Reducción de espacios de entrenamiento de HMMS empleando DPCA

Training space reduction for HMMS employing DPCA

Johanna Carvajal- González[1] Ing., Milton Orlando Sarria Paja[1] Ing., Álvaro Ángel Orozco Gutiérrez[2] MSc., Germán Castellanos Domínguez[1] PhD.
1. Grupo de Control y Procesamiento Digital de Señales, Universidad Nacional de Colombia-Manizales
2. Grupo de Control e Instrumentación, Universidad Tecnológica de Pereira
jpcarvajalg@unal.edu.co, mosarriap@unal.edu.co, aaog@utp.edu.co, cgcastellanosd@unal.edu.co

Recibido para revisión: 5 de Octubre de 2007, Aceptado: 28 de Noviembre de 2008, Versión final: 15 de Diciembre de 2008

Resumen—**Se desarrolla una metodología de reducción de espacios de entrenamiento, basado en la extracción dinámica de características, para obtener un mayor desempeño en un sistema de clasificación con un costo computacional bajo, empleando DPCA, técnica de análisis que muestra ser eficiente en la selección de las componentes principales que tienen mayor influencia en el desempeño del sistema de clasificación con problemas de alta dimensionalidad. Para los experimentos se emplea un sistema de reconocimiento basado en modelos ocultos de Markov (Hidden Markov Models – HMMs), para modelar la dinámica estocástica presente en señales acústicas y en señales MER (Micro Electrode Recording). Como resultado se obtiene un mejor rendimiento del sistema de clasificación con un conjunto de características reducido, en comparación con el desempeño que presenta el sistema cuando se emplea el espacio de características completo.**

Palabras Clave—**Características dinámicas, HMM, MER, PCA, Reconocimiento de patrones, Selección de características dinámicas, Voces patológicas**

Abstract—**A methology to reduce training spaces based on the dynamic features extraction is developed. The methodology is aimed at improving the performance of a classification system with low computational cost by using DPCA (dynamic principal components analysis). This analysis technique has shown to be efficient in the selection of principal components that expose a greater influence in the classification performance on high dimensionality problems. A recognition system based on Hidden Markov Models (HMMs) is employed to model the stochastic dynamics present in both acoustic and MER(Micro Electrode Recording) signals. As result, a better rate of classification is obtained by the system with reduced dimensionality.**

Keywords—**Dynamic features, Dynamic feature selection, HMM,MER, Pathological voices, Pattern recognition, PCA**

I. INTRODUCCIÓN

En los últimos años se ha mostrado que las señales biomédicas se describen mejor mediante un conjunto de características que varían temporalmente [1] [2], a diferencia de otras tareas de reconocimiento de patrones, en las cuales los objetos a clasificar se pueden describir usando características estáticas. Esta variación temporal hace que las propiedades estadísticas cambien, y es precisamente esa variabilidad estocástica la que contiene la información necesaria para la clasificación [3], en este sentido se hace necesario usar una técnica que permita modelar la dinámica estocástica que presentan este tipo de señales.

Los Modelos Ocultos de Markov (HMMs) son una clase de procesos estocásticos que permiten modelar series de tiempo y han sido empleados en el procesamiento de secuencias de datos temporales aplicados entre otras tareas para modelar las variaciones en señales acústicas (reconocimiento de voz) y para a la detección de patologías de voz [2], también en la identificación de firmas temporales, en electrocardiografía (ECG) [4], en patrones de emisión cerebral mediante el análisis en encefalograma (EEG) [5] y en el diagnóstico de enfermedades cardíacas usando registros fonocardiográficos (PCG) [6].

De otra parte, la extracción de características se emplea como una forma de representación del espacio inicial de entrenamiento, a objeto de mejorar sus cualidades de discriminancia. Las técnicas de análisis de componentes, por ejemplo PCA (*Principal Component Analysis*), son un caso de amplio uso, en que el espacio trasformado se forma de elementos ortogonales y aunque no se considera ningún modelo de aleatoriedad, en la estimación de las respectivas matrices de

covarianza, es conocido que la efectividad de la técnica es más alta entre más se ajuste la condición de Gaussividad para las distribuciones de las características, justo para el caso que los modelos de Markov han sido desarrollados [7].

En este trabajo se presenta una metodología para analizar la dinámica oculta en señales acústicas y señales MER (Micro Electrode Recording). La caracterización de las señales de voz se realiza usando *NNE*, *HNR*, *MFCC*, la relación *GNE* y la energía medida por trama de la señal, además de las primeras derivadas de cada uno de los contornos. La caracterización de señales MER aplicada al reconocimiento de zonas cerebrales se realiza utilizando esquemas de actualización adaptativos con umbrales autoajustables [8]. Por otro lado, se emplea la metodología propuesta en [9], con el fin de identificar las características acústicas dinámicas que contribuyen de mayor forma en el proceso de clasificación, con las cuales se entrena un clasificador basado en HMMs, para realizar una comparación y evaluación entre los conjuntos de características y determinar cuales son las mas adecuadas para un tarea especifica.

II. MODELOS MATEMÁTICOS

A. Modelos ocultos de Markov

Una cadena de Markov es un proceso aleatorio $\theta(t)$ que puede tomar una cantidad finita K de valores discretos dentro del conjunto $\{\vartheta_1,\cdots,\vartheta_K\}$, tal que en los momentos determinados del tiempo $(t_o < t_1 < t_2 < \cdots)$ los valores del proceso aleatorio cambien (con probabilidades de cambio conocidas), esto es, se efectúan los cambios en forma de secuencia aleatoria $\theta_0 \to \theta_1 \to \theta_2, \cdots$, siendo $\theta_n = \theta(t_n)$ el valor de la secuencia después del intervalo n de tiempo. Las cadenas de Markov asumen una cantidad finita de valores discretos o estados para la representación de una señal aleatoria. En particular, cada estado de manera directa se asocia a un evento fisico observable. Sin embargo, en la práctica, se tienen aplicaciones con señales que no presentan de forma evidente los eventos sobre los cuales se construye el modelo. En este sentido, se debe construir un modelo probabilística sobre los estados no observables u *ocultos*. Como resultado las cadenas construidas por este principio, corresponden a un proceso estocástico doble incrustado; la función probabilística de los estados ocultos y el mismo modelo de aleatoriedad de Markov impuesto sobre la señal [10].

Un modelo de Markov se denota por $\lambda = \{\boldsymbol{\Pi}, \mathbf{B}, \mathbf{p}_{\theta_1}\}$, igualmente se define el estado en el tiempo t_n como $\theta_n \in \boldsymbol{\vartheta} = \{\vartheta_k : k = 1,\cdots,n_\vartheta\}$, donde n_ϑ es el numero de estados del modelo. La matriz $\boldsymbol{\Pi} = \{\pi_{mn} : m,n = 1,\cdots,n_\vartheta\}$, es la matriz de transición de estados, esta compuesta por las probabilidades discretas π_{mn} que corresponde con la probabilidad:

$$\pi_{mn}(k) = P(\theta_{k+1} = \vartheta_n \mid \theta_k = \vartheta_m) \tag{1}$$

$B = \{b_j(\cdot)\}$ esta asociado con la función densidad de probabilidad de observación de cada uno de los estados, $b_j(\varphi_n)$ corresponde a la probabilidad de producir una observación en particular φ_n en el estado ϑ_j. Existen dos estructuras para describir esta densidad, el primer caso es el discreto, es una forma eficiente de modelar los datos que son naturalmente simbólicos y hacen parte de un libro de códigos de longitud n_v. El segundo caso es el caso continuo b_j es una función paramétrica, comúnmente Gaussiana o una mezcla de Gaussianas, definida como:

$$b_j(\varphi_n) = P(\varphi_n \mid q_t = s_j) = \sum_{m=1}^{M} c_{jm} N(\varphi_n, \mu_{jm}, \Sigma_{jm}) \tag{2}$$

Donde el vector de observación φ_n tiene dimensión P, M es el número de mezclas, el coeficiente de las mezclas c_{jm} es un parámetro que controla la m-esima mezcla, μ_{jm} y Σ_{jm} corresponde al vector de medias y la matriz de covarianzas respectivamente de la m-esima mezcla.

El vector $p_{\theta_1}(i) = P(\theta_1 = \vartheta_1) \quad 1 \le i \le n_\vartheta$ corresponde con la probabilidad inicial de estados.

La probabilidad de que una secuencia de observaciones de longitud n_φ, $\boldsymbol{\varphi} = \{\varphi_1,\cdots,\varphi_{n\varphi}\}$, sea generada por un modelo λ esta definida como:

$$P(\boldsymbol{\varphi} \mid \lambda) = \sum_{\forall \boldsymbol{\theta}} \left(p_{\theta_1} \prod_{n=2}^{n_\varphi} \pi_{\theta_{n-1}\theta_n} \prod_{n=1}^{n_\varphi} b_{\theta_n}(\varphi_n) \right) \tag{3}$$

Donde $\boldsymbol{\theta} = \{\theta_0, \theta_1, \cdots, \theta_{n\theta}\}$ corresponde con una secuencia de estados y el modelo λ esta asociado con una clase en particular.

A.1. Estimación de Máxima Verosimilitud – MLE

El método mas comúnmente empleado en el entrenamiento de de un HMM es la estimación de Máxima Verosimilitud (Maximum Likelihood Estimation- MLE), que trata de ajustar los parámetros de un modelo de tal forma que se maximice la

probabilidad (3) de que las muestras de entrenamiento sean generadas por dicho modelo, esto es equivalente a optimizar la función la objetivo ML.

Considerar un conjunto de R muestras de entrenamiento $\mathbf{Y} = \left\{ \boldsymbol{\varphi}_1^{n\varphi_1}, .., \boldsymbol{\varphi}_r^{n\varphi_r}, .., \boldsymbol{\varphi}_R^{n\varphi_R} \right\}$, y sus categorías o etiquetas, $\hat{C} = \left\{ c^1, .., c^r, .., c^R \right\}$ donde $c^r \in \left\{ C_1, .., C_m, .., C_M \right\}$, y M es el numero total de clases. Cada muestra o registro $\boldsymbol{\varphi}_r^{n\varphi_r}$ es representada por una secuencia de longitud $n\varphi_r$ de vectores de características $\boldsymbol{\varphi}_r^{n\varphi_r} = \left\{ \varphi_{r,1}, .., \varphi_{r,t}, .., \varphi_{r,n\varphi_r} \right\}$. El conjunto total de parámetros de los HMM se denota por Θ y esta compuesto por M modelos, es decir $\Theta = \left\{ \lambda_1, .., \lambda_m, .., \lambda_M \right\}$, donde λ_m denota los parámetros del HMM que representa la categoría o clase C_m. La función objetivo ML se puede escribir como:

$$f_{ML}(\Theta) = \sum_{r=1}^{R} \log\left(P\left(\boldsymbol{\varphi}_r^{n\varphi_r} \middle| c^r \right) \right) = \mathrm{L}\left(\mathbf{Y} \mid \Theta \right) \tag{4}$$

Optimizar la función (4) implica ajustar los parámetros de cada modelo de forma separada con los datos de entrenamiento de cada clase de tal forma que la verosimilitud se maximice.

Para lograr este objetivo se emplea un método iterativo eficiente conocido como el algoritmo de Baum-Welch es la forma convencional de estimar los parámetros de un HMM y una de las más importantes razones para que el entrenamiento basado en MLE sea tan popular, este es esencialmente un algoritmo de optimización iterativo que, bajo ciertas restricciones puede hacer converger los valores de los parámetros a un máximo local de la función de verosimilitud [11].

Si Ω es el espacio de los parámetros, la estimación de máxima verosimilitud, consiste en ajustar el estimado Θ_{ML} tal que:

$$\Theta_{ML} = \arg\max_{\Theta \in \Omega} \mathrm{L}\left(\mathbf{Y} | \Theta \right) \tag{5}$$

El algoritmo primero encuentra el valor esperado del *log*-verosimilitud de los datos completos $\mathrm{L}\left(\mathbf{Y} | \Theta \right)$ con respecto a los datos desconocidos, por medio de los datos observados $\mathbf{Y}$ y de los actuales parámetros estimados. Se define [12]:

$$Q\left(\Theta, \Theta^{(i-1)} \right) = E\left[\log p\left(\mathbf{Y} | \Theta \right) | \, \mathbf{Y}, \Theta^{(i-1)} \right] \tag{6}$$

donde $\Theta^{(i-1)}$ son los parámetros actuales que se emplean para evaluar la esperanza y Θ son los nuevos parámetros que se estiman para incrementar Q. La clave para entender la expresión (6), está en que $\mathbf{Y}$ y $\Theta^{(i-1)}$ son constantes y Θ es una variable que se desea ajustar. El segundo paso (paso M) del algoritmo EM busca maximizar la esperanza calculada en el primer paso. Esto es encontrar:

$$\Theta^{(i)} = \arg\max_{\Theta} \mathbf{Q}\left(\Theta, \Theta^{(i-1)} \right) \tag{7}$$

Estos dos pasos son repetidos cuanto sea necesario. El objetivo es que el algoritmo haga converger los parámetros $\Theta^{(i)} = \Theta_{ML}$. En cada iteración del algoritmo está garantizado el incremento del *log*-verosimilitud; por lo tanto el algoritmo converge a un máximo local de la función de verosimilitud [13].

III. ANÁLISIS DE VARIABLES DINÁMICAS

Aproximaciones ampliamente conocidas como PCA [14][15][16] y métodos de búsqueda secuencial, han sido adaptados como métodos de selección de características para el uso de sistemas de clasificación basados en HMMs. Asumiendo que los datos de un contorno de entrada están altamente correlacionados, métodos que emplean transformaciones lineales tales como PCA tratan de explotar la correlación presente en los datos proyectándolos a un nuevo espacio donde los ejes son ortonormales.

Sea $\xi_{ij}[k], k = 1, \ldots, m$ la *j*-esima característica dinámica que pertenece a la *i*-esima observación, donde $j = 1, \ldots, p$, $i = 1, \ldots, n$; siendo n el numero de observaciones y p el numero de características o variables por observación, las cuales cambian sobre el tiempo *k*. Cada vector de observación $\boldsymbol{\xi}_i$ puede ser representado por un supersector de tamaño $mp \times 1$:

$$\boldsymbol{\xi}_i = \begin{matrix} [\xi_{i1}[1], \xi_{i1}[2], \cdots, \xi_{i1}[m], \xi_{i2}[1], \cdots \\ \cdots, \xi_{i2}[m], \cdots, \xi_{ip}[m]]^{\mathrm{T}} \end{matrix} \tag{8}$$

Después de centrar cada una de los supervectores, la matriz de covarianzas se calcula de la siguiente forma:

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\xi}_i^0 \boldsymbol{\xi}_i^{0T} = \frac{1}{n} \mathbf{G}\mathbf{G}^{\mathrm{T}} \tag{9}$$

Donde $\mathbf{G}$ representa la matriz $\mathbf{G} = \left[\boldsymbol{\xi}_1^0 \; \boldsymbol{\xi}_2^0 \cdots \boldsymbol{\xi}_n^0 \right]$. En la mayoría de casos es muy difícil calcular los vectores propios $\mathbf{v}$ y valores propios λ de una matriz tan grande. Sin embargo, teniendo en cuenta las propiedades de , en especial, aquella que dice que tiene los mismos valores propios no-nulos que y la ventaja de que , como se da en [17]:

$$\mathbf{G}^{\mathrm{T}} \mathbf{G} \hat{\mathbf{v}}_i = \lambda \hat{\mathbf{v}}_i \tag{10}$$

Siendo $\hat{\mathbf{v}}_i$ los valores propios de $\mathbf{G}^T\mathbf{G}$, tal que $\mathbf{v}_i = \mathbf{G}\hat{\mathbf{v}}_i$. Entonces, los vectores propios correspondientes a los valores propios de $\mathbf{S}$ diferentes de cero son $\mathbf{v}_i = \mathbf{G}\hat{\mathbf{v}}_i / \|\mathbf{G}\hat{\mathbf{v}}_i\|$. Los vectores propios asociados con los r valores propios de mas grandes, son seleccionados como las Direcciones Principales [7], los cuales generan una base ortonormal para un subespacio que contiene la mayor parte de la información ofrecida por las observaciones. Por otra parte, debido a que PCA es una transformación lineal de los datos, es posible reconstruir una observación a partir de la suma ponderada de los vectores propios o direcciones principales, por medio de

$$\hat{\boldsymbol{\xi}}_i^0 = \sum_{k=1}^{r} w_k \mathbf{v}_k^T \tag{11}$$

Donde los pesos de reconstrucción están dados por $w_k = \mathbf{v}_k^T \boldsymbol{\xi}_i^0$ y pueden ser vistos como un nuevo conjunto de características, y teniendo en cuenta la propiedad de ortonormalidad de la base, las observaciones pueden ser clasificadas empleando cualquier criterio geométrico sobre el subespacio particionado (KNN).

Por otro lado, el método propuesto permite identificar y seleccionar las características dinámicas que de mayor forma contribuyen al proceso de reconocimiento. Las magnitudes de los valores propios que generan la base de representación nos dicen las variables que se deben tomar. Sea $\boldsymbol{\rho}$ el vector expresado como $\boldsymbol{\rho} = \sum_{k=1}^{r} \lambda_k |\mathbf{v}_k|$, los mayores valores dentro del vector $\boldsymbol{\rho}$ señalan cada uno de los puntos de las variables dinámicas que más influencia tienen en el proceso. Esta suma de valores absolutos es una aproximación debido ala equivalencia de normas en espacios finitos (L^1 y L^2). Reordenando de la siguiente forma:

$$\boldsymbol{\rho} = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1m} & \rho_{21} & \cdots & \rho_{2m} & \cdots & \rho_{p1} & \cdots & \rho_{pm} \end{bmatrix}^{\mathrm{T}}$$

$$\Rightarrow \mathbf{P} = \begin{bmatrix} \rho_{11} & \rho_{21} & \cdots & \rho_{p1} \\ \rho_{12} & \rho_{22} & \cdots & \rho_{p2} \\ \vdots & \vdots & & \vdots \\ \rho_{1m} & \rho_{2m} & \cdots & \rho_{pm} \end{bmatrix} \tag{12}$$

Es posible obtener un escalar $\hat{\rho}_j = \sum_{k=1}^{m} \rho_{jk}$, $j = 1, \ldots, p$ que es la suma de cada uno de los elementos de cada columna j de la matriz $\mathbf{P}$. En consecuencia, la suposición mas importante es que los valores mas grandes de $\hat{\rho}_j$ indican los mejores atributos de entrada puesto estas características son las que tienen más alta correlación con los componentes principales.

VI. MATERIALES Y MÉTODOS

La metodología empleada se puede resumir en el diagrama de bloques mostrado en la Figura 1.

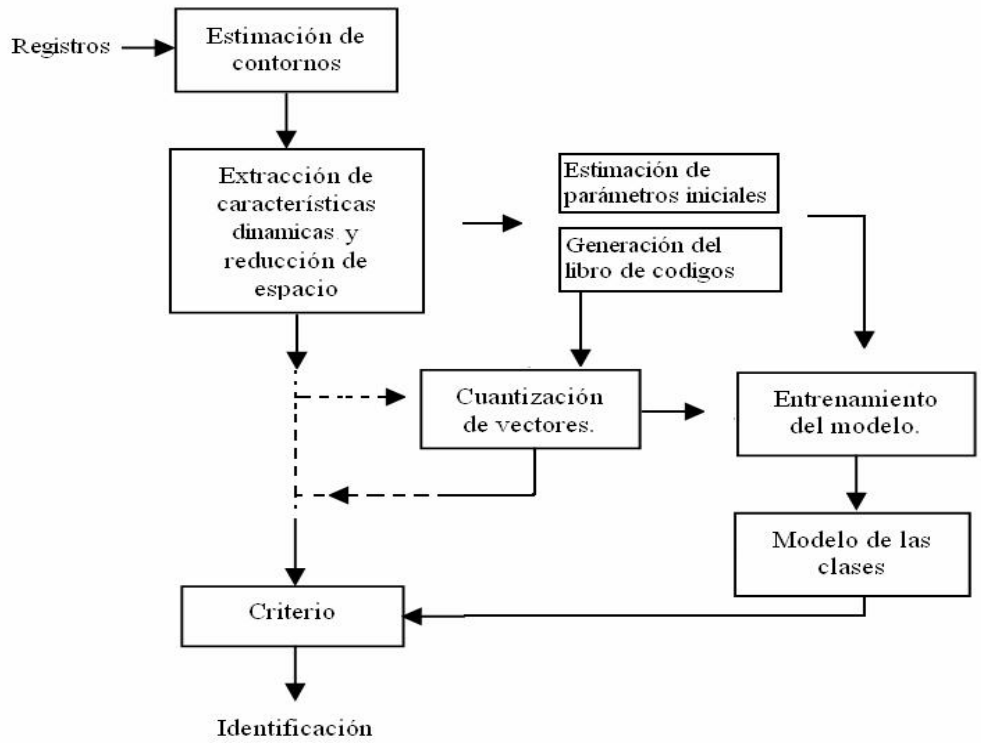


Figura 1. Diagrama de bloques de la metodología propuesta

Teniendo un conjunto de registros de señales o registros, pertenecientes a un conjunto de clases $\{C_1,..,C_m,..,C_M\}$, se debe buscar un espacio de representación adecuado que estará formado por parámetros que permitan diferenciar entre cada una de las clases, reduciendo la cantidad de datos necesarios para representar cada registros sin pérdida de la información relevante necesaria para la clasificación, a esta etapa se le conoce como extracción de parámetros o estimación de contornos.

La siguiente etapa es la extracción de características, esta etapa se emplea para reducir el tamaño del espacio de representación de los datos, proporcionados en la etapa de extracción de parámetros, de tal manera que se utilicen únicamente las variables que mayor información aportan al proceso de clasificación. Aunque teóricamente esta puede no ser necesaria si la etapa anterior proporciona un espacio de representación de baja dimensión y alta discriminancia, en la práctica se emplea a menudo por problemas de alta dimensionalidad.

En esta etapa se usa la técnica descrita en la sección **II-2**. Teniendo en cuenta el peso que se tiene se puede reducir el espacio generando un espacio de representación conformado sólo por las características que tienen más peso o relevancia.

Teniendo el conjunto de características con las cuales se va a realizar el entrenamiento el objetivo es generar un modelo de Markov por cada clase, tal que se describa de forma óptima las secuencias de entrenamiento tal como se describe en la sección **II-1.1**. Los parámetros a estimar son la matriz de probabilidad de transición de estados $\prod$, la distribución inicial de estados $\mathbf{p}_{\theta_1}$ y la distribución de probabilidad de observaciones $\mathbf{B}$ que dependiendo del caso, continua o discreta, se deberán estimar los parámetros necesarios para su correcta descripción, para el caso continuo esta distribución se modela mediante una mezcla de gaussianas estimando los pesos de ponderación, el vector de medias y la matriz de covarianzas por estado, para el caso discreto se estima la probabilidad de emitir el símbolo *k* en el estado *i*. generando un vector de probabilidades por cada estado.

A. Descripción de las bases de datos

A.1. Registros de voz

La base de datos BD1 fue desarrollada por el *Massachusetts Eye and Ear Infirmary* [18]. Debido a la heterogeneidad de la base de datos (diferente frecuencia de muestreo en la adquisición de los registros), los registros utilizados fueron remuestreados a una frecuencia de muestreo de 25 kHz y con una resolución de 16 bits. Corresponden a pronunciaciones de la vocal sostenida /ah/. Se utilizaron 173 registros de pacientes patológicos (con una amplia gama de patologías vocales orgánicas, neurológicas, traumáticas y psíquicas) y 53 registros de pacientes normales, de acuerdo con los registros enumerados en [19]. Los registros de pacientes patológicos tienen una duración aproximada de 1s, mientras que en los registros de pacientes normales la duración es de unos 3s.

La base de datos DB2 pertenece al Grupo de Control y Procesamiento Digital de Señales de la Universidad Nacional de Colombia sede Manizales y contiene 80 registros de la vocal sostenida /a/, pronunciadas por 40 pacientes con voz normal y 40 pacientes con voz disfónica. Los registros fueron adquiridos con frecuencia de muestreo de 22050 Hz.

A.2. Registros MER

La base de datos de la UPV de señales MER provenientes de microelectrodos de registro son grabaciones de intervenciones quirúrgicas sobre cinco pacientes etiquetadas por médicos especialistas en el área quienes identificaron la zona en la cual se encuentra el microelectrodo. Los registros se encuentran a diferentes profundidades de acuerdo al equipo estereotáxico. Estos registros se obtuvieron a partir del proyecto Sistema asistido para la toma de decisiones en la cirugía de la enfermedad de Parkinson, código PI031546, financiado por el Ministerio de Sanidad y Consumo, Instituto de Salud Carlos III Fondo de Investigaciones Sanitarias, España. La frecuencia de muestreo de todas las señales es de 24000Hz. Cada registro tiene una duración de 10s. En total existen 177 registros discriminados así: 66 señales de tálamo, 25 de subtálamo, 38 de sustancia negra y 87 de zona incierta.

B. 1. Caracterización de señales

B.1. Señales MER

La caracterización de señales MER está orientada a la identificación de zonas cerebrales (tálamo y subtálamo). Se utilizaron los esquemas de actualización adaptativos propuestos en [8]. Las características seleccionadas son los máximos coeficientes, y la varianza de los niveles de descomposición (aproximación y detalle) obtenidos al aplicar los esquemas adaptativos.

B.2. Registros de voz

La caracterización de señales de voz se realiza en base a la escala de frecuencias *mel*, que está basada en la representación perceptual de los MFCC [20][21] y se han considerado, además, dentro de los vectores de características, parámetros relacionados con mediciones de ruido, diseñados para medir la componente de ruido relativo en las señales de voz. En particular se utilizó la relación armónico ruido (Harmonic to Noise Ratio - HNR) [22], la energía de ruido normalizada (Normalized Noise Energy - NNE) [23] y la relación excitación glottal ruido (Glottal to Noise Excitation Ratio) [24], debido a que estas medidas dan una idea de la calidad y grado de normalidad de la voz.

$$\underbrace{HNE\,|\,NNE\,|\,GNE\,|\,En\,|\,MFCC}_{s1}\;\;\underbrace{\Delta HNE\,|\,\Delta NNE\,|\,\Delta GNE\,|\,\Delta En\,|\,\Delta MFCC}_{s2}$$

Figura. 2: Parámetros contenidos en el vector de características.

El vector de características se forma concatenando el conjunto de parámetros de ruido mencionados, además de su primera derivada temporal debido, a que la velocidad de los cambios en los coeficientes da información importante de su comportamiento dinámico [25]. La Figura 2 muestra gráficamente la composición del vector de características empleado. En la celda 4, el parámetro *En* es la energía medida por trama de la señal. El número de coeficientes MFCC utilizados en el vector para las pruebas realizadas es igual a 12. $s2$ Es el conjunto de derivadas de cada uno de los parámetros anteriores. El cálculo de Δ fue realizado por medio de un filtro FIR antisimétrico de respuesta al impulso finita y de longitud 9, para evitar la distorsión de fase de la secuencia temporal [26].

C. Extracción dinámica de características

La variabilidad presente en el conjunto de características considerado, puede ser asociada a la cantidad de información que dicho conjunto contiene. Es posible plantear un criterio de selección, que permita la identificación de aquellas variables que más peso o relevancia aportan a la variabilidad total, examinando el nivel de correlación del conjunto de características dinámicas con respecto a las componentes que maximizan la variabilidad [7]. Debido a que la magnitud absoluta de los vectores propios, ponderados por sus respectivos valores propios, determinan el nivel de correlación entre las variables originales y las componentes principales, se pueden identificar como variables relevantes aquellas asociadas a las mayores magnitudes absolutas anteriormente mencionadas [9]. El conjunto de variables dinámicas obtenidas en la etapa de parametrización, fue reducido empleando una metodología de selección que hace uso del criterio antes mencionado.

D. Entrenamiento y clasificación mediante HMMs

Las pruebas se realizan inicialmente con una sola característica evaluando el desempeño de clasificación cuando se tiene una sola coordenada, luego se incrementa la dimensión del espacio añadiendo otra característica y evaluando de nuevo el desempeño de clasificación, este proceso se repite hasta completar el total de características con las que se cuenta. La forma de incrementar la dimensión del espacio se realiza teniendo en cuenta el orden de importancia que tiene cada de las características, inicialmente orden ascendente y luego en orden descendente, para el caso específico de señales acústicas también se realizan las pruebas en orden aleatorio.

Sobre el conjunto total de registros se realiza de forma aleatoria dos particiones una contiene el 70% que se emplea para entrenamiento y el 30% restante para validación, este proceso se realiza 11 veces. Se debe tener en cuenta que tipo de distribución que se empleará para modelar la estadística de observación por estado que puede ser discreta o continua, en el caso discreto se debe realizar la estimación de un libro de códigos, procedimiento realizado mediante el algoritmo de K-medias, que permite estimar K centroides en el espacio de representación, para luego asociar un vector de observación a un centroide determinado(el mas cercano), este proceso se conoce como cuantización vectorial, y se ilustra en la Figura 3.

Para el caso continuo se debe tener en cuenta que la función que modelará las observaciones es una función paramétrica y de tipo gaussiana, en esta etapa se debe estimar las matrices de covarianzas y vectores de medias para cada uno de los estados, procedimiento realizado mediante el algoritmo *K-means clustering*.

Figura 3. Diagrama de bloques de la estructura de entrenamiento y clasificación básica del VQ.

Teniendo las observaciones se procede a estimar el modelo para cada una de las clases a ser reconocidas, generando un conjunto de modelos $\Theta = \{\lambda_1, .., \lambda_m, .., \lambda_M\}$. Para clasificación se emplea el criterio de *máxima probabilidad a posteriori* (MAP) el cual consiste en asociar un registro a la clase cuyo modelo tiene mayor probabilidad de haber generado esa secuencia de observaciones.

V. RESULTADOS

A. Resultados sobre señales acústicas -DB1

En la Figura 4 se muestran los pesos de cada una de las características, por el eje de la abcisa se tiene el cardinal de cada característica, mientras por el eje de la ordenada se muestra el peso que tiene cada una de las características a lo largo de todo el tiempo de observación. Cabe anotar que la importancia de cada característica está dada por su magnitud en el eje de las ordenadas.

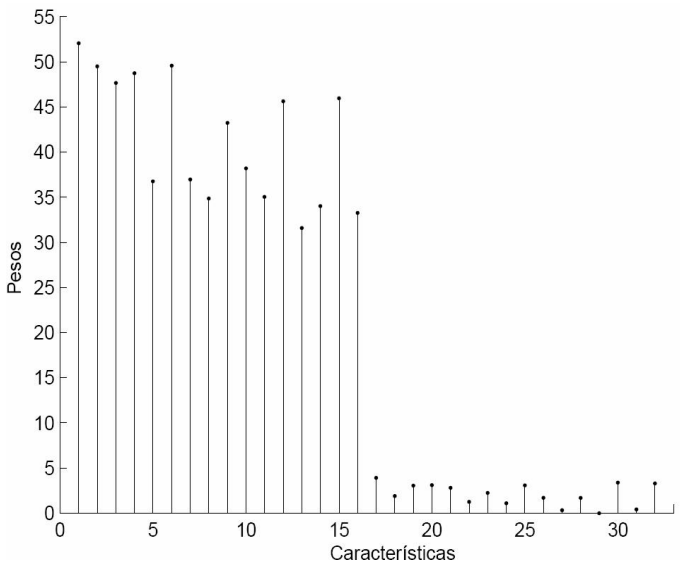


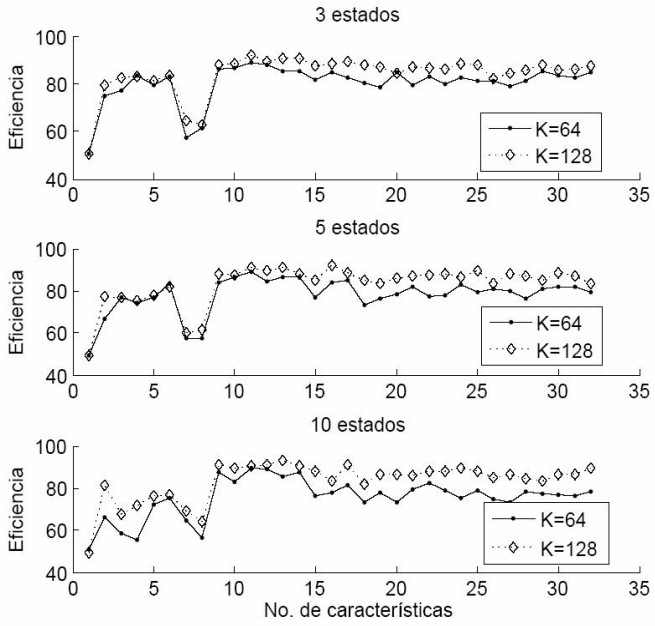
Figura 4. Resultados de la extracción dinámica de características – DB1

A.1. Densidad de probabilidad de observación discreta

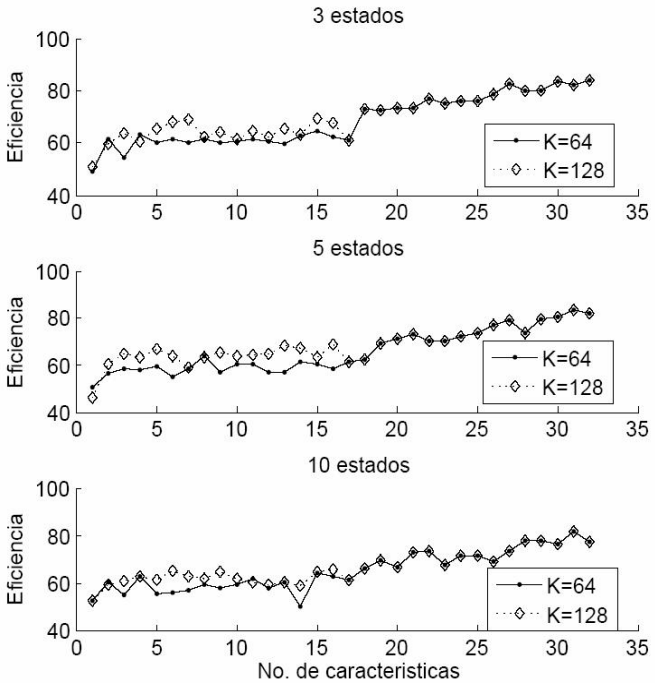
Las pruebas son realizadas inicialmente para la base de datos de voz DB1 empleando HMMs con una distribución de observación discreta, como se mencionó anteriormente en la medida que se incrementa la dimensión de espacio de características se realiza una evaluación del sistema de clasificación obteniéndose el rendimiento para diferentes casos, variando el número de estados y el tamaño del libro de códigos.

Los primeros experimentos se realizan incrementando el espacio sin tener en cuenta el peso de las características, es decir, en orden aleatorio los resultados se pueden ver en la Figura 5(a) donde se muestra que el máximo desempeño se alcanza con pocas características, y se puede concluir que no es necesario tener un espacio de representación demasiado grande, pero este conjunto puede verse afectado por características que desmejoran el desempeño del sistema de clasificación.

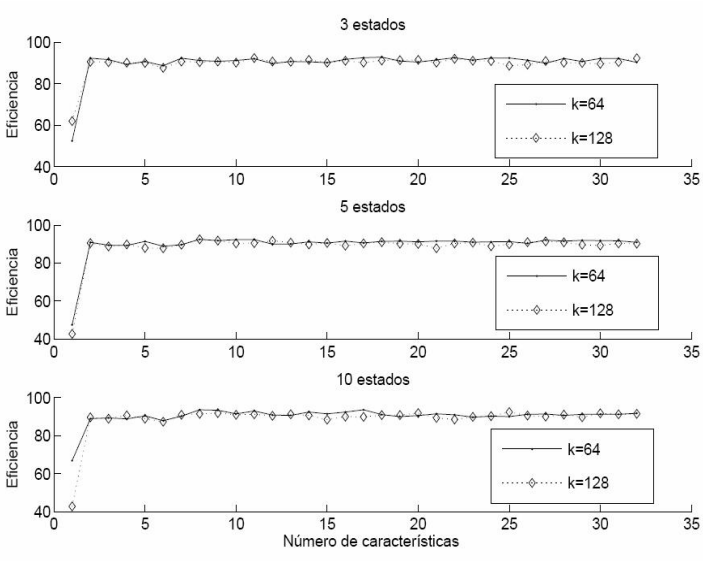
En la medida en que se añaden características se nota que el rendimiento no es uniforme y no tiene una tendencia definida, es decir, en algunos casos tiende a aumentar y si se añade otra característica tiende a disminuir, de esta forma no se puede concluir que parámetros de entrenamiento son los más adecuados aunque se tiene mayor eficiencia cuando se usan 128 clusters para el libro de códigos. La explicación a este fenómeno puede estar en que la adición de nuevas características sin un aporte significativo de relevancia puede empeorar el rendimiento de los algoritmos de estimación de parámetros para los modelos de HMM.



(a)



(b)



(c)

Figura 5. Resultado del entrenamiento con características en orden (a) Aleatorio, (b) Ascendente y (c) Descendente

En la Figura 5(b) se muestran los experimentos que se realizan incrementando el espacio teniendo en cuenta los pesos que se muestran en la Figura 4 de forma ascendente. Se nota que existe una tendencia ascendente en la tasa de rendimiento del sistema de clasificación, pero su valor no supera el 85%, teniendo como máximo un 84,09%, cuando se entrena con 3 estados, independiente del tamaño del libro de códigos y con el espacio completo, es decir, con las 32 características, aun si la tendencia es ascendente el resultado obtenido no es el mas alto posible, lo cual se hace evidente al analizar los resultados en orden aleatorio donde se muestra que el rendimiento mas alto se alcanza con 11 características, 3 estados, un libro de códigos de 128 centroides, este caso es del 92%.

Los resultados mas interesantes se presentan cuando el espacio de características se incrementa teniendo en cuenta los pesos de las características en orden descendente, es decir, se añaden primero las características con una mayor importancia y progresivamente se añaden las que tienen menos peso, los resultados se muestran en la Figura 5(c), como se observa con muy pocas características se alcanza el mejor rendimiento para todos los casos de análisis, alrededor del 95%; aunque al agregar características adicionales el rendimiento tienda a disminuir, esto muestra que es posible alcanzar rendimientos bastante aceptables con un conjunto de características muy reducido y no es necesario emplear el conjunto completo características.

Tabla. 1 Mejores resultados y el número de características empleado.

C. B.	Numero de estados					
	3		5		10	
	Ef.	N. C.	Ef.	N. C.	Ef.	N. C.
64	95.4±2.6	18	95.0±2.8	8	96.0±2.6	17
128	95.0±3.3	11	95.0±1.7	8	94.1±2.6	25

A.2. Densidad de probabilidad de observación continua

Empleando la misma base de datos se realizan pruebas para cuando se usan HMMs con distribución de observación continua, variando el numero de mezclas y el numero de estados, realizando los experimentos de igual forma que en el caso discreto, incrementando la dimensión del espacio de entrenamiento gradualmente pero en este caso solo hasta 16 características, teniendo en cuenta que para el caso discreto se obtienen buenos resultados con un conjunto de 8 y 11 características, los resultados se muestran en la Figura 6 y la Tabla 2.

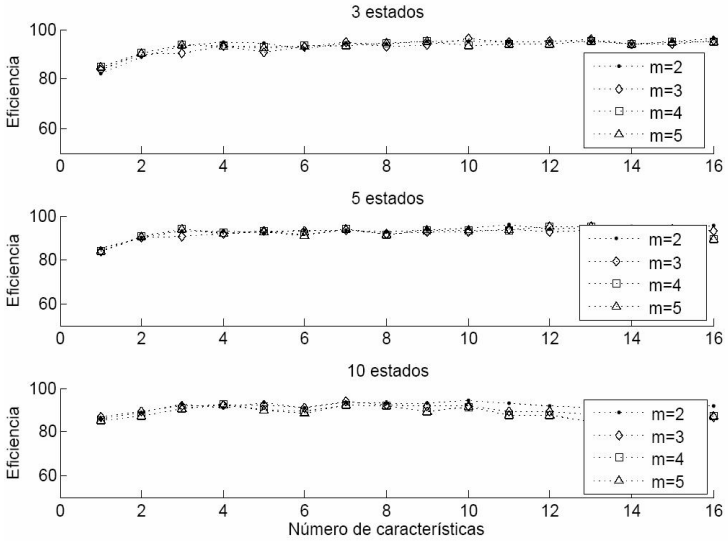


Figura 6. Resultados con distribución de observación continua en orden descendente

Tabla. 2. Mejores resultados empleando HMM continuos

N. de G.	Numero de estados					
	3		5		10	
	Ef.(%)	N. C.	Ef.(%)	N. C.	Ef.(%)	N. C.
2	96.3±2.8	16	95.9±2.8	11	94.1±2.7	10
3	96.3±2.8	10	94.0±2.7	11	93.7±2.7	7
4	95.5±2.7	13	95.7±2.7	12	92.7±2.9	4
5	94.1±2.8	10	94.5±2.6	11	92.9±2.5	9

Se puede notar que cuando se emplean dos gaussianas por mezcla se tienen buenos resultados, aunque en general para los demás casos el rendimiento obtenido no supera el caso discreto, y para el caso particular de esta base de datos se puede decir que la clasificación se realiza de una forma óptima cuando se emplean HMMs discretos.

B. Resultados sobre señales acústicas –DB2

Empleando la DB2, y teniendo en cuenta los resultados obtenidos para la base de datos anterior, se realizan los experimentos añadiendo las características de forma descendente, los pesos de las características empleadas se muestran en la Figura 7.

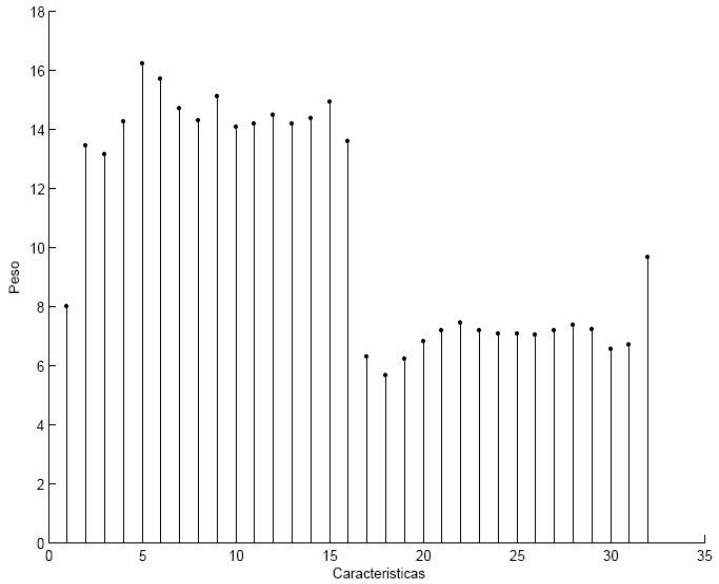


Figura 7. Resultados de la extracción dinámica de características – DB2

B.1. Densidad de probabilidad de observación discreta

Los resultados cuando se emplean HMMs con distribución de observación discreta se muestran en la Figura 8.

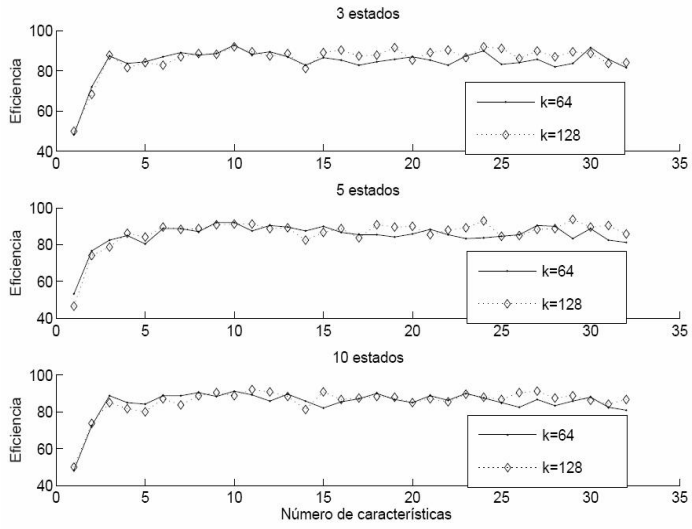


Figura 8. Resultados con distribución de observación discreta en orden descendente

Donde se presentan resultados similares a los obtenidos con la base de datos DB1. En este caso los mejores resultados y el número de características se muestra en la Tabla 3, y se puede observar que el rendimiento más alto se presenta con 29 características, 5 estados y un libro de códigos de 128 centroides pero al comparar este rendimiento con el que se presenta cuando se tienen 10 características, 3 estados y un libro de códigos de 64 centroides se puede notar que la diferencia no es significativa con relación al rendimiento, pero el coste computacional es mucho mas reducido en el último caso.

Tabla 3. Mejores resultados HMM discretos en DB2

C. B.	Numero de estados					
	3		5		10	
	Ef.(%)	N. C.	Ef.(%)	N. C.	Ef.(%)	N. C.
64	92.9±2.2	10	92.0±2.5	10	91.2±2.3	10
128	92.0±2.8	10	93.3±2.8	29	92.0±2.8	11

B.2. Densidad de probabilidad de observación continua

En este caso se puede obtener una mejora en cuanto al desempeño del sistema de clasificación, y los experimentos se realizan solo para 16 características, ya que los resultados significativos se obtienen dentro de este conjunto para el caso discreto. En la Figura 9 se muestran los resultados obtenidos, donde se muestra un comportamiento similar al caso discreto. Se nota que independientemente del número de gaussianas empleadas los resultados son similares, para mayor claridad en la Tabla 4 se presentan los mejores resultados obtenidos, donde se observa que se presenta una mejora aunque no muy significativa con relación al caso discreto, el rendimiento del sistema no presenta tantas fluctuaciones cuando se añaden características al espacio de entrenamiento.

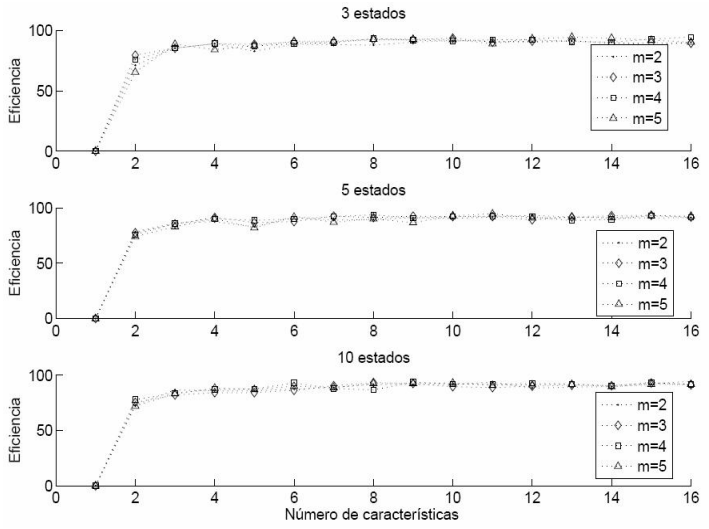


Figura 9. Resultados con distribución de observación continua en orden descendente

En la Tabla 4 se muestra el rendimiento más alto alcanzado para cada uno de los casos, también se muestra el número de características empleadas, cabe anotar que se puede mejorar el resultado obtenido en el caso discreto cuando se tiene 29 características, en este caso con un espacio de dimensión 11, 5 estados y 5 gaussianas.

Tabla 4. Mejores resultados DB2, HMMs con distribución continua, orden descendente

N. de G.	Numero de estados					
	3		5		10	
	Ef.(%)	N. C.	Ef.(%)	N. C.	Ef.(%)	N. C.
2	91.6±2.7	10	93.3±2.8	12	94.5±2.8	16
3	93.7±2.8	10	93.3±2.8	15	93.3±2.8	8
4	94.5±2.9	16	93.7±2.8	8	93.7±2.8	9
5	94.5±2.8	13	94.5±2.8	11	92.9±2.8	10

C. Resultados sobre señales MER

Esta metodología no es posible aplicarla cuando se cuenta con un espacio de características reducido y donde los pesos o ponderaciones para cada una de las características son muy similares o con diferencias poco significativas, este es el caso de señales MER, en la Figura 10 se muestran los pesos obtenidos para el conjunto de características asociado a esta base de datos.

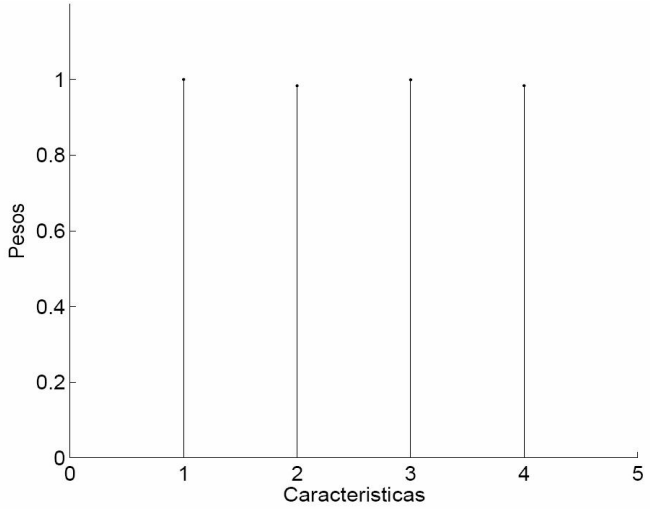


Figura 10. Resultados de la extracción dinámica de características – MER

C.1. Densidad de probabilidad de observación discreta

Para este caso y teniendo en cuenta que las características tienen prácticamente el mismo nivel de importancia no se pueden esperar los mismos resultados obtenidos anteriormente para voz, en este caso los rendimientos mas altos se alcanzan cuando se añaden las características en orden ascendente, si la dimensión del espacio de representación se incrementa orden descendente los resultados mas elevados se obtienen cuando se trabaja con el espacio completo es decir las cuatro características para todos los casos evaluados, en la Tabla 5.

Tabla 5. Mejores resultados con distribución de observación discreta

C. B.	Numero de estados					
	3		5		10	
	Ef.(%)	N. C.	Ef.(%)	N. C.	Ef.(%)	N. C.
64	87.6±4.2	4	87.0±3.4	4	86.8±3.2	4
128	87.9±2.1	4	87.3±3.7	4	88.0±2.6	4

Los resultados que se obtienen cuando se agregan las características en orden ascendente son los mostrados en la Figura 11 y Tabla 6.

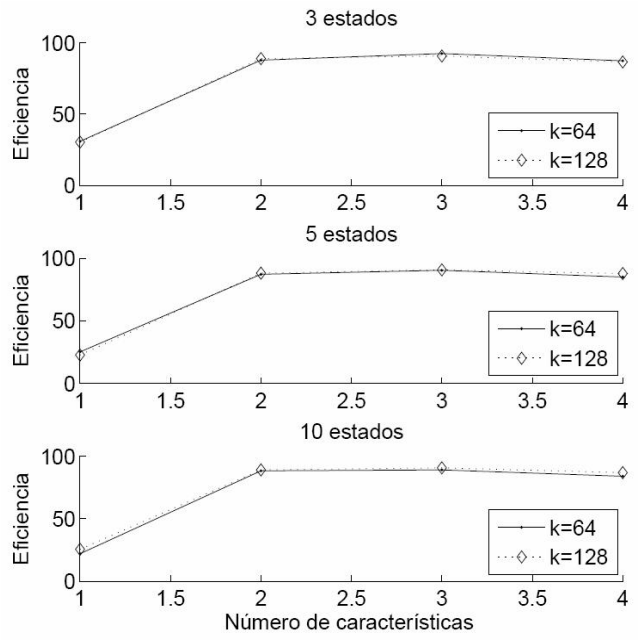


Figura 11. Resultados con distribución de observación discreta en orden ascendente

Tabla 6. Mejores resultados caso discreto orden ascendente

C. B.	Numero de estados					
	3		5		10	
	Ef.(%)	N. C.	Ef.(%)	N. C.	Ef.(%)	N. C.
64	92.2±3.2	3	90.2±2.7	3	89.0±3.4	3
128	90.6±2.7	3	90.6±1.9	3	90.7±2.0	3

C.2. Densidad de probabilidad de observación continua

Empleando la misma base de datos se realizan los experimentos con HMMs continuos, los resultados en orden descendente se muestran en la Figura 12, y en la Tabla 7.

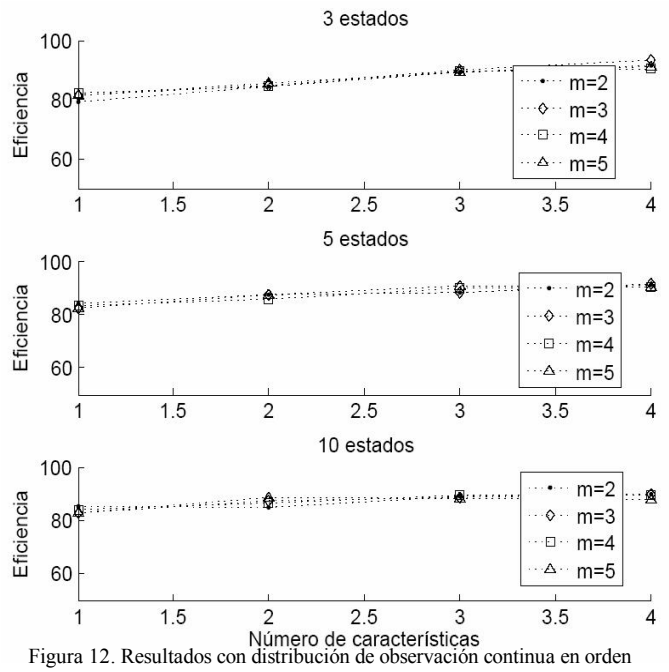


Figura 12. Resultados con distribución de observación continua en orden ascendente

Tabla 7. Mejores resultados, diferentes configuraciones HMM continuos, orden ascendente

N. de G.	Numero de estados					
	3		5		10	
	Ef.(%)	N. C.	Ef.(%)	N. C.	Ef.(%)	N. C.
2	91.6±3.3	4	91.2±3.1	4	89.3±3.2	4
3	93.1±2.4	4	91.5±2.6	4	89.6±2.8	4
4	90.5±3.5	4	90.4±3.2	4	89.5±2.8	3
5	91.0±3.2	4	90.5±4.0	3	88.9±2.8	2

Si el espacio de características se incrementa de forma ascendente se tienen los resultados mostrados en la Figura 13 y en la Tabla 8.

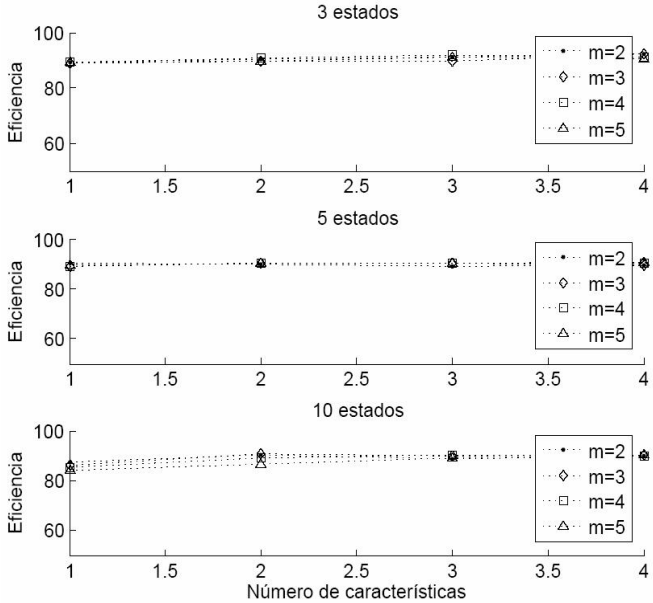


Figura 13. Resultados con distribución de observación continua en orden descendente

Tabla 8. Mejores resultados, diferentes configuraciones HMM continuos, orden descendente

N. de G.	Numero de estados					
	3		5		10	
	Ef.(%)	N. C.	Ef.(%)	N. C.	Ef.(%)	N. C.
2	92.4±2.7	4	90.2±4.3	1	90.1±3.1	2
3	92.4±2.6	4	90.4±3.6	2	90.6±3.1	2
4	92.1±3.1	3	90.2±2.2	2	90.4±2.8	3
5	91.3±3.1	3	90.7±3.3	4	90.1±2.8	4

En las tablas se puede notar que empleando 3 estados y 2 o 3 gaussianas por mezcla se mejora el rendimiento alcanzado cuando se emplea HMMs discretos, pero se debe notar que en este caso se debe tener el espacio de representación completo y la mejora no es significativa.

VI. CONCLUSIONES

La metodología propuesta para reducir el número de características dinámicas en identificación de patologías de voz, demostró que puede ser de gran utilidad teniendo en cuenta los resultados de los experimentos realizados. Como resultado se obtiene un desempeño satisfactorio cuando se emplea un conjunto de características considerablemente reducido y una arquitectura HMM relativamente simple. Esto demuestra que para mejorar el desempeño de un sistema de detección de patologías de voz se debe empezar por encontrar un buen conjunto de características (las de mayor relevancia) en lugar de incrementar la complejidad del modelo que se emplea, lo que permite que la etapa de entrenamiento sea más eficiente.

Del entrenamiento incremental es posible notar que si el conjunto de características inicial no es adecuado, el modelo no puede discriminar correctamente y por lo tanto el desempeño no es bueno, incluso si después se agregan características que aportan significativamente al sistema, estas pueden tener una tendencia a mejorar pero el desempeño alcanzado cuando se emplea el total de características no es el máximo posible. Caso contrario, si se emplea un conjunto inicial apropiado, se observa un rápido aumento en el desempeño del sistema de clasificación en la medida que se agregan las características.

En el último caso la reducción del costo computacional es evidente. Los resultados mostraron que para el caso discreto el tiempo necesario para generar el libro de códigos esta altamente relacionado con el número de características empleado, además el coste computacional se ve ampliamente reducido cuando se emplea un modelo de Markov discreto, que es una arquitectura más simple pero muestra un desempeño igual o superior a los HMM con distribución continua.

Por otro lado, en señales MER, se puede notar que no hay gran diferencia significativa entre los pesos calculados, es decir las características tienen una relevancia similar y por lo tanto no se puede tomar una decisión sobre que características

emplear para generar el modelo, esto indica que no se debe reducir el espacio de entrenamiento, puesto que en cualquier caso el desempeño máximo se alcanza cuando se tiene el espacio de características completo, sin embargo se puede notar que para el caso discreto se tiene un tendencia similar a la observada en la detección de patologías de voz cuando se hace el entrenamiento de forma incremental tomando los pesos en orden ascendente. Otro aspecto a tener en cuenta que es el número de características empleado en esta tarea es pequeño (4 características) y por los resultados se puede concluir que el conjunto de características es óptimo y representa de una forma muy eficiente la dinámica presente en las señales.

VI. TRABAJO FUTURO

Como trabajo futuro se propone el uso de modelos de representación que incluyan la optimización de una medida de separabilidad, también probar otros modelos de variables latentes tales como PCA probabilistico, ICA etc; De igual forma se propone emplear una medida de separabilidad combinada con el análisis de relevancia presentado de tal forma que la reducción pueda hacerse de forma automática, y poder comparar con otros análisis de reducción de espacios dinámicos.

RECONOCIMIENTOS

Este trabajo se enmarca dentro del proyecto "*Clasificación de bioseñales utilizando modelos ocultos de Markov entrenados con métodos discriminativos*" de la la Universidad Tecnológica de Pereira UTP, y "*Detección de los niveles de compromiso de resonancia en niños con labio y/o paladar hendido*" de la Universidad Nacional de Colombia – sede Manizales, financiados por Colciencias dentro del programa *Jóvenes Investigadores*.

Los autores agradecen a los ingenieros Genaro Daza, Franklin Sepúlveda y Eduardo Giraldo por la fundamentación teórica y algunos cálculos realizados en el desarrollo de este trabajo.

BIBLIOGRAFÍA

[1]. Wester M., 1998. Automatic classification of voice quality: Comparing regression models and hidden markov models. En: Symposium on Databases in Voice Quality Research and Education.

[2]. Dibazar A. y Narayanan S., 1998. A system for automatic detection of pathological speech. En: Proceedings of the 36th Asilomar Conf. Signals, Systems and Computers.

[3]. Kadous M. W., 2002. Temporal classification: Extending the classification paradigm to multivariate time series. PhD thesis, School of Computer Science and Engineering.

[4]. Andreão R., Dorizzi B. y Boudy J., 2006. ECG Signal Analysis Through Hidden Markov Models. En: IEEE Transactions on Biomedical Engineering, Vol. 53, No. 8, pp. 1541-1550.

[5]. Lotte F., Congedo M., Lécuyer A., Lamarche F. y Arnaldo B., 2007. A review of classification algorithms for EEG-based brain-computer interfaces. En: Journal of Neural Engineering.

[6]. El-Hanjouri M., Alkhaldi W., Hamdy N. y Abdel Alim O., 2002. Heart diseases diagnosis using HMM. En: IEEE MELECON, Cairo, Egypt, pp. 489-493.

[7]. Jolliffe I.T., 2002. Principal component analysis, 2nd ed. New York: Springer series in statistics.

[8]. Giraldo E., Orozco A. y Castellanos G., 2007. Extracción de características en señales MER para el reconocimiento de zonas cerebrales. En: IV Congreso Latinoamericano de Medicina Biomédica CLAIB, Venezuela.

[9]. Daza G, Arias J.D., Godino J.I., Sáenz N., Osma V. y Castellanos G., 2009. Dynamic feature extraction: an application to voice pathology detection. En: Intelligent automation and soft computing (in press).

[10]. Rabiner L., 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. En: IEEE, Vol. 77.

[11]. Collins M., 1997. The EM algorithm. University of Pennsylvania, Tech.

[12]. Moon T.K., 1966. The expectation-maximization algorithm. En: IEEE Signal Processing Magazine, pp. 47–60.

[13]. Bilmes J., 1998. A gentle tutorial of the EM algorithm and its application to parameter stimation for Gaussian mixture and Hidden Markov Models. Berkeley CA: International Computer Science Institute.

[14]. Rankinea L., Mesbaha M. y Boashash B., 2007. IF estimation for multicomponent signals using image processing techniques in the time–frequency domain, Signal Processing, Vol. 87, pp. 1234-1250.

[15]. Stemmer G., Hacker C., Noth E. y Niemann H., 2002. Multiple Time Resolutions for Derivative s of Mel-Frequency Cepstral Coefficients, Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on, pp. 37-40.

[16]. Navin-Gupta C., Palaniappan R., Rajan S., Swaminathan S. y Krishnan S.M., 2005. Segmentation and Classification of Heart Sounds, CCECE/ CCGEI, IEEE, pp. 1674-1677.

[17]. Turk M. y Pentland A., 1991. Eigenfaces for recognition. En: Cognitive Neuroscience, Vol. 3, No. 1, pp. 71–86.

[18]. 1994. Voice disorders database. Versión 1.03 [CDROM]. Massachusetts Eye and Ear Infirmary. Lincoln Park, NJ: Kay Elemetrics Corp.

[19]. Parsa V. y Jamieson D., 2000. Identification of pathological voices using glottal noise measures. En: Journal of Speech Language and Hearing Research, Vol. 43, No. 2, pp. 469–485.

[20]. Gold B. y Morgan N., 2000. Speech and Audio Signal Processing. John Wiley & Sons, INC.

[21]. Deller J.R., Proakis J.G. y Hansen J.H., 1993. Discrete-Time Processing of Speech Signals. J. Griffin, Ed. Macmillan Publishing Company.

[22]. Deliyski D., 1993. Acoustic model and evaluation of pathological voice production, En: Proceedings of Eurospeech , Vol. 3, Berlin, Germany, pp. 1969–1972.

[23]. Kasuya H., Ogawa S., Mashima K. y Ebihara S., 1986. Normalized noise energy as an acoustic measure to evaluate pathologic voice. En: Journal of the Acoustical Society of America, Vol. 80, No. 5, pp. 1329–1334.

[24].Michaelis D., Gramms T. y Strube H.W., 1997. Glottal-to-noise excitation ratio – A new measure for describing pathological voices. En: Acustica/ Acta acustica, Vol. 83, pp. 700–706.

[25]. Godino J.I., Gómez P., Sáenz N., Blanco M., Cruz F. y Ferrer M.A., 2005. Discriminative methods for the detection of voice disorders. En: Proceedings of the 3th International Conference on Non-Linear speech processing.

[26].Godino J., Gómez P., M. y Blanco M., 2006. Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters. En: IEEE Transactions on Biomedical Engineering, Vol. 53, No. 10, pp. 1943–1953.