

# DIMENSIONALIDAD INTRÍNSECA DE UN CONJUNTO DE PALABRAS AISLADAS

\*E. BAYDAL, \*G. ANDREU, \*H. RULOT, \*\*E. VIDAL  
UNIV. DE VALENCIA

*La Dimensionalidad Intrínseca (DI) de un conjunto de objetos hace referencia al número mínimo de parámetros necesarios para la "generación" de dichos objetos. La estimación de la DI aplicada al caso de pronunciaciones de palabras presenta interés porque permite obtener una medida escalar conveniente de la dificultad de las distintas tareas de Reconocimiento de Palabras Aisladas. Debido a que el conjunto de objetos a tratar (palabras pronunciadas) no presenta estructura de espacio vectorial, el método escogido para el cálculo de la DI ha sido el propuesto por K.W. Pettis et al., el cual basa la estimación únicamente en la información proporcionada por las disimilitudes (distancias) entre cada objeto y sus k-vecinos más próximos.*

**Keywords:** Intrinsic Dimensionality Estimation, Metric Spaces, Isolated Word Recognition, Pattern Recognition.

## 1. INTRODUCCIÓN

En Reconocimiento Automático del Habla los objetos considerados (palabras o frases pronunciadas) se suelen representar mediante *sucesiones* de vectores de parámetros (típicamente de dimensión 8-16), donde cada vector representa las características de la señal vocal en un intervalo de análisis de duración fija (típicamente 10-20 ms.) /3/. Así pues, como las longitudes de las sucesiones de estos vectores representan la duración total de la señal analizada, éstas pueden variar, no sólo de una palabra (o frase) a otra, sino incluso entre las distintas pronunciaciones de una misma palabra (o frase).

La información contenida en este modo de representación es sustancialmente inferior a la que posefa originalmente la representación directa de la señal en el dominio temporal. A pesar de ello, los vectores de parámetros normalmente utilizados suelen garantizar la conservación de la información psicoacústica más relevante, de forma que una "parametrización inversa" conduce a una señal acústica inteligible por un oyente humano.

No obstante, si nos ceñimos al Reconocimiento de Palabras Aisladas (RPA) /3/, la representación arriba indicada puede resultar aún altamente redundante. En el marco del Reconocimiento Geométrico de Formas (también llamado Reconocimiento basado en la Teoría de la Decisión) existen diversos métodos clásicos para analizar la redundancia de una representación (ver, por ej., /10/), aunque para ello es necesario asumir que dicha re-

• \*E. Baydal, \*G. Andreu, \*H. Rulot, \*\*E. Vidal

\*Centro de Informática de la Univ. de Valencia - \*\*Centro de Informática de la U.P.V. Camino de Vera, s/n. Valencia.

• Article rebut el novembre de 1986.

presentación es de tipo *vectorial*. Dada la longitud variable (y por tanto el número variable de parámetros) de la representación de las palabras pronunciadas, dicha representación puede difícilmente considerarse una representación vectorial /5/. Sin embargo, sí existe(n) Medida(s) de Disimilitud adecuada(s) entre palabras pronunciadas /3/, lo que, en la mayoría de los casos, confiere *estructura métrica* al espacio de representación disponible /5/ y /7/. Intuitivamente, parece claro que la redundancia de la representación de un conjunto de palabras pronunciadas debe depender no sólo del tipo de representación adoptada, sino también de las palabras realmente incluidas en dicho conjunto. Así por ejemplo, si se trata de un conjunto de pronunciaciones de un diccionario compuesto por tan sólo 2 palabras (por ej. "sí", "no") cabe esperar que la redundancia sea alta, o lo que es lo mismo, que el número de parámetros *intrínsecamente* necesarios para representar los objetos de un universo tan limitado, sea muy pequeño.

Cuando se dispone de una representación vectorial, la estimación del "número mínimo de parámetros intrínsecamente necesarios" recibe el nombre de estimación de la "*dimensionalidad intrínseca*" del conjunto de los objetos considerados. En este caso, se pueden aplicar directamente algunas de las técnicas de reducción de la redundancia (reducción de la dimensionalidad) arriba referidas. Sin embargo, aún cuando los objetos no vengán representados vectorialmente, existen métodos para realizar una estimación del número de parámetros intrínsecamente necesarios, lo que por extensión se denomina, análogamente, estimación de la dimensionalidad intrínseca. La dimensionalidad intrínseca de un conjunto de objetos puede interpretarse también como el número mínimo de parámetros necesarios para la "generación" de dichos objetos. Por ejemplo, los puntos distribuidos a lo largo de una hélice en un espacio de tres dimensiones pueden ser expresados en función de un único parámetro, con lo que su dimensionalidad intrínseca sería uno.

La estimación de la dimensionalidad intrínseca de un conjunto de pronunciaciones de palabras tiene interés en sí misma ya que permite obtener una conveniente medida escalar de la dificultad de las distintas tareas de RPA (dificultad del vocabulario, dificultad del conjunto de locutores en RPA multilocutor, etc.). Además, hay algunos temas en los que dicha estimación puede tener un interés indirecto. Citaremos dos de estos temas.

El primero está relacionado con un algoritmo recientemente propuesto ("AESAS") para reducir el número de comparaciones (cálculos de la Medida de Disimilitud) requerido en la búsqueda del vecino más próximo normalmente utilizada en RPA /5/ /6/ /8/. En un espacio vectorial de dimensión  $d$ , este algoritmo presenta una complejidad temporal cuya cota inferior absoluta podría estimarse, para cualquier talla del diccionario, en  $d+1$  comparaciones entre la muestra a reconocer y  $d+1$  prototipos /6/. Una estimación de la dimensionalidad intrínseca de un conjunto de muestras vocales nos aportaría, en este caso, información sobre cuan ajustados están los resultados experimentales obtenidos para dichas muestras /5/ /8/ con respecto a la máxima eficacia alcanzable con el AESAS.

El segundo tema está relacionado con la posibilidad de encontrar realmente una representación de tipo vectorial para la pronunciación de palabras de cierto vocabulario. En este sentido podría ser útil la formulación recientemente propuesta por L. Goldfarb /2/, según la cual un conjunto de objetos cuyas disimilitudes (distancias) cruzadas son conocidas, puede representarse en un espacio "pseudoeuclídeo" (un espacio de Minkowski modificado) isométrico con el espacio (métrico) original; es decir, conservando los

valores originales de disimilitud entre objetos. Si se aplica esta formulación al problema que nos ocupa (RPA), se presenta la dificultad de que la dimensión del espacio pseudoeuclídeo resultante es generalmente  $n+1$ , donde  $n$  es el número de objetos (palabras pronunciadas) considerados. Aunque en /2/ se sugieren métodos similares a los clásicos para reducir dicha dimensionalidad, desafortunadamente, estos métodos requieren de la introducción de ciertos umbrales para cuya estimación no se dispone de métodos precisos. En este caso, una estimación previa de la "dimensionalidad intrínseca" podría hacer innecesaria la introducción de los umbrales indicados, pudiendo procederse directamente a reducir la dimensionalidad a su valor más conveniente.

Entre los métodos propuestos para la estimación de la dimensionalidad intrínseca de un conjunto de objetos, tan sólo son utilizables en nuestro caso aquellos en los que no se hace uso de estructura vectorial alguna en dicho conjunto. Esto reduce drásticamente el número de métodos utilizables, siendo el propuesto por K.W.Pettis et.al. /1/ prácticamente el único que puede ser aplicado a nuestro problema. Este método se basa únicamente en la información suministrada por la densidad de vecinos más próximos alrededor de cada objeto (punto); consecuentemente, tan sólo una medida de disimilitud entre objetos es necesaria, y la aplicación al problema que nos ocupa es inmediata.

## 2. BASE MATEMATICA

Dado un conjunto de puntos  $(X_1, \dots, X_n)$  en un espacio  $L$ -dimensional distribuidos según una densidad desconocida  $p(\cdot)$ , se va a calcular su dimensionalidad intrínseca  $d$  mediante una aproximación  $\hat{d}$  basada en el cálculo de una estimación de  $p(x)$  dada por:

$$\hat{p}(x) = \frac{k/n}{V} \tag{1}$$

donde  $k$  es el número de vecinos más próximos a  $x$  (con respecto a cierta métrica) dentro de la esfera centrada en  $x$  de radio  $R_k$  y volumen  $V = V_d (R_k)^d$ , y  $V_d$  es el volumen de una hiperesfera unidad en dimensión  $d$ , el cual viene dado por:

$$V_d = \frac{(\pi)^{d/2}}{\pi \left(\frac{d}{2} + 1\right)}$$

Sustituyendo en (1) la expresión de  $V$  y tomando logaritmos se obtiene:

$$\log(R_k) = (1/d) \log(k) + \log \left[ (n V_d \hat{p}(x))^{-1/d} \right] \tag{2}$$

Si en el último término en (2) fuera independiente de  $k$ , habría una relación lineal entre  $\log(k)$  y  $\log(R_k)$  con una pendiente de  $(1/d)$  y podría usarse (2) para calcular  $d$ . Pero  $\hat{p}(x)$  no es independiente de  $k$ , y  $R_k$  no está unívocamente determinado, con lo que la resolución de (2) se hace difícil. Se usará una ecuación similar a (2) que permita obtener  $d$  /1/.

Se define una distancia media sobre el conjunto de objetos al k-ésimo vecino más cercano como:

$$\bar{r}_k = (1/n) \sum_{i=1}^n r_{k x_i}$$

donde  $r_{k x_i}$  es la distancia desde  $x_i$  a su k-ésimo vecino más próximo.

Se puede demostrar que, bajo ciertas condiciones /1/, la esperanza de  $\bar{r}_k$  viene dada por:

$$E(\bar{r}_k) = (1/n) \sum_{i=1}^n E(r_{k x_i}) = \frac{1}{G_{kd}} k^{1/d} C_n \quad (3)$$

donde

$$G_{kd} = \frac{k^{1/d} \Gamma(k)}{\Gamma(k+1/d)}$$

y

$$C_n = (1/n) \sum_{i=1}^n \left[ n p(x_i) V_d \right]$$

Aunque  $C_n$  depende del conjunto de puntos, es independiente de k. Tomando logaritmos en (3) se obtiene una ecuación similar a (2):

$$\log(G_{kd}) + \log E(\bar{r}_k) = (1/d) \log(k) + \log(C_n) \quad (4)$$

El término  $\log(G_{kd})$  aunque no es independiente de k se aproxima a cero para todo k y todo d/1/. Se tomará  $E(\bar{r}_k) \simeq \bar{r}_k$  y  $d \simeq \hat{d}$  en (4), obteniéndose:

$$\log(G_{k\hat{d}}) + \log(\bar{r}_k) = (1/\hat{d}) \log(k) + \log(C_n) \quad (5)$$

A partir de (5) se puede obtener  $\hat{d}$  (y, si se desea,  $C_n$ ) mediante el siguiente algoritmo iterativo:

La estimación inicial  $\hat{d}_0$  se obtiene asumiendo  $\log(G_{k\hat{d}}) = 0$  y ajustando por mínimos cuadrados una recta de  $\log(\bar{r}_k)$  en función de  $\log(k)$ , desde  $k=1..K$ , siendo K el número máximo de vecinos tomados. Con el valor obtenido se calcula  $\log(G_{k\hat{d}_0})$  y se vuelve a ajustar la ecuación (5) para obtener  $\hat{d}_1$ . Se continua hasta alcanzar un i para el cual  $|\hat{d}_i - \hat{d}_{i-1}| < \epsilon$ , para un  $\epsilon$  dado. La estimación es  $d = \hat{d}_i$ . En la práctica  $\hat{d}_i$  se redondea al entero más próximo.

La siguiente aproximación para el primer término de la expresión (5) está basada en el desarrollo en serie de Taylor para el logaritmo de la función gamma:

$$\log(G_{kd}) = \frac{\hat{d}-1}{2 k \hat{d}^2} + \frac{(\hat{d}-1)(\hat{d}-2)}{12 k^2 \hat{d}^3} - \frac{(\hat{d}-1)^2}{12 k^3 \hat{d}} \quad (6)$$

Teniendo en cuenta la ecuación de la pendiente de una recta obtenida mediante un ajuste por mínimos cuadrados, se ha obtenido para  $\hat{d}_j$  la expresión:

$$\hat{d}_j = \left[ \frac{k \sum_{k=1}^K (\log k) (F_j(k)) - \left( \sum_{k=1}^K \log k \right) (F_j(k))}{k \sum_{k=1}^K (\log k)^2 - \left( \sum_{k=1}^K \log k \right)} \right] \quad (7)$$

donde  $F_j(k) = \log \bar{r}_k + (\log (G_{kd}))_j$

Las expresiones (6) y (7) son las utilizadas en el algoritmo indicado, cuya presentación formal se omite por considerarse evidente.

En el método propuesto originalmente /1/ existe un apartado en el que se eliminan los puntos fronterizos o "discordantes", dicho apartado no se ha implementado ya que el efecto de estos puntos puede reducirse si se eligen adecuadamente los valores de  $K$  y el número de puntos utilizados. No obstante, en algunos casos se espera que puedan aparecer ciertos efectos de bordes.

### 3. RESULTADOS EXPERIMENTALES

Hemos estudiado 2 tipos de datos obtenidos de forma diferente:

- 1) Por simulación.
- 2) Muestras de palabra real.

En todos los casos se presentarán los resultados en función de dos parámetros:  $n$  (número de objetos -puntos- considerados) y  $K$  (número máximo de vecinos más próximos). Asimismo se ha elegido  $\epsilon = 0.01$ , con lo que el número de iteraciones producidas por el algoritmo ha sido siempre inferior a cinco. Para valores de  $K$  próximos a  $n$  aparecen siempre efectos de bordes debido a lo cual la discrepancia entre  $d$  y  $\hat{d}$  llega a ser grande. Este efecto es especialmente evidente para pequeños valores de  $n$ , en los cuales los bordes adquieren una gran influencia.

#### 3.1. Simulación.

En este apartado realizaremos el estudio de un conjunto de puntos de dimensionalidad conocida, para comprobar el buen funcionamiento del método.

Los puntos han sido escogidos de forma aleatoria en una hélice descrita por la ecuación:

$$x = \cos \sigma ; y = \sin \sigma ; z = 0.1(\sigma) ; \text{ con } 0 \leq \sigma \leq 4.$$

Obviamente, la dimensionalidad intrínseca de este conjunto de puntos es uno. Hemos generado tres conjuntos diferentes de puntos, de tallas 20, 50 y 100 y a estos conjuntos se les ha aplicado el método propuesto utilizando la métrica  $L_2$ (euclídea). De esta forma se han obtenido diversas estimaciones de  $d$ . Con estas estimaciones se ha construido una gráfica (Fig.1), representándose  $\hat{d}$  en función de  $k$ , y una curva para cada  $n$  (20,50,100). Para evitar los efectos de bordes no se han utilizado valores de  $K$  superiores a 19. La estimación obtenida para la dimensionalidad comienza a ser buena a partir de  $n \geq 50$ .

La dimensionalidad que se ha obtenido coincide con la esperada. Los resultados globales son muy similares a los presentados por W.Pettis et.al. /1/, lo que corrobora el buen funcionamiento del método.

### 3.2. Experiencias con muestras de palabra real.

Entre la amplia gama de experimentos que se pueden realizar sobre la dimensionalidad de un conjunto de palabras, vamos a estudiar únicamente dos casos:

A) Palabras correspondientes a un diccionario de nombres de plantas aromáticas y medicinales. Se trata de 200 palabras pronunciadas una única vez por un mismo locutor (masculino), y parametrizadas mediante 16 filtros pasa-banda distribuidos según la escala Mel /4/. Las distancias entre palabras se miden mediante un procedimiento usual de alineamiento temporal no lineal /3/ /7/, aplicándose convenientemente al algoritmo de estimación propuesto. Los resultados se presentan en la (Fig.2) donde se representa  $\hat{d}$  en función de  $K$ . La estimación se estabiliza para un valor de  $\hat{d}$  entre 14 y 15.

B) Palabras correspondientes a varias pronunciaciones de los diez dígitos castellanos. Se han considerado cuatro conjuntos, tomando en cada una de ellos diferente número de locutores, de distinto sexo, y variando el número de repeticiones de cada dígito. Las estimaciones de la dimensionalidad obtenidas para cada conjunto se han representado en una misma gráfica (fig.3), donde el eje de abscisas corresponde al número de vecinos  $K$ , y el de ordenadas a  $d$ . Las distintas curvas representadas corresponden a: A) 3 locutores femeninos y 6 repeticiones (180 palabras); B) 3 locutores masculinos y 6 repeticiones (180 palabras); C) 6 locutores mixtos y 3 repeticiones (180 palabras); D) 10 locutores mixtos y 2 repeticiones (200 palabras).

Los efectos de bordes adquieren gran importancia para valores de  $K$  inferiores a 19, por lo que se ha omitido la representación de éstos. A partir de  $K=49$  la estimación de  $\hat{d}$  se estabiliza, obteniéndose  $d=5$  para los datos mixtos y  $d=4$  para los otros dos.

Se observa que la dimensión aumenta de forma directa al número de locutores e inversamente al de repeticiones. Para los datos correspondientes a locutores de un mismo sexo se ha obtenido una dimensión ligeramente menor que para el caso mixto, lo que es debido a que en el caso mixto hay mayor variación en la forma de pronunciar las palabras.

Si se comparan estos resultados con los obtenidos en el experimento anterior (A), se observa claramente una mayor dimensionalidad en el caso anterior. Este resultado es consistente con la mayor variabilidad en la for-

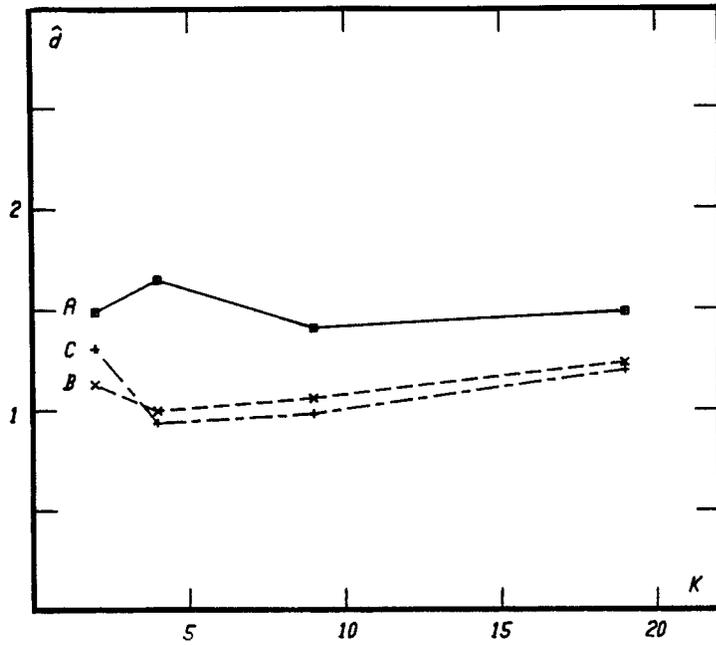


Fig.1. Estimación de la dimensionalidad de un conjunto de puntos (N) correspondientes a una hélice: A. N=20, B. N=50, C. N=100.

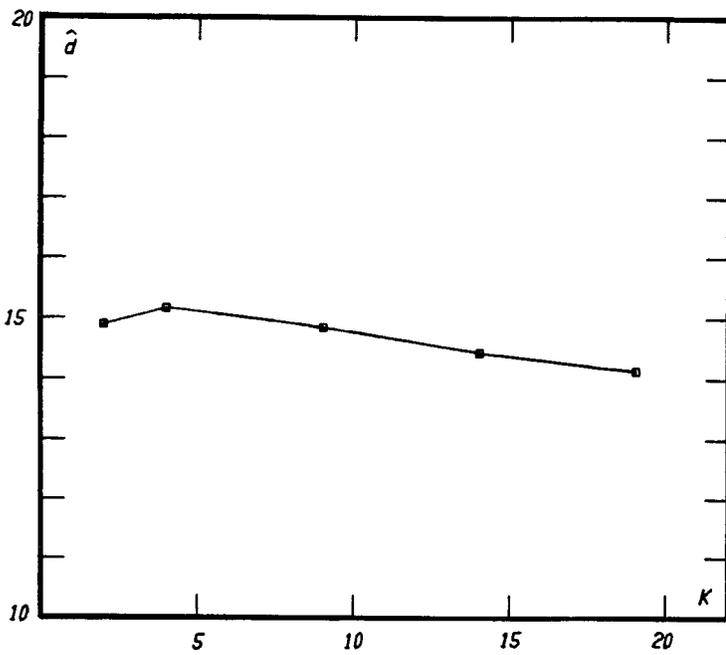


Fig. 2. Estimación de la dimensionalidad de un conjunto de 200 palabras correspondientes a una repetición de un diccionario de hierbas aromáticas pronunciadas por un único locutor.

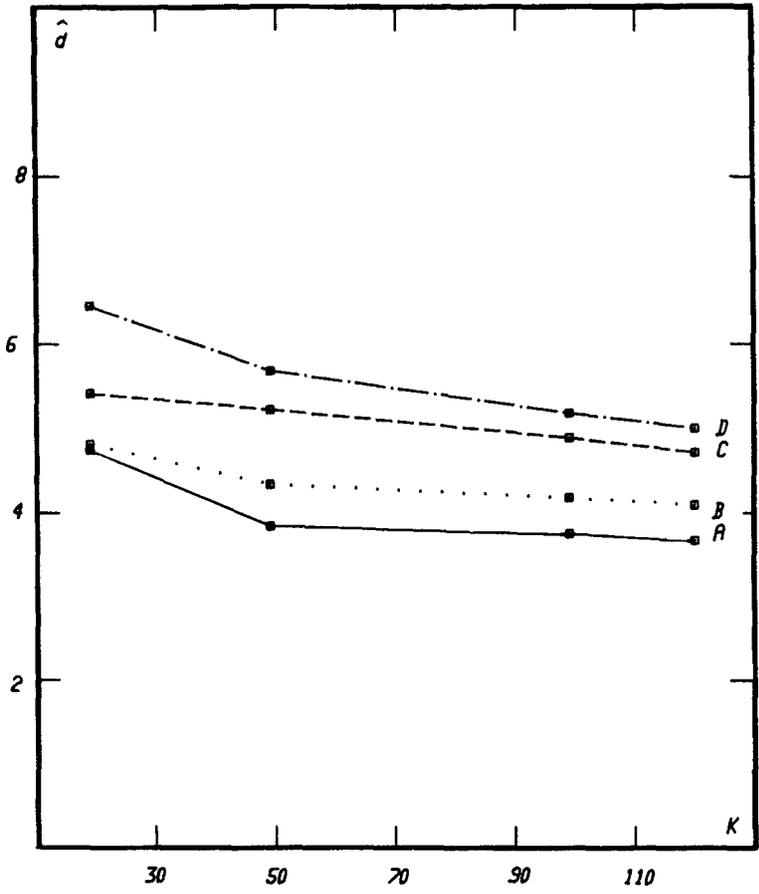


Fig. 3. Estimación de la dimensionalidad de un conjunto de palabras aisladas compuesto por distintas repeticiones de los dígitos castellanos: A. 3 locutores femeninos y 6 repeticiones. B. 3 locutores masculinos y 6 repeticiones. C. 6 locutores mixtos y 3 repeticiones. D. 10 locutores mixtos y 2 repeticiones.

ma de los objetos (200 palabras diferentes), en (A), y la mayor homogeneidad de estas formas en el caso de los dígitos (10 palabras diferentes).

#### 4. CONCLUSIONES

Para estimar la dimensionalidad intrínseca de un conjunto de muestras vocales, se ha propuesto el uso de un método /1/ originalmente desarrollado para realizar dicha estimación en espacios (pseudo)métricos. Este método se ha implementado, y aplicado a diversos conjuntos de palabras pronunciadas aisladamente. Los resultados obtenidos arrojan valores de dimensionalidad comprendidos entre 4 y 15, los cuales son consistentes con la idea intuitiva del grado relativo de dificultad asociado a cada uno de los conjuntos considerados.

Por otra parte, es interesante destacar la gran correlación existente entre estos resultados, y el número de comparaciones medio requerido por el algoritmo de búsqueda AESA /6/ para los mismos conjuntos de datos /8/ /9/. Según se discute en /6/, en un espacio vectorial de dimensión  $d$  este número de comparaciones debe ser siempre (no muy) superior a  $d+1$ , mientras que si tomamos como ciertos los valores de dimensionalidad obtenidos en el trabajo aquí presentado, los números medios de comparaciones obtenidos en /8/ y /9/ resultan excederla en aproximadamente 2-5 unidades. Esta ajustada coincidencia indica pues, tanto la optimalidad (con respecto a la máxima eficacia alcanzable con el AESA) de los resultados obtenidos en /8/ y /9/, como la fiabilidad de los métodos introducidos en este trabajo.

## BIBLIOGRAFIA

- /1/ Pettis, K.W., etc..., "An intrinsic dimensionality estimator from near-neighbor information", IEEE Trans. PAMI, vol.pa 41-1, NO.1, january 1979.
- /2/ Goldfarb, L., "A unified approach to pattern recognition", Pattern Recognition, vol. 17, NO.5, pp. 575-582, 1984.
- /3/ Casacuberta, F. y E.Vidal, "Reconocimiento automático del habla". Ed. Marcombo, En prensa, 1987.
- /4/ Benedí, J.M., F.Casacuberta, E.Vidal, "Un nuevo nivel de etiquetado microfonético difuso para un sistema multinivel difuso de reconocimiento automático del habla". Rev. Informática y Automática. Pendiente de publicación, 1987.
- /5/ Vidal, E., F.Casacuberta, H.Rulot, "Is the DTW "distance" really a metric? -An algorithm reducing the number of DTW comparisons in Isolated Word Recognition". Speech Communication, NO.4, pp.333-334.
- /6/ Vidal, E., "An Algorithm for finding nearest neighbours in (approximately) constant average time". Pattern Recognition Letters, NO.4, pp. 145-157, 1986.
- /7/ Vidal, E., F.Casacuberta, H.Rulot, J.Benedí, M.J.Lloret, "On the verification of the triangle inequality by DTW dissimilarity measures". Speech Communication. Pendiente de Publicación, 1987.
- /8/ Vidal, E., H.Rulot, F.Casacuberta, J.Benedí, "On the Use of a Metric-Space Search Algorithm (AESAs), for Fast DTW-Based Recognition of Isolated Words", A publicar en IEEE Trans. Acoustica Speech and Signal Proc. (1987).
- /9/ Lloret, M.J., "Aplicación de un algoritmo de búsqueda en espacios métricos al Reconocimiento de Palabras Aisladas multilocutor", Tesina Lic., Univ. Valencia, 1986.
- /10/ Duda, R.O., P.E.Hart, "Pattern Classification and Scene Analysis". Ed. J.Wiley and sons, New York, 1973.

