Entrenamiento discriminativo maximizando una distancia entre modelos de clases

Milton O. Sarria-Paja[1]
Cesar G. Castellanos-Domínguez[2]

Resumen:

En este trabajo se presenta una técnica de entrenamiento discriminativo para modelos ocultos de Markov, orientado a la detección de patologías en señales de voz. La técnica busca maximizar el área que encierra la curva ROC (Receiver Operating Characteristic curve) ajustando los parámetros de modelo, empleando como función objetivo la distancia entre las medias de las funciones de densidad de probabilidad subyacentes asociadas a cada clase. Como resultado se obtiene una mejora en el desempeño del sistema de clasificación comparada con diferentes criterios de entrenamiento.

Palabras clave:

Modelos Ocultos de Markov, Detección de patologías, Entrenamiento discriminativo, Curvas de desempeño.

Abstract:

This paper presents an approach that improves discriminative training criterion for Hidden Markov Models, and it is oriented to voice

1 M.Sc. Ingeniero Electrónico, Docente Ocasional **Instituto Tecnológico Metropolitano**, Grupo MIRP, miltonsarria@itm.edu.co.

2 Ph. D. Ingeniero en Telecomunicaciones, Docente asociado al departamento de Ingeniería Eléctrica, Electrónica y Computación de la Universidad Nacional de Colombia – Manizales, Grupo de control y procesamiento digital de señales, cgcastellanosd@unal.edu.co.

Fecha de recepción: 11 de marzo de 2010
Fecha de aceptación: 07 de junio de 2010

pathological identification. This technique aims at maximizing the Area under the Receiver Operating Characteristic curve by adjusting the model parameters using as objective function the distance between the means of the underlying probability densities functions associated with each class. As result we obtain an improvement in the performance of the classification system compared with different training criteria.

Keywords:

Hidden Markov Models, Detection of pathology, Discriminative training, Performance curves.

1. Introducción

En este trabajo se aborda el problema de entrenamiento en Modelos Ocultos de Markov (*Hidden Markov Models – HMM*). Modelos ampliamente utilizados en sistemas de reconocimiento de voz, ubicándose como una herramienta estándar para modelar las variaciones estocásticas presentes en este tipo de bioseñales (Rabiner, 1989). En este caso específico se estudia el problema de entrenamiento en una aplicación de detección de patologías en señales de voz, que comprende uno de los problemas mas importantes en aplicaciones biomédicas de las tecnologías de reconocimiento de voz (Wang & Jo, 2007).

El entrenamiento de los HMM implica el ajuste de los parámetros de un modelo, tal que se extraiga la máxima información de las secuencias de observación. Entre los métodos más conocidos están el criterio basado en la estimación de máxima verosimilitud (*Maximum Likelihood Estimation* - MLE) (Bilmes, 1998), la técnica de Máxima Información Mutua (*Maximum Mutual Information* – MMI) (Bahl et al., 1986), o el criterio de Mínimo Error de Clasificación (*Minimum Classification Error* - MCE) (Juang et al., 1997). El primer caso se conoce como un criterio de entrenamiento generativo y los dos últimos como criterios discriminativos.

En el desarrollo de herramientas de diagnostico asistido, otro aspecto importante es el uso de métricas de desempeño robustas, por ejemplo, sensitividad, especificidad o el uso de curvas de rendimiento como la curva ROC (*Receiver Operating Characteristic*), más específicamente, se ha empleado el área que encierra (***área bajo la curva -*** *ABC*) como un buen indicador del desempeño de un sistema de clasificación; y por lo tanto se puede plantear un criterio de entrenamiento que busque maximizar el ABC.

Así por ejemplo en (Li et al., 2002) se presenta una primera aproximación de entrenamiento discriminativo conocido como *FOM-training*, la técnica consiste en es ajustar los parámetros de un modelo de mezclas de Gaussianas (GMM) en el cual se

busca maximizar el área que encierra la curva ROC, obteniendo los gradientes de forma empírica. Como resultado, la capacidad discriminante del clasificador mejora en comparación con el criterio MLE. Siguiendo esta idea se presenta otra aproximación en (Gao et al., 2003) donde se incorpora algunas medidas de interés en el diseño de un clasificador, este criterio se conoce como MFOM. Sin embargo el criterio de entrenamiento no trabaja directamente sobre la curva como se propuso originalmente.

El principal inconveniente en la implementación del criterio *FOM-training* es la ausencia de una expresión cerrada que se pueda relacionar bien a la curva ROC o al ***área*** que encierra. Para superar el anterior problema, este trabajo propone optimizar una función analítica que se pueda relacionar con una curva de desempeño. Dado que el área que encierra la curva ROC es directamente proporcional a la separación entre las funciones densidad de probabilidad de cada una de las clases, entonces se puede emplear una medida de separación como función a optimizar. De esta forma, maximizar la separación entre las densidades de probabilidad debería llevar a que el área que encierra la curva ROC se maximice, y por lo tanto el rendimiento del sistema de clasificación sea mejorado.

El experimento que se detalla en este trabajo consiste en realizar una comparación entre los métodos de entrenamiento clásicos (MLE, MMI y MCE) y el criterio de entrenamiento propuesto, empleando la base de datos de patologías de voz desarrollada por *The Massachusetts Eye and Ear Infirmary Voice Laboratory (MEEIVL)* (MEEIVL, 1994) y empleando diferentes métricas de desempeño para realizar la comparación en un marco suficientemente amplio.

Este manuscrito está estructurado de la siguiente forma: En la sección 2 se hace una descripción detallada de las técnicas de entrenamiento generativo y discriminativo aplicables a HMM, a continuación se explica la construcción de la curva ROC a modo de introducción para describir el método de entrenamiento propuesto. En la sección 3 se describe el ajuste experimental, como las bases

de datos, la parametrización, la metodología de validación y la arquitectura del modelo. Las dos últimas secciones presentan los resultados y conclusiones del trabajo.

2. Materiales y Métodos

Asuma la siguiente descripción para el conjunto de entrenamiento $\mathbf{Y} = \{\boldsymbol{\varphi}_r^{n\varphi_r}: r=1,..., R\}$, compuesto por R observaciones con sus respectivas etiquetas, $\mathbf{C} = \{c^r: r=1,..., R\}$, donde $\boldsymbol{c}^r \in \{c_m : m=1,...,M\}$, siendo M el número total de clases. Cada registro $\boldsymbol{\varphi}_r^{n\varphi_r}$ se representa por una secuencia de longitud $n\varphi$ de vectores de características $\boldsymbol{\varphi}_r^{n\varphi_r} = \{\boldsymbol{\varphi}_{r,t} \in^{\rho} : t=1,..., n\varphi_r\}$ Siendo ρ la dimensión del espacio de características.

El conjunto total de parámetros de los HMM se denota por Θ y se compone por M modelos, es decir, $\Theta=\{\lambda_m\}$, donde λ_m denota los parámetros del HMM que representa la categoría o clase c_m y esta definido por el conjunto de parámetros $\lambda_m = \{\boldsymbol{A}^{(m)}, \boldsymbol{B}^{(m)}, \boldsymbol{\pi}^{(m)}\}$, donde $\boldsymbol{A}^{(m)}$ es la matriz de transición de estados, y está compuesta por las probabilidades discretas $\boldsymbol{a}_{ij}^{(m)}$ que representa la probabilidad de pasar del estado $\boldsymbol{s}_i$ al estado $\boldsymbol{s}_j$, $\boldsymbol{B}^{(m)}$ corresponde a la función densidad de probabilidad de observación que en este caso particular se describe mediante un modelo de mezclas de Gaussianas por estado, definido en (1):

$$b_j^{(m)}(\varphi_{r,t}) = \sum_{k=1}^{K} c^{(m)}{}_{jk} \, \mathcal{N}\left[\varphi_{r,t}, \mu^{(m)}{}_{jk}, \Sigma^{(m)}{}_{jk}\right] \qquad (1)$$

Donde $\boldsymbol{\mu}^{(m)}{}_{jk}$ es el vector de medias y $\boldsymbol{\Sigma}^{(m)}{}_{jk}$ la matriz de covarianzas de la k-ésima mezcla del estado s_j, que por simplicidad se asume diagonal, es decir, $\boldsymbol{\Sigma}^{(m)}{}_{jk}$, además, $\boldsymbol{\pi}^{(m)}$ corresponde al vector de probabilidad de estado inicial (Bilmes, 1998).

Los modelos ocultos de Markov describen procesos estocásticos doblemente anidados, compuestos de una capa oculta que controla la evolución temporal de las características espectrales de una capa observable (Rabiner, 1989).

2.1 Criterios de entrenamiento

2.1.1 Criterio MLE

Se asume que la forma funcional de $P\left(\boldsymbol{\varphi}_r^{n\varphi_r} \mid \boldsymbol{c}^r\right)$ es conocida, y puede estimarse al ajustar el conjunto de parámetros del modelo para optimizar la descripción de un conjunto de observaciones dado. La función objetivo ML se define en (2):

$$f_{ML}(\Theta) = \sum_{r=1}^{R} \log\left(P\left(\varphi_r^{n\varphi_r} \left| \mathbf{c}^r \right.\right)\right) \qquad (2)$$

Cuya optimización se alcanza ajustando los parámetros de cada modelo por separado, con los datos de entrenamiento de la clase correspondiente (Bilmes, 1998; Rabiner, 1989).

2.1.2 Criterio MMI

Dada una secuencia de observación, se debe escoger la clase c_m que garantice un mínimo de incertidumbre. Ésta condición puede alcanzarse minimizando la entropía condicional, $H(\mathbf{C} \mid \mathbf{Y})=H(\mathbf{C})-I(\mathbf{C};\mathbf{Y})$, cuya optimización implica minimizar la entropía $H(\mathbf{C})$, o bien maximizar la información mutua $I(\mathbf{C};\mathbf{Y})$. La primera tarea corresponde a hallar el modelo con el mínimo de entropía, que analíticamente es complejo e intratable. En la segunda aproximación, se maximiza la información mutua (Bahl et al., 1986), dada en (3):

$$f_{MMI}(\Theta) = \frac{1}{R}\sum_{r=1}^{R}\left(\log P\left(\varphi_r^{n\varphi_r} \left| \mathbf{c}^r \right.\right) - \log \sum_{i=1}^{M} P\left(\varphi_r^{n\varphi_r} \left| c_i \right.\right) P(c_i)\right) \qquad (3)$$

Criterio MCE. Incluye una función de pérdida, proporcional al error de clasificación, $f_{MCE}(\Theta) = l_i(\boldsymbol{\varphi}_r^{n\varphi_r};\Theta)$, y que se asocia al costo de asignar la secuencia $\boldsymbol{\varphi}_r^{n\varphi_r}$ a la clase c_i, se define como:

$$l_i\left(\varphi_r^{n\varphi_r};\Theta\right) = \begin{cases} 0, & \varphi_r^{n\varphi_r} \text{ asignado correctamente a } c_i \\ 1, & \varphi_r^{n\varphi_r} \text{ asignado incorrectamente a } c_i \end{cases} \qquad (3)$$

Debido a que ésta no es una función derivable, se ha propuesto en cambio emplear una función sigmoidal, (4):

$$l_i\left(d_i\left(\varphi\right)\right)=\frac{1}{1+\exp\left(-\gamma d\left(\varphi\right)+\alpha\right)} \qquad (4)$$

Donde $d_i(\boldsymbol{\varphi})$ se da en (5) y corresponde a una medida de mala clasificación:

$$d_i\left(\varphi\right)=-\mathrm{g}_i\left(\varphi;\lambda_i\right)+\log\left[\frac{1}{M-1}\sum_{j,j\neq i}\exp\left[\mathrm{g}_j\left(\varphi;\lambda_j\right)\eta\right]\right]^{\frac{1}{\eta}} \qquad (5)$$

con $\mathrm{g}_i(\boldsymbol{\varphi};\lambda_i)$ definido como la función de verosimilitud condicional para la clase c_i y η es una constante positiva (Juang et al.,1997).

2.2 Curva ROC - Receiver Operating Characteristic

La toma de decisiones clínicas exige la valoración de la utilidad de cualquier prueba diagnóstica, es decir, su capacidad para clasificar correctamente a los pacientes en categorías o estados en relación con la enfermedad (típicamente dos: estar o no estar enfermo, respuesta positiva o negativa). El análisis de su validez puede obtenerse calculando valores como error de clasificación, sensibilidad y especificidad. Sin embargo La curva más utilizada en la literatura médica para la toma de decisiones es la curva ROC (Receiver Operating Characteristic), que representa la tasa de falso acierto o falsa aceptación (FP) en función de la tasa de acierto o aceptación verdadera (VP), para diferentes valores del umbral de decisión (γ). La disposición de la ROC (Fig. 1) depende de la forma y del solapamiento de las distribuciones subyacentes de las clases (patológica, normal – positiva, negativa) (Hanley & McNeil, 1982).

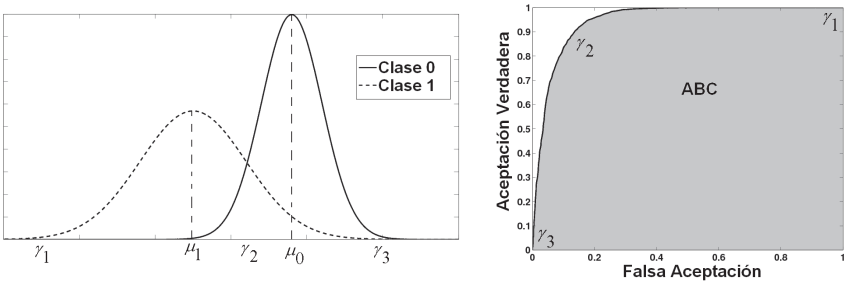


FIG. 1. CONSTRUCCIÓN DE LA CURVA ROC

En el caso de HMM, el cálculo de la curva ROC se hace mediante el uso de los cocientes (puntajes) de verosimilitud estimados de cada registro con los modelos asociados a cada clase, y construir así las densidades de probabilidad subyacentes generadas por cada uno de los modelos. Con los puntajes obtenidos se crea un histograma, que para los registros pertenecientes a la clase positiva (clase 0) debería estar situado en su mayor parte a la derecha y para los que pertenecen a la clase negativa (clase 1) en su mayor parte a la izquierda (Fawcett, 2006). Así, la puntuación para la secuencia $\boldsymbol{\varphi}_r^{n\varphi_r}$ se calcula mediante (6):

$$s_r = \log\left(P\left(\varphi_r^{n\varphi_r} \middle| \lambda_0\right)\right) - \log\left(P\left(\varphi_r^{n\varphi_r} \middle| \lambda_1\right)\right) \qquad (6)$$

Donde λ_i representa el modelo asociado a la clase c_i,i=0,1. El histograma normalizado se puede interpretar como una versión discreta de las funciones densidad de probabilidad de las clases.

Una mayor precisión diagnóstica de la prueba se traduce en el desplazamiento hacia arriba y a la izquierda de la curva ROC, lo que sugiere que el ***área bajo la curva (ABC)*** se puede emplear como un índice conveniente de la exactitud global de la prueba; el mejor indicador correspondería a un valor de 1 y el mínimo a uno de 0,5 (si fuera menor de 0,5 debería invertirse el criterio de decisión de la prueba) (Hanley & McNeil, 1982; Fawcett, 2006). En este sentido, se ha propuesto emplear como criterio de entrenamiento la optimización del área que encierra la ROC, teniendo como

restricción la no existencia de una función analítica que represente el ABC (Li et al., 2002).

2.3 Criterio de entrenamiento propuesto

Para resolver el problema planteado es necesario hacer un análisis sobre la construcción de la curva ROC, se puede notar que cuando las distribuciones de probabilidad están separadas, tanto como es posible (Fig. 2) se puede asumir que el ABC alcanzara un valor máximo, por lo tanto, se propone utilizar una medida de distancia entre las distribuciones, cuya optimización indirectamente debe mejorar el ABC.

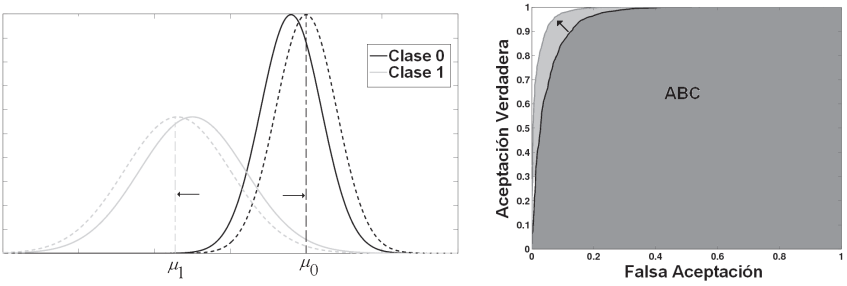


FIG. 2. MAYOR SEPARACIÓN ENTRE DENSIDADES DE PROBABILIDAD, MEJOR RENDIMIENTO

La distancia de Mahalanobis es la opción más clara, y la distancia entre distribuciones que mejor se ajusta a los requerimientos planteados:

$$D^2 = (\mu_0 - \mu_1)^2 S^{-1} \tag{7}$$

S corresponde a la varianza promedio de las distribuciones y se calcula de la siguiente forma:

$$S = ((n_0 - 1)\ S_0 + (n_1 - 1)S_1)/N \tag{8}$$

Así mismo μ_1 y S_i son las medias y varianzas respectivamente, de las distribuciones de cada clase (i = 0,1), n_0 es el número de registros de la clase 0, n_1 los de la clase 1, y $N = n_0 + n_1 - 2$.

Analizando (7) es claro que existen por lo menos tres formas de hacer que la distancia D^2 sea máxima, bien maximizando la distancia entre medias (μ_i) de cada distribución, minimizando S definida en (8); y por último, maximizando directamente la distancia D^2, tal como está definida en (7). Teniendo en cuenta estos aspectos y con el fin de simplificar el problema desde el punto de vista matemático, se asume que la dispersión para cada clase permanece constante y por lo tanto se debe maximizar la distancia euclídea entre medias, así como se ilustra en la Fig. 2. La media de muestra para la densidad de probabilidad asociada a la clase i se puede estimar de la siguiente forma:

$$\mu_i = \frac{1}{n_i}\sum_{r=1}^{n_i} s_r 1\left(\varphi_r^{\,n\varphi_r} \in c_i\right) \tag{9}$$

En (9) el operador 1(.) es una función que toma el valor de 1 si (.) es verdadero y 0 de otra forma. Para realizar el cálculo de $\log\left(P\left(\boldsymbol{\varphi}_r^{\,n\varphi_r} \mid \lambda_l\right)\right)$ se emplea el algoritmo de Viterbi (Rabiner, 1989).

2.3.1 Proceso de optimización

Para actualizar los parámetros de cada uno de los modelos se emplea el algoritmo GPD (Generalized Probabilistic Descend) (Juang & Katagiri, 1992), es una técnica de optimización basada en el cálculo de gradientes, donde se definen las siguientes transformaciones sobre los parámetros a actualizar, que permiten mantener las restricciones probabilísticas de los HMM durante la adaptación, dichas restricciones se expresan en (10):

$$\begin{aligned} &\pi_j \to \hat{\pi}_j \text{ donde } \pi_j = \frac{e^{\hat{\pi}_j}}{\sum_k e^{\hat{\pi}_k}} \quad a) \\ &a_{ij} \to \hat{a}_{ij} \text{ donde } a_{ij} = \frac{e^{\hat{a}_{ij}}}{\sum_k e^{\hat{a}_{kj}}} \quad b) \end{aligned} \tag{10}$$

Las transformaciones que se hacen sobre las componentes Gaussianas del modelo, se definen en (11):

$$
\begin{aligned}
c_{jk} \rightarrow \hat{c}_{jk} \text{ donde } c_{jk} &= \frac{e^{\hat{c}_{ij}}}{\sum_{k} e^{\hat{c}_{kj}}} \quad a) \\
\mu_{jkl} \rightarrow \hat{\mu}_{jkl} &= \frac{\mu_{jkl}}{\sigma_{jkl}} \quad b) \\
\sigma_{jkl} \rightarrow \hat{\sigma}_{jkl} &= \log \sigma_{jkl} \quad c)
\end{aligned}
\tag{11}
$$

La actualización de un parámetro θ en particular, se realiza como está indicado en (12):

$$
\hat{\theta}(n+1) = \hat{\theta}(n) + \varepsilon \frac{\partial f(\Theta)}{\partial \hat{\theta}} \tag{12}
$$

Donde ε es la tasa de aprendizaje, n indica la iteración actual y $\partial f(\Theta)/\partial \hat{\theta}$ es la derivada parcial de la función objetivo con respecto al parámetro $\hat{\theta}$. Finalmente para calcular el parámetro θ se emplea el conjunto de ecuaciones (10) y (11).

3. Marco Experimental

3.1 Base de datos

La base de datos sobre la cual se realizan las pruebas fue desarrollada por *The Massachusetts Eye and Ear Infirmary Voice Laboratory* (MEEIVL, 1994). 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 normales o sanos (Saénz-Lechon et al., 2006).

Cada registro fue ventaneado uniformemente con una ventana tipo Hanning de 40 ms, y con traslape del 50%. A cada ventana se le extrae un vector de 48 características, compuesto por: la relación armónico ruido *(Harmonic-to-Noise Ratio - HNR)* (De Krom, 1993), la energía de ruido normalizada (*Normalized Noise Energy* - NNE) (Kasuya et al.,1986) la relación excitación glottal a ruido (*Glottal to Noise Excitation Ratio* - GNE) (Michaelis et al., 1997), la energía de la ventana (*En*) y 12 MFCC (*Mel-Frequency Cepstrum Coefficients*) (Rabiner & Juang, 1993).

Correspondientes a 16 características que conforman el conjunto *s1*, dicho conjunto se complementa al concatenar su primera (Δ s1) y segunda (Δ(Δ s1)) derivada temporal, debido a que la velocidad de los cambios en los coeficientes dan información importante de su comportamiento dinámico (Godino-Llorente et al., 2005), la primera y segunda derivada de cada uno de los parámetros medidos conforman el conjunto *s2*. Finalmente el vector por cada ventana de análisis esta conformado por los conjuntos *s1* y *s2* con la estructura que se muestra en la Fig. 3.

$$\underbrace{\boxed{HNR \mid NNE \mid GNE \mid En \mid 12 \;\; MFCC}}_{s1} \; \underbrace{\boxed{\Delta s1 \mid \Delta(\Delta s1)}}_{s2}$$

Fig. 3. Estructura del vector de parámetros

Los MFCC son derivados del cálculo de la FFT (*Fast Fourier Transform*) (Rabiner & Juang, 1993). Esta aproximación no paramétrica permite modelar los efectos de las patologías en la excitación (pliegues vocales) y en el sistema (tracto vocal), mientras que los parámetros relacionados con mediciones de ruido (HNR, NNE, GNE) están diseñados para medir la componente de ruido relativo en las señales de voz, y estas medidas dan una idea de la calidad y grado de normalidad de la voz (Godino-Llorente et al., 2005).

3.2 Entrenamiento, arquitectura del modelo y evaluación de desempeño

Como el caso considerado corresponde a observaciones de tipo continuo, estas se modelan mediante un modelo de mezclas de Gaussianas donde es necesario estimar los pesos de ponderación, el vector de medias y la matriz de covarianzas por estado. Es posible por lo tanto variar el número de estados y el número de Gaussianas que conforman la mezcla de cada estado, y de esta forma determinar la arquitectura más adecuada del modelo de Markov en el sistema de clasificación. En la Fig. 4 se muestra la estructura general del sistema implementado.

En la primera etapa se encuentra la estimación de características dinámicas, como se menciono anteriormente para registros de voz se estiman 48 características por ventana. Posterior a esto, sigue la etapa de entrenamiento, que es básicamente el ajuste de los parámetros de los HMM mediante un criterio de entrenamiento dado. Sin embargo antes se debe inicializar los parámetros de los modelos de mezclas de Gaussianas asociados a cada estado, para lo cual se requiere emplear algoritmos de agrupamiento, en este caso se emplea el algoritmo de *k-medias*. Los parámetros adicionales del modelo de Markov tales como matriz de transición y vector de probabilidad inicial se inicializan de forma aleatoria de tal forma que cumplan las restricciones estadísticas.

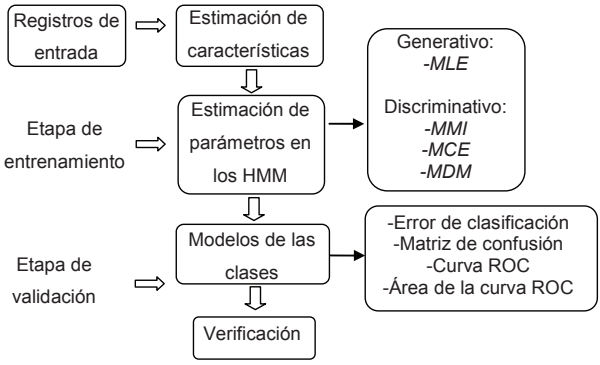


FIG. 4. ESTRUCTURA DEL SISTEMA DE CLASIFICACIÓN

En esta etapa se cuenta con dos enfoques de entrenamiento diferentes, entrenamiento generativo donde el ajuste se realiza mediante la estimación de máxima verosimilitud (MLE) entre el modelo y los datos, y el entrenamiento discriminativo donde se cuentan con diferentes técnicas o criterios siendo los más representativos el criterio de Mínimo Error de Clasificación (MCE) y Máxima Información Mutua (MMI) y donde se incluye el criterio de entrenamiento propuesto.

El siguiente paso es la evaluación del sistema de clasificación para lo cual se debe contar con un conjunto de registros que no han sido tenidos en cuenta en ninguna de las etapas de generación de los modelos. En esta fase de evaluación o validación se tendrá en cuenta básicamente tres indicadores para medir el desempeño del sistema. El indicador más conocido y ampliamente empleado es la precisión o tasa de acierto y se refiere a la porción de patrones clasificados correctamente por el sistema, para calcular la tasa de acierto se utiliza el umbral de mínimo costo (Duda et al., 2001). Utilizando el mismo umbral se calcula la matriz de confusión (Fig. 5), para el caso en el cual se tienen dos clases (0 y 1), se deben calcular los siguientes parámetros (Saénz-Lechon et al., 2006):

Detección correcta o aceptación verdadera (TP, true positive): el número (porcentaje) de patrones de la clase 0 que el clasificador asigna correctamente como pertenecientes a la clase 0, esta medida es llamada también sensibilidad; falso rechazo (FN, false negative): el número (porcentaje) de patrones de clase 0 que el clasificador asigna incorrectamente como pertenecientes a la clase 1; falsa aceptación (FP, false positive): el número (porcentaje) de patrones de clase 1 que el clasificador asigna incorrectamente como pertenecientes a la clase 0; y rechazo correcto o verdadero (TN, true negative): el número (porcentaje) de patrones de clase 1 que el clasificador asigna correctamente como pertenecientes a la clase 1. Esta medida es llamada también especificidad.

		Clase correcta	
		Clase 0	Clase 1
Clase estimada por el clasificador	Clase 0	**VP**	**FP**
	Clase 1	**FN**	**VN**

Fig. 5. Estructura del vector de parámetros

Finalmente se tiene en cuenta la curva ROC, más precisamente el área que encierra dicha curva (ABC). Para darle validez estadística a la prueba y determinar la capacidad de generalización del sistema se emplea un esquema de validación cruzada, con diferentes conjuntos de entrenamiento-validación (*k-fold*), escogidos de forma aleatoria del conjunto completo de datos. En este trabajo se emplean 11 conjuntos, utilizando para el entrenamiento el 70% de los registros y para la validación el 30% restante.

4. Resultados

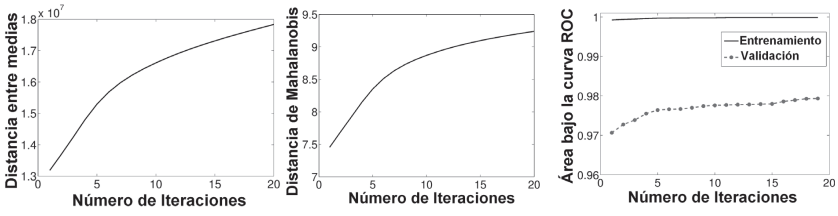
Las pruebas iniciales se realizan con la técnica de entrenamiento estándar (MLE) que constituye la base de comparación para los demás criterios de entrenamiento. El análisis se realiza cambiando la estructura del modelo, es decir, variando el número de estados y el número de Gaussianas por estado (NG), se obtienen los resultados que se muestran en la Tabla 1, donde T.A. denota tasa de acierto.

Para esta base de datos, incrementar la complejidad del modelo no implica un incremento en el desempeño del sistema de clasificación. Por lo tanto y para efectos de comparar el criterio de entrenamiento estándar con los criterios de entrenamiento discriminativos se adopta una arquitectura con 2 estados y 3 Gaussianas por estado, ya que esta configuración presenta el mejor desempeño en la Tabla 1, es necesario aclarar que la elección de arquitectura se hace de forma arbitraria, ya que dicho desempeño no presenta una diferencia estadística significativa con respecto otras arquitecturas.

Tabla 1. Resultados HMM continuo - MLE

NG	Número de estados			
	2	3	5	10
	T.A.	T.A.	T.A.	T.A.
2	94,1 ± 1,1	94,3 ± 2,6	92,7 ± 3,0	84,7 ± 3,0
3	94,6 ± 1,3	91,1 ± 1,8	90,5 ± 2,9	82,3 ± 2,3
4	91,5 ± 3,4	91,4 ± 2,8	90,6 ± 2,8	81,0 ± 2,8

Una vez definida la arquitectura el siguiente paso es realizar la estimación de parámetros empleando el criterio de entrenamiento propuesto. En principio la tarea consiste en maximizar por medio de un algoritmo iterativo la distancia entre las medias de las densidades de probabilidad asociadas a cada clase, en la Fig. 6 se puede observar como la optimización de dicha distancia (a), indirectamente hace que la distancia de Mahalanobis también se maximice (b) y finalmente como se esperaba este proceso se refleja en el desempeño del clasificador, donde se nota un incremento progresivo en el área que encierra la curva ROC (c), especialmente en el conjunto de validación.



a) b) c)

Fig. 6. Proceso de aprendizaje – Optimización de la distancia entre medias

Teniendo en cuenta los resultados anteriores, y seleccionando la mejor arquitectura para MLE, se realiza la comparación entre los diferentes criterios de entrenamiento como se muestra en la Tabla 2, las siglas MDM denota el criterio de entrenamiento propuesto a saber *Máxima Distancia entre Medias*.

Tabla 2. Resultados comparación criterios de entrenamiento

Entrenamiento	ABC	T.A.	M. de C.	
MLE	0,9604±0,02	94,6±1,3	98,5	15,3
			1,50	84,7
MMI	0,9690±0,02	95,8±1,9	99,6	16,4
			0,40	83,6
MCE	0,9701±0,02	96,3±1,4	99,1	13,6
			0,90	83,4
MDM	0,9802±0,01	97,3±1,7	99,8	10,7
			0,20	89,3

En la Tabla 2 se observa que en general todas las técnicas de entrenamiento discriminativo superan la técnica de entrenamiento generativo y adicionalmente que la técnica de entrenamiento propuesta es superior a todas las demás técnicas, mostrando que el ABC y la precisión son los más grandes y una clara superioridad en cuanto a capacidad de generalización.

Además de las cifras mostradas los resultados se complementan con la Fig. 7 donde se muestra las curvas ROC de cada uno de los criterios de entrenamiento que se han tenido en cuenta. Se puede notar la clara diferencia entre los criterios de entrenamiento discriminativo y el criterio de entrenamiento estándar. Las mediciones del área que encierran cada una de las curvas se presentan en la Tabla 2 en la primera columna.

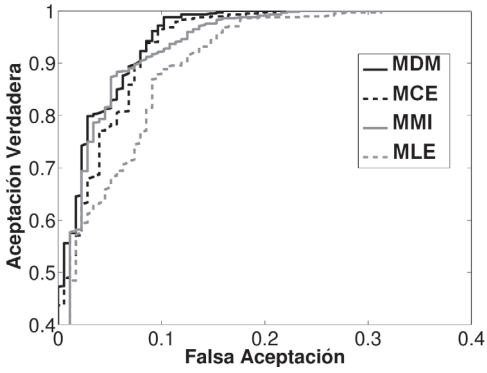


Fig. 7. Curva ROC para cada uno de los criterios de entrenamiento

5. Conclusiones

Se mejora el desempeño de clasificación del método básico de entrenamiento MLE, mediante el uso de un criterio de entrenamiento discriminativo, para el cual se sugiere el empleo de una función de costo que relaciona indirectamente el área que encierra una curva de desempeño, en particular se propone la curva ROC, con una distancia entre modelos de clases.

La función de costo empleada es la distancia entre las medias asociadas a las densidades de probabilidad subyacentes de dos clases, generadas a partir de modelos ocultos de Markov, para su optimización se utiliza un algoritmo iterativo basado en el calculo de gradientes y que tiene en cuenta la naturaleza estocástica del modelo. Mostrando de forma satisfactoria la estrecha relación que existe entre la medida de distancia empleada y el área de la curva ROC.

Las pruebas realizadas presentan como resultado un desempeño satisfactorio empleando una arquitectura HMM relativamente simple, mejorando el desempeño del método de entrenamiento estándar y el de los métodos de entrenamiento discriminativo. Esto demuestra que para mejorar el desempeño de un sistema de detección de patologías además tener un buen conjunto de características, también se debe tener un criterio de entrenamiento adecuado que se enfoque en la generación de una frontera de decisión óptima, buscando de esta forma que no sea necesario incrementar la complejidad del modelo, y que la etapa de entrenamiento sea más eficiente.

6. Agradecimientos

Este trabajo se enmarca dentro del proyecto P09225 financiado por el **Instituto Tecnológico Metropolitano** - Medellín y la Institución Universitaria Salazar Herrera. Y el proyecto 1127-405-20332 financiado por la Universidad Nacional de Colombia – sede Manizales y la Universidad de Caldas.

7. Referencias

Bahl, L.R., Brown, P.F., Souza, P.V., Mercer, R.L., (1986); Maximum mutual information estimation of Hidden Markov Models parameters for speech recognition. Proceedings of the IEEE international Conference on Acoustic, Speech and Signal Processing, 11, 49-52.

Bilmes, J.A., (1998); A gentle tutorial of the EM algorithm and its applications to parameter estimation for Gaussian mixture and Hidden Markov Models. International Computer Science Institute, Bekerly CA, USA.

De Krom, G., (1993); A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. Journal of Speech and Hearing Research, 36(2), 254-266.

Duda, R.O., Hart, P.E., Stork, D.G., (2001); Pattern Classification, Segunda edición. John Wiley & Sons, INC.

Fawcett, T., (2006); An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

Gao, S., Wu, W., Lee, C.H., Chua, T.S., (2003); A Maximal Figure-of-Merit Learning Approach to Text Categorization. Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval. 1, 174-181.

Godino-Llorente, J.I., Gómez-Vilda, P., Sáenz-Lechón, N., Blanco-Velasco, M., Cruz-Roldán, F., Ferrer-Ballester, M.A., (2005); Discriminative methods for the detection of voice disorders. Proceedings of the 3th International Conference on Non-Linear speech processing – NOLISP 2005, 158 - 167.

Hanley, J.A., McNeil, B.J., (1982); The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology, 143(1), 29–36.

Juang, B.H., Katagiri, S., (1992); Discriminative learning for minimum error classification, IEEE Transactions on Signal Processing, 40(12), 3043-3053.

Juang, B.H., Hou, W., Lee, C.H., (1997); Minimum classification error rate methods for speech recognition. IEEE transaction on Speech and Audio Processing, 5(3), 257-265.

Kasuya, H., Ogawa, S., Mashima, K., Ebihara, S., (1986); Normalized noise energy as an acoustic measure to evaluate pathologic voice. Acoustical Society of America, 80(5), 1329-1334.

Li, X., Chang, E., Dai, B.Q., (2002); Improving speaker verification with figure of Merit training. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1, 693- 696.

MEEIVL, (1994); Voice disorders database, version 1.03 *[CDROM]. Lincoln Park, NJ: Kay Elemetrics Corp.*

Michaelis, D., Gramms, T., Strube, H.W., (1997); Glottal to Noise Excitation ratio - a new measure for describing pathological voices. Acta Acustica united with Acustica, 83(4), 700-706.

Rabiner, L., (1989); A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.

Rabiner, L., Juang B.H., (1993); Fundamentals of Speech Recognition. PTR Prentice Hall.

Saénz-Lechon, N., Godino-Llorente, J.I., Osma-Ruiz, V., Gomez-Vilda, P., (2006); Methodological issues in the development of automatic systems for voice pathology detection, Biomedical Signal Processing and Control, 1, 120-128.

Wang, J., Jo, C., (2007); Vocal Folds Disorder Detection using Pattern Recognition Methods. Proceedings of the 29th Annual International Conference of the IEEE EMBS'07, 3253–3256.