema de Reconocimiento de Palabras Aisladas con ciertos diccionarios difíciles.

El método descrito en este trabajo está relacionado con las ideas de otro recientemente propuesto /7/. Pero en nuestro caso se ha modificado el concepto de frontera de decisión, obteniendo, en el aprendizaje, unas funciones discriminantes que permiten una mayor separabilidad entre clases.

Los resultados experimentales obtenidos confirman en primer lugar los resultados previos /7/ (margen = 0), aportando aún mejores tasas de reconocimiento siempre y cuando se utilicen un número suficiente de muestras en el aprendizaje, el incremento sea variable y el factor de corrección suficientemente pequeño (menos de 0.01).

En segundo lugar, la modificación del concepto de frontera de decisión (que en lugar de ser una "hipersuperficie" es un "hipervolumen"), permite reducir la tasa de error (margen = 3).

En tercer lugar, la utilización del factor de escala variable en el aprendizaje da lugar a tasas de error inferiores, en general, a la utilización del factor de escala constante.

Por último solo destacar un hecho bastante conocido y confirmado en estos experimentos: el utilizar centroides como prototipos produce mejores resultados que si se utilizan como prototipos muestras escogidas al azar.

El método propuesto en este trabajo no supone un aumento significativo de la complejidad temporal en el reconocimiento. El número de operaciones extra a realizar por prototipo es (tamaño(patrón) + tamaño(muestras)) sumas y productos con las producciones utilizadas, frente a (tamaño(patrón) . . tamaño (muestras)) necesario para realizar el alineamiento temporal. En cambio, para factores de escala muy pequeños, la fase de aprendizaje puede necesitar mucho tiempo de cálculo.

En cuanto a la complejidad espacial, el aumento que necesita este método tampoco es considerable. Esto supone (3 . tamaño(patrón)).

En la actualidad existen varios proyectos, en la linea del presentado en este trabajo, con el objetivo de reducir aún más las tasas de error obtenidas. Otros, están en estudio para la utilización de técnicas de aprendizaje más potentes, aunque más costosas, como por ejemplo las basadas en mínimos cuadrados.

TABLA 5.1.a: Resultados del Experimento 1: Porcentaje de error.

MARGEN	INCREMENTO		
	0.05	0.01	0.005
0	35	20	30
3	55	25	20

TABLA 5.1.b: Resultados del Experimento 2: Porcentaje de error.

MARGEN	INCREMENTO		
	0.05	0.01	0.005
0	25	20	20
3	20	15	15

TABLA 5.2.a: Resultados del Experimento 3: Porcentajes de error.

MARGEN	INCREMENTO		
	0.05	0.01	0.005
0	45	25	30
3	45	20	15

TABLA 5.2.b: Resultados del Experimento 4. Porcentajes de error.

MARGEN	INCREMENTO		
	0.05	0.01	0.005
0	20	15	20
3	15	15	10

TABLA 5.3.a: Resultados del Experimento 5: Porcentaje de Error.

MARGEN	INCREMENTO		
	0.05	0.01	0.005
0	30	25	10
3	30	20	14

TABLA 5.3.b: Resultados del Experimento 6: Porcentaje de Error.

MARGEN	INCREMENTO		
	0.05	0.01	0.005
0	20	15	10
3	20	10	10

TABLA 5.4.a: Resultados del Experimento 7: Porcentaje de error.

MARGEN	INCREMENTO		
	0.05	0.01	0.005
0	35	20	10
3	30	15	10

TABLA 5.4.b: Resultados del Experimento 8: Porcentaje de error.

MARGEN	INCREMENTO		
	0.05	0.01	0.005
0	11	6	9
3	14	6	6

7. BIBLIOGRAFIA

/1/ R.K.MOORE: "Systems for Isolated and Connected Word Recognition" en "New Systems and Architectures for Automatic Speech Recognition an Synthesis". Springer Verlag 1985.

/2/ F.CASACUBERTA y E.VIDAL: "Reconocimiento automático del habla" Marcombo, 1987.

/3/ R.K.MOORE, M.J.RUSSELL, M.J.TOMLINSON: "Locally constrained Dynamic Programming in Automatic Speech Recognition". ICASSP-82, pp.1270-1273. 1982.

/4/ R.K.MOORE, M.J.RUSSELL, M.J.TOMLINSON: "The Discriminative Network: A Mechanism for focusing Recognition in Whole-Pattern Matching". IEEE. Trans. on ASSP. pp. 1041-1044. 1983.

/5/ L.RABINER, J.G.WILPON: "Isolated Word Recognition using a two-pass Pattern Recognition Approach". ICASSP-81 pp. 724-727.

/6/ R.O.DUDA y P.E.HART: "Pattern Classification and Scene Analysis". Wiley. 1973.

/7/ A.FERNANDEZ, P.RODRIGUEZ, E.VIDAL y H.RULOT: "Reconocimiento de Palabras muy semejantes mediante un Método de Aprendizaje basado en el Criterio Perceptrón". Pendiente de publicación.

/8/ J.M.BENEDI, F.CASACUBERTA y E.VIDAL: "Un nuevo nivel de etiquetado microfonético difuso para un sistema multinivel difuso de Reconocimiento Automático del habla". Pendiente de publicación.