Entrenamiento discriminativo de componentes principales ocultas de Markov aplicado a detección de estados funcionales en bioseñales

Bonie Johana Restrepo Cuestas

Universidad Tecnológica de Pereira

Facultad de Ingenierías

Pereira

2009

Entrenamiento discriminativo de componentes principales ocultas de Markov aplicado a detección de estados funcionales en bioseñales

Bonie Johana Restrepo Cuestas

Trabajo de grado para optar al título de
Magíster en Ingeniería Eléctrica

Director
Prof. Julián David Echeverry

Universidad Tecnológica de Pereira
Facultad de Ingenierías
Ingeniería Eléctrica,
Pereira
2009

Contenido

Lista de tablas

1. Resumen

Se propone desarrollar una metodología de carácter discriminativo para la estimación de los parámetros de un modelo de clasificación basado en modelos ocultos de Markov, en donde las observaciones estén explicadas por un modelo de variable latente (componentes principales ocultas de Markov). El modelo de variable latente dinámico obtenido (HMPCA), se utiliza para realizar una reducción de dimensionalidad localizada que permite disminuir el número de características dinámicas necesarias para la etapa de entrenamiento. Con el enfoque de entrenamiento discriminativo, se pretende mejorar la precisión del clasificador de forma directa, implementando la estimación de parámetros mediante técnicas de gradiente estocástico o secuencial para minimizar una función que penaliza el error de clasificación en el contexto de la teoría de decisión bayesiana. Se compara el desempeño del modelo HMPCA optimizado con el criterio discriminativo MCE y el desempeño obtenido con la estimación de máxima verosimilitud. El modelo discriminativo propuesto se emplea en tareas de análisis de señales biológicas con patologías, específicamente, en clasificación de señales de ECG y voz.

2. Introducción

Un modelo oculto de Markov (HMM - *hidden Markov model*) es un proceso doblemente estocástico empleado en la descripción de patrones secuenciales [34]. Se ha utilizado con éxito en aplicaciones de voz [34], [1, 11], en reconocimiento de secuencias de ADN [13] y en el reconocimiento de caracteres escritos [26].

La descripción probabilística del HMM se realiza a través de una cadena de Markov a cuyos estados se asocia una función de probabilidad que describe las observaciones o en forma general un modelo de observación. De esta manera, la descripción completa del HMM requiere de la especificación de tres distribuciones de probabilidad, a saber, la función de probabilidad de estado inicial, la matriz de transición entre los estados de la cadena de Markov y la función de probabilidad o el modelo de observación asociado a cada estado.

En un sistema de clasificación que utiliza teoría de decisión Bayesiana, el HMM puede emplearse bien como un modelo generativo, bien como un modelo discriminativo o bien como una función discriminante. Si se usa como un modelo generativo, el HMM describe una función de probabilidad condicional de las observaciones dadas las clases o, en otras palabras, la verosimilitud de las observaciones. Los parámetros de las funciones de probabilidad asociadas al modelo, se estiman empleando el criterio de máxima verosimilitud a través del algoritmo EM (*Expectation - Maximization*) y se asume que la clase correcta es aquella para la cual la probabilidad condicional es mayor. Si se emplea como modelo discriminativo, el HMM representa la función de probabilidad posterior de la clases dadas las observaciones. Los parámetros del modelo se estiman usando criterios como máxima información mutua (MMI - *maximum mutual information*) [5] o error medio cuadrático (MSE - *mean squared error*) [28] y se asume que la clase correcta es aquella para la cual la probabilidad posterior es mayor. Finalmente, si se emplea como función discriminante, el HMM describe una función que mapea las observaciones a la etiqueta de clase. Los parámetros del modelo se obtienen optimizando un función de costo. Ejemplos de este enfoque incluyen el mínimo error de clasificación (MCE - *minimum classification error*) [22], el entrenamiento correctivo [6] y el criterio de margen máximo [40].

En un problema de clasificación supervisada, emplear el HMM como función de probabilidad condicional (modelo generativo) o como función de probabilidad posterior (modelo discriminativo) no garantiza que se encuentre una relación directa con el error de clasificación [37]. Sin embargo, al emplear el HMM como función discriminante, se busca minimizar el error de clasificación al modelar directamente las superficies de decisión entre clases.

Por otro lado, un problema recurrente en el reconocimiento estadístico de patrones [12] tiene que ver con la dimensionalidad del vector de características que se emplea para representar las observaciones. Este problema se conoce con el nombre de la "maldición de la dimensionalidad" [17] e implica que el número de observaciones necesario para entrenar un clasificador crece de forma exponencial a medida que aumenta la dimensionalidad del espacio de entrada. Una solución aproximada al problema se obtiene al representar el espacio inicial de alta dimensionalidad, con un espacio de dimensión menor que explote ciertas propiedades del espacio de entrada. El análisis de componentes principales (PCA - *principal component analysis*) es una de las técnicas más empleadas para hacer reducción de dimensionalidad. PCA transforma la información de un conjunto conformado por un alto número de variables correlacionadas en un conjunto más pequeño, reteniendo la mayor varianza del conjunto inicial. Aunque el esquema de reducción PCA no cuenta con un modelo generativo asociado, el análisis de componentes probabilístico (PPCA - *probabilistic principal component analysis*) [42] supera este problema y permite utilizar el criterio de máxima verosimilitud para estimar los parámetros del modelo.

PCA y PPCA asumen que las observaciones son independientes y normalmente multivariadas, lo que hace inapropiado su uso en aplicaciones con series de tiempo. Una forma de dar solución a este inconveniente consiste en utilizar una versión temporal del modelo de componentes principales probabilístico conocido como componentes principales ocultas de Markov (HMPCA) [3]; esta metodología consiste en utilizar un HMM que emplea como densidades de observación el modelo de variable latente PPCA. El uso de esta formulación permite obtener una representación optimizada de los datos observados a través del tiempo y mejorar las capacidades de transformación, reducción y clasificación de series de tiempo. En [2], la estimación de los parámetros del modelo HMPCA se realizó empleando el criterio de máxima verosimilitud (ML). Aunque éste método de optimización garantiza convergencia y es eficiente computacionalmente, no asegura que se minimice el error de clasificación; además, requiere de un conjunto de entrenamiento suficientemente robusto para asegurar una estimación óptima de los parámetros del modelo.

Por tanto, se propone un esquema de reconocimiento de patrones con carácter discriminativo basado en el criterio MCE orientado a realizar reconocimiento de patrones en bioseñales, que permita obtener una representación optimizada de los datos observados a través del tiempo, que relacione de forma directa el rendimiento del modelo con el ajuste de sus parámetros; y adicionalmente, permita disminuir el número de características dinámicas necesarias para la etapa de entrenamiento.

3. Objetivos

3.1. Objetivo general

Desarrollar y analizar una metodología de estimación de parámetros a partir de métodos discriminativos, basada en el análisis de su dependencia estadística empleando procesos markovianos y modelos de observación de variable latente, con aplicación en el reconocimiento de disfunciones de bioseñales (voz y ECG).

3.2. Objetivos específicos

1. Implementar el algoritmo de componentes principales ocultas de Markov (basado en máxima verosimilitud).
2. Establecer el modelo discriminativo del algoritmo de componentes principales ocultas de Markov.
3. Estimar los parámetros del modelo discriminativo utilizando técnicas directas de minimización de error como MCE, mediante el algoritmo de gradiente secuencial.
4. Comparar el desempeño de ambos algoritmos en tareas de reconocimiento de patologías y estados funcionales.

4. Marco teórico

Un HMM es un proceso doblemente estocástico empleado en la caracterización de muestras de datos observadas de una serie de tiempo discreta, por lo general de alta dimensionalidad. No sólo es una manera eficiente de construir modelos paramétricos, sino que también incorpora principios de programación dinámica en su núcleo para una unificada clasificación y segmentación de patrones de secuencias de datos variantes con el tiempo. La suposición subyacente del HMM es que las muestras de los datos pueden ser caracterizadas como un proceso aleatorio paramétrico, y los parámetros del proceso estocástico pueden estimarse en un marco de trabajo preciso y adecuadamente definido [17]. Los parámetros de los modelos ocultos de Markov pueden estimarse usando criterios de máxima verosimilitud [34] o usando esquemas de entrenamiento discriminativo [23, 29].

Por otro lado, existen varias razones para mantener la dimensionalidad del espacio de características tan pequeño como sea posible, entre ellas están la reducción en el costo de la medición de las variables, la precisión del clasificador, y la disminución en el número de muestras necesarias para entrenar el clasificador [18, 41]. En el análisis de series de tiempo multivariadas, la clave se encuentra en la determinación de los retardos (*lags*) entre las diferentes variables. Esto difiere de la caracterización del orden o grado de libertad de una serie de tiempo univariada, donde el objetivo es la estimación de la dimensionalidad intrínseca de los datos [18].

Hasta el momento se han utilizado diversos métodos de selección de características en el reconocimiento de bioseñales, todos orientados a trabajar con características estáticas para obtener aquellas con la más alta capacidad discriminante. Por ejemplo, en el procesamiento de patologías de voz, algunos esquemas de selección tales como PCA y el análisis discriminante lineal (LDA - *linear discriminant analysis*) se han utilizado con éxito [30, 19, 15]. Sin embargo, estos esquemas están orientados a características estáticas donde no se tiene en cuenta el cambio dinámico de las mismas.

4.1. Modelos ocultos de Markov

Una cadena de Markov es un proceso aleatorio $\theta(t)$ que puede tomar una cantidad finita K de valores discretos o estados para la representación de una señal aleatoria,

dentro del conjunto $\{\boldsymbol{\vartheta}_1, ..., \boldsymbol{\vartheta}_K\}$, tal que en los momentos determinados del tiempo $(t_0 < t_1 < t_2 < ...)$ los valores del proceso aleatorio cambien (con probabilidades de cambio conocidas), esto es, se efectúan los cambios en forma de secuencia aleatoria $\theta_0 \rightarrow \theta_1 \rightarrow \theta_2, ...,$ siendo $\theta_n = \theta(t_n)$ el valor de la secuencia después del intervalo n de tiempo. En particular, cada estado de manera directa se asocia a un evento físico 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 probabilístico sobre los estados no observables u ocultos. Como resultado las cadenas construidas por este principio, corresponden a un doble proceso estocástico; dados por la función probabilística de los estados ocultos y el mismo modelo de aleatoriedad de Markov impuesto sobre la señal.

Los modelos ocultos de Markov se pueden caracterizar mediante el siguiente conjunto de parámetros [34]:

(a). Los símbolos de observación corresponden a la salida física del sistema en análisis y conforman la secuencia aleatoria $\boldsymbol{\varphi} = \{\varphi_1, \ldots, \varphi_{n_\varphi}\}$ en los momentos definidos de tiempo $1, 2, \ldots, n_\varphi$, donde n_φ es la longitud de la secuencia de observación.

(b). El número de los estados ocultos del modelo, $\boldsymbol{\vartheta} = \{\vartheta_i : i = 1, \ldots, n_\vartheta\} \in \mathfrak{V}$, que siendo no observables, pueden ser relacionados con algún sentido físico del proceso. Los estados ocultos conforman la secuencia aleatoria $\boldsymbol{\theta} = \{\theta_0, \theta_1, \ldots, \theta_{n_\theta}\}$ en los momentos definidos de tiempo $= 1, 2, \ldots, n_\theta$, donde n_θ es la longitud de la secuencia de estados en análisis, que es igual a la longitud de la secuencia de observación $n_\varphi = n_\theta$.

(c). La matriz probabilidad de transición de estados, $\boldsymbol{\Pi} = \{\pi_{ij} : i, j = 1, \ldots, n_\vartheta\}$, en la que cada elemento se determina como sigue

$$\pi_{ij} = P\left(\theta_{n+1} = \vartheta_j | \theta_n = \vartheta_i\right), \ \pi_{ij} \geq 0, \ \sum_{j=1}^{n_\vartheta} \pi_{ij} = 1.$$

(d). El conjunto completo de parámetros que representa la distribución de las observaciones por cada estado del modelo $\mathbf{B} = b_j(\cdot)$. Existen dos formas de distribuciones de salida que pueden ser consideradas. La primera es una suposición de observación discreta donde se asume que una observación es una de n_υ posibles símbolos de observación $\boldsymbol{\upsilon} = \{\upsilon_k : k = 1, \ldots, n_\upsilon\} \in \mathfrak{U}$. En este caso $b_j(\varphi_n = \upsilon_k) = p(\upsilon_k | \theta_n = \vartheta_j)$. La segunda forma de emisión, está conformada por una mezcla de M funciones de distribución para cada estado. Convencionalmente las funciones utilizadas son gaussianas multivariadas, debido a sus

propiedades y a que su tratamiento matemático está descrito de forma amplia en la literatura. En este caso $b_j(\varphi_n) = \sum_{m=1}^{M} c_{jm}\mathcal{N}_{\varphi_n}(\boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm})$; donde $\boldsymbol{\mu}_{jm}$, $\boldsymbol{\Sigma}_{jm}$, y c_{jm} representan respectivamente el vector de medias, la matriz de covarianza y el peso de la componente gaussiana m en el estado j.

Los coeficientes de la mezcla c_{jm} satisfacen la restricción estocástica

$$\sum\nolimits_{m=1}^{M} c_{jm} = 1, \quad 1 \leq j \leq n_\vartheta.$$

(e). El vector probabilidad de estado inicial $\mathbf{p}_{\theta_1}$ con elementos $\{P_{\theta_1}(i)\}$, donde

$$P_{\theta_1}(i) = P\left(\theta_1 = \vartheta_i\right),\ 1 \leqslant i \leqslant n_\vartheta.$$

Los valores de aleatoriedad $\mathbf{\Pi}$, $\mathbf{B}$ y $\mathbf{p}_{\theta_1}$, notados en conjunto como

$$\lambda = \{\mathbf{\Pi}, \mathbf{B}, \mathbf{p}_{\theta_1}\},$$

conforman los parámetros de un modelo oculto de Markov, el cual se puede emplear para generar la estimación de la secuencia de observación, $\boldsymbol{\varphi} = \{\varphi_1, \varphi_2, \ldots, \varphi_{n_\varphi}\}$.

El desarrollo de los HMMs está relacionado con las siguientes tres tareas estadísticas a resolver [17]:

1. La primera, conocida como el problema de evaluación, consiste en calcular de manera eficiente la probabilidad $P(\boldsymbol{\varphi}|\lambda)$ dados la secuencia de observación $\boldsymbol{\varphi} = \{\varphi_1, \varphi_2, \ldots, \varphi_{n_\varphi}\}$ con longitud n_φ y el modelo $\lambda = \{\mathbf{\Pi}, \mathbf{B}, \mathbf{p}_{\theta_1}\}$. La solución se realiza aplicando el algoritmo recursivo de propagación, que analiza el desarrollo de la cadena de estados en el sentido natural del tiempo, y realiza el desdoblamiento de las secuencias de observación, empleando la relación inherente entre los elementos contiguos de las cadenas.
2. La segunda tarea estadística se conoce como el problema de decodificación, que busca elegir de forma óptima la correspondiente secuencia de estados $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_{n_\varphi}\}$; dada una secuencia de observación con longitud n_φ y el modelo conocido λ, para un criterio de medida fijado a priori. Aun cuando no se tiene una solución que de forma universal sea aceptada, el problema puede ser resuelto eficientemente utilizando el algoritmo de Viterbi.
3. La tercera tarea estadística (problema de aprendizaje), busca estimar un conjunto de parámetros que brinden el máximo valor de $P(\boldsymbol{\varphi}|\lambda)$. Este es el problema más difícil de resolver, porque no se conoce ningún método de análisis que maximice la probabilidad conjunta de los datos de entrenamiento en una forma cerrada. Sin embargo, se puede determinar el modelo $\lambda = \{\mathbf{\Pi}, \mathbf{B}, \mathbf{p}_{\theta_1}\}$,

tal que $P(\boldsymbol{\varphi}|\lambda)$ sea maximizada localmente usando procedimientos iterativos de estimación.

4.2. Ajuste de los parámetros del HMM

En forma general, el problema de clasificación puede dividirse en dos tareas importantes: la inferencia y la decisión. En la inferencia se pretende encontrar un modelo de representación que incorpore la incertidumbre de los datos y luego mediante una regla de decisión, basada en la evaluación del modelo, a cada observación se le asigna una clase. La teoría de decisión bayesiana es un análisis estadístico fundamental para resolver el problema de clasificación [12]. Aquí, el problema de clasificación se plantea en términos probabilísticos y se suponen conocidos todos los valores relevantes de probabilidad. Dadas $\nu = 1, \ldots, V$ clases, se asume que existe una probabilidad a priori por clase $P(\nu)$, que hace referencia al conocimiento previo de la probabilidad de ocurrencia de cada clase; en este caso el efecto de los datos observados $\boldsymbol{\varphi}$ es expresado a través de la probabilidad condicional $p(\boldsymbol{\varphi}|\nu)$, conocida como función de verosimilitud; así, la probabilidad conjunta puede ser expresada como $p(\boldsymbol{\varphi}, \nu) = P(\nu|\boldsymbol{\varphi})\, p(\boldsymbol{\varphi}) = p(\boldsymbol{\varphi}|\nu)\, P(\nu)$. Al ordenar la igualdad anterior se puede obtener la fórmula de Bayes:

$$P(\nu|\boldsymbol{\varphi}) = \frac{p(\boldsymbol{\varphi}|\nu)\, P(\nu)}{p(\boldsymbol{\varphi})}, \tag{1}$$

donde el denominador, conocido como la evidencia, es una constante de normalización que asegura que la probabilidad posterior es una densidad de probabilidad válida:

$$p(\boldsymbol{\varphi}) = \sum_{\nu=1}^{V} p(\boldsymbol{\varphi}|\nu)\, P(\nu).$$

La probabilidad conjunta, proporciona un panorama completo de la incertidumbre asociada a las dos variables. La evaluación de dicha probabilidad puede verse en forma general como el problema de inferencia, que usualmente no cuenta con una solución simple. Una vez se resuelva el problema de inferencia, la etapa de decisión puede verse como un paso simple.

El problema de decisión consiste en tomar acciones específicas, basadas en la probabilidad de cada clase condicional, para definir cual de las V clases debe ser asignada a cada uno de los datos observados. Utilizando el teorema de Bayes se puede encontrar dicha probabilidad como se muestra en la ecuación (1), $P(\nu|\boldsymbol{\varphi})$ corresponde a la probabilidad posterior. El objetivo es reducir la posibilidad de asignar una clase incorrecta a cada muestra, es decir, minimizar el error de clasificación; por tanto se elige como la clase correcta aquella que tenga mayor probabilidad posterior.

Existen 3 enfoques distintos para resolver el problema de clasificación [9], como se muestra a continuación:

Modelos generativos: que reciben este nombre debido a que permiten producir datos sintéticos en el espacio de entrada. Aquí el problema de inferencia se soluciona encontrando un modelo de distribución de los datos por clase; y la decisión se toma a partir de la evaluación de la probabilidad de un nuevo dato de entrada dados los parámetros que definen las distribuciones por clase. Este tipo de tratamiento requiere un conjunto de entrenamiento grande que se refleja directamente en un costo computacional elevado.

Modelos discriminativos: este tipo de esquemas en la etapa de inferencia modelan directamente la probabilidad condicional de clase dados los datos y luego utilizando teoría de decisión se le asigna una clase a cada nueva entrada.

Funciones discriminantes: toma las entrada y las relaciona directamente a una clase específica; esta solución combina la inferencia y la decisión en un único problema de aprendizaje, por tanto es más simple que las anteriores.

Cuando se utiliza un HMM para resolver un problema de clasificación; la variación entre los enfoques mencionados, está directamente relacionada con la elección del criterio en que se basa la estimación de parámetros que conforman el modelo [21]. Como se muestra a continuación: para obtener un modelo generativo, la estimación debe apuntar a la maximización de la probabilidad condicional de los datos dados los parámetros que definen cada clase; entre los que se encuentra el criterio ML. Si se requiere un modelo discriminativo, el ajuste de parámetros debe estar asociado a la reducción del error de clasificación indirectamente, entre estas técnicas se pueden nombrar MMI, mínima información discriminante (MDI - *minimum discriminative information*), CMLE y MSE. En el caso de considerar HMM como una función discriminante, la optimización debe reducir de forma directa el error de clasificación; como lo hacen las técnicas MCE y de entrenamiento correctivo. En la tabla 4.1 se resumen los enfoques presentados.

4.2.1. Máxima esperanza EM

El método EM, implica la estimación de máxima verosimilitud y se emplea cuando la información necesaria para estimar los parámetros del modelo está incompleta, como es el caso de los HMMs, en el que se conocen las secuencias de observación, pero no se conocen las secuencias de estados. El algoritmo se compone de dos pasos [10]:

- *Cálculo de la esperanza* o promedio, $\mathscr{Q}(\widetilde{\lambda}|\lambda) = E\{\log P(\boldsymbol{\theta}, \boldsymbol{\varphi}|\widetilde{\lambda})|\boldsymbol{\varphi}, \lambda\},\ \lambda, \widetilde{\lambda} \in \mathfrak{M}$
- *Maximización*, se escogen los valores de $\widetilde{\lambda}$ que maximicen $\mathscr{Q}(\widetilde{\lambda}|\lambda)$

TABLA 4.1. Enfoques para la clasificación

Enfoque	Criterio
Modelos Generativos	Máxima verosimilitud (ML) [7]
Modelos Discriminativos	Máxima verosimilitud condicional (CML) [14] Máxima información mutua (MMI) [5] Error medio cuadrático (MSE) [28] Mínimo riesgo de Bayes (MBR) [24]
Funciones Discriminativas	Entrenamiento correctivo [6] Mínima razón de error empírico [25] Mínimo error de clasificación (MCE) [22] Criterio de margen máximo [40]

donde $\widetilde{\lambda}$ representa el nuevo conjunto de parámetros después de una iteración y $\mathfrak{M}$ es el espacio de los parámetros del modelo. El algoritmo de máxima esperanza se basa en la búsqueda de un conjunto de parámetros $\widetilde{\lambda}$ que maximice $\log P(\boldsymbol{\theta}, \boldsymbol{\varphi}|\widetilde{\lambda})$. Sin embargo, debido a que no se conoce la relación $\log P(\boldsymbol{\theta}, \boldsymbol{\varphi}|\widetilde{\lambda})$ se maximiza su valor esperado actual, dados tanto las observaciones $\boldsymbol{\varphi}$ como el modelo λ. La función $\mathscr{Q}(\widetilde{\lambda}|\lambda)$ se conoce como la función auxiliar.

Con el fin de describir el procedimiento de estimación recursiva de los parámetros del modelo oculto de Markov, se define $\zeta_n(i,j)$ como la probabilidad de estar en el estado ϑ_i en el tiempo n y en el estado ϑ_j en el tiempo $n+1$, dado el modelo y la secuencia de observación

$$\zeta_n(i,j) = P(\theta_n = \vartheta_i, \theta_{n+1} = \vartheta_j|\boldsymbol{\varphi}, \widetilde{\lambda}),$$

que puede reescribirse como

$$\zeta_n(i,j) = \frac{P(\theta_n = \vartheta_i, \theta_{n+1} = \vartheta_j, \boldsymbol{\varphi}|\widetilde{\lambda})}{P(\boldsymbol{\varphi}|\widetilde{\lambda})}.$$

Esta probabilidad se puede relacionar con la medida de probabilidad

$$\gamma_n(i) = P(\theta_n = \vartheta_i|\boldsymbol{\varphi}, \widetilde{\lambda}) = \sum_{j=1}^{n_\vartheta} \zeta_n(i,j), \tag{2}$$

que se define como la probabilidad de encontrarse en el estado ϑ_i en el tiempo n, dada la secuencia observación $\boldsymbol{\varphi}$ y el modelo $\widetilde{\lambda}$.

La probabilidad conjunta de la secuencia de observación $\boldsymbol{\varphi}$ y la secuencia de estados $\boldsymbol{\theta}$ bajo los parámetros del modelo $\widetilde{\lambda}$, está dada por

$$P\left(\boldsymbol{\varphi}, \boldsymbol{\theta}|\widetilde{\lambda}\right) = p_{\theta_1}\left(\theta_1\right) \prod_{n=2}^{n_\varphi} P\left(\theta_n|\theta_{n-1}, \widetilde{\lambda}\right) \prod_{n=1}^{n_\varphi} P\left(\varphi_n|\theta_n, \widetilde{\lambda}\right).$$

El logaritmo de la función de probabilidad conjunta está dado por

$$\log P\left(\boldsymbol{\varphi}, \boldsymbol{\theta}|\widetilde{\lambda}\right) = \log p_{\theta_1}\left(\theta_1\right) + \sum_{n=2}^{n_\varphi} \log P\left(\theta_n|\theta_{n-1}, \widetilde{\lambda}\right) + \sum_{n=1}^{n_\varphi} \log P\left(\varphi_n|\theta_n, \widetilde{\lambda}\right).$$

El algoritmo de máxima esperanza requiere la maximización de $\mathscr{Q}(\widetilde{\lambda}|\lambda)$, el valor esperado del logaritmo de la probabilidad conjunta, donde el valor esperado se toma con respecto a la distribución anterior de los estados dadas las observaciones, $P\left(\boldsymbol{\theta}|\boldsymbol{\varphi}, \lambda\right)$. Así,

$$\mathscr{Q}(\widetilde{\lambda}|\lambda) = \sum_{\boldsymbol{\theta}} P\left(\boldsymbol{\theta}|\boldsymbol{\varphi}, \lambda\right) \log \widetilde{p}_{\theta_1}\left(\theta_1\right) + \sum_{\boldsymbol{\theta}} P\left(\boldsymbol{\theta}|\boldsymbol{\varphi}, \lambda\right) \sum_{n=2}^{n_\varphi} \log P\left(\theta_n|\theta_{n-1}, \widetilde{\lambda}\right) + \\ + \sum_{\boldsymbol{\theta}} P\left(\boldsymbol{\theta}|\boldsymbol{\varphi}, \lambda\right) \sum_{n=1}^{n_\varphi} \log P\left(\varphi_n|\theta_n, \widetilde{\lambda}\right), \quad (3)$$

donde $\sum_{\boldsymbol{\theta}}$ denota la suma sobre todas las posibles secuencias de estados. La ecuación (3) está compuesta de tres términos; el primero para los parámetros de la probabilidad inicial $\mathbf{p}_{\theta_1}$, el segundo para los parámetros de la matriz probabilidad de transición de estados $\boldsymbol{\Pi}$ y el tercero para la matriz de probabilidad del modelo de observación $\mathbf{P}(\boldsymbol{\upsilon}|\boldsymbol{\vartheta})$

$$\mathscr{Q}(\widetilde{\lambda}|\lambda) = \mathscr{Q}(\widetilde{\mathbf{p}}_{\theta_1}|\mathbf{p}_{\theta_1}) + \mathscr{Q}(\widetilde{\boldsymbol{\Pi}}|\boldsymbol{\Pi}) + \mathscr{Q}(\widetilde{\mathbf{P}}(\boldsymbol{\upsilon}|\boldsymbol{\vartheta})|\mathbf{P}(\boldsymbol{\upsilon}|\boldsymbol{\vartheta})). \quad (4)$$

Los términos en la ecuación (4) pueden maximizarse de forma separada como se muestra a continuación:

– La expresión de la estimación recursiva para la distribución inicial de probabilidades $\mathbf{p}_{\theta_1}$ puede obtenerse como

$$\mathscr{Q}(\widetilde{\mathbf{p}}_{\theta_1}|\mathbf{p}_{\theta_1}) = \sum_{\boldsymbol{\theta}} P\left(\boldsymbol{\theta}|\boldsymbol{\varphi}, \lambda\right) \log \widetilde{p}_{\theta_1}\left(\theta_1\right),$$

que equivale a la expresión

$$\mathscr{Q}(\widetilde{\mathbf{p}}_{\theta_1}|\mathbf{p}_{\theta_1}) = \sum_{i=1}^{n_\vartheta} P\left(\theta_1 = \vartheta_i|\boldsymbol{\varphi}, \lambda\right) \log \widetilde{p}_{\theta_1}\left(\theta_1 = \vartheta_i\right),$$

que usando la ecuación (2) es igual a

$$\mathscr{Q}(\widetilde{\mathbf{p}}_{\theta_1}|\mathbf{p}_{\theta_1}) = \sum_{i=1}^{n_\vartheta} \gamma_1(i) \log \widetilde{P}_{\theta_1}(i),$$

sumando el factor de Lagrange ϱ en la expresión anterior, sujeto a la restricción $\sum_i P_{\theta_1}(i) = 1$ y al hacer la derivada igual a cero se obtiene

$$\frac{\partial}{\partial \widetilde{P}_{\theta_1}(i)} \left(\sum_{i=1}^{n_\vartheta} \gamma_1(i) \log \widetilde{P}_{\theta_1}(i) + \varrho \left(\sum_i \widetilde{P}_{\theta_1}(i) - 1 \right) \right) = 0,$$

del desarrollo de la derivada, realizando la suma sobre i para obtener ϱ y resolviendo para $\widetilde{P}_{\theta_1}(i)$, se obtiene la solución [34]

$$\widetilde{P}_{\theta_1}(i) = \gamma_1(i).$$

– La expresión para la actualización de los parámetros de la matriz de transición de estados $\widetilde{\mathbf{\Pi}}$ puede encontrase al maximizar la función

$$\mathscr{Q}(\widetilde{\mathbf{\Pi}}|\mathbf{\Pi}) = \sum_{\boldsymbol{\theta}} P\left(\boldsymbol{\theta}|\boldsymbol{\varphi}, \lambda\right) \sum_{n=2}^{n_\varphi} \log P\left(\theta_n|\theta_{n-1}, \widetilde{\lambda}\right),$$

la suma puede organizarse como

$$\begin{aligned}\mathscr{Q}(\widetilde{\mathbf{\Pi}}|\mathbf{\Pi}) &= \sum_{i=1}^{n_\vartheta} \sum_{j=1}^{n_\vartheta} \sum_{n=2}^{n_\varphi} P\left(\theta_n = \vartheta_j, \theta_{n-1} = \vartheta_i|\boldsymbol{\varphi}, \lambda\right) \log P\left(\theta_n = \vartheta_j|\theta_{n-1} = \vartheta_i, \widetilde{\lambda}\right) \\ &= \sum_{i=1}^{n_\vartheta} \sum_{j=1}^{n_\vartheta} \sum_{n=2}^{n_\varphi} \zeta_n(i,j) \log(\widetilde{\pi}_{ij}),\end{aligned}$$

donde la maximización de la expresión anterior con la restricción $\sum_{j=1}^{n_\vartheta} \pi_{ij} = 1$ conduce a la siguiente fórmula de actualización [34]

$$\widetilde{\pi}_{ij} = \frac{\sum_{n=2}^{n_\varphi} \zeta_n(i,j)}{\sum_{n=2}^{n_\varphi} \gamma_n(i)}. \tag{5}$$

En este caso, durante la etapa de promedio, los valores de $\zeta_n(i,j)$ y $\gamma_n(i,j)$ se estiman usando los parámetros del modelo λ. En la maximización, se utiliza la ecuación (5) para calcular los nuevos parámetros $\widetilde{\lambda}$.

– La expresión para la actualización de los parámetros del modelo de observación $\widetilde{\mathbf{P}}(\boldsymbol{\upsilon}|\boldsymbol{\vartheta})$ puede encontrarse maximizando

$$\mathscr{Q}(\widetilde{\mathbf{P}}(\boldsymbol{\upsilon}|\boldsymbol{\vartheta})|\mathbf{P}(\boldsymbol{\upsilon}|\boldsymbol{\vartheta})) = \sum_{\boldsymbol{\theta}} P\left(\boldsymbol{\theta}|\boldsymbol{\varphi},\lambda\right) \sum_{n=1}^{n_\varphi} \log P\left(\varphi_n|\theta_n,\widetilde{\lambda}\right),$$

la suma anterior puede organizarse como

$$\begin{aligned}\mathscr{Q}(\widetilde{\mathbf{P}}(\boldsymbol{\upsilon}|\boldsymbol{\vartheta})|\mathbf{P}(\boldsymbol{\upsilon}|\boldsymbol{\vartheta})) &= \sum_{i=1}^{n_\vartheta}\sum_{n=1}^{n_\varphi} P\left(\theta_n=\vartheta_i|\boldsymbol{\varphi},\lambda\right) \log P\left(\varphi_n=\upsilon_k|\theta_n=\vartheta_i,\widetilde{\lambda}\right)\\ &= \sum_{i=1}^{n_\vartheta}\sum_{n=1}^{n_\varphi} \gamma_n(i)\log\widetilde{p}_{ik}.\end{aligned}$$

La maximización de la expresión anterior, con la restricción $\sum_{k=1}^{n_\varphi}\widetilde{P}_{ik}=1$, da como resultado

$$\widetilde{P}_{ik} = \frac{\sum\limits_{n=1}^{n_\varphi}\gamma_n(i)\delta(\varphi_n,\upsilon_k)}{\sum\limits_{n=1}^{n_\varphi}\gamma_n(i)},\ \delta(\varphi_n,\upsilon_k) = \begin{cases}1, & \varphi_n=\upsilon_k\\ 0, & \varphi_n\neq\upsilon_k\end{cases}$$

Para distribuciones continuas (mezcla de gaussianas), la forma de la función $\mathscr{Q}$ es ligeramente diferente, las variables ocultas deben incluir no sólo la secuencia de estados, sino también una variable que indica la componente de la mezcla para cada estado en cada tiempo. Por consiguiente, se puede escribir $\mathscr{Q}$ como:

$$\mathscr{Q}(\widetilde{\lambda}|\lambda) = \sum_{\boldsymbol{\theta}}\sum_{\boldsymbol{z}} P\left(\boldsymbol{\theta}|\boldsymbol{\varphi},\lambda\right)\log P\left(\boldsymbol{\theta},\boldsymbol{\varphi},\boldsymbol{z}|\widetilde{\lambda}\right).$$

Siendo $\boldsymbol{z} = \left\{\boldsymbol{z}_{\boldsymbol{\theta}_1}^{k_1}, \boldsymbol{z}_{\boldsymbol{\theta}_2}^{k_2}, \boldsymbol{z}_{\boldsymbol{\theta}_3}^{k_3}, \ldots, \boldsymbol{z}_{\boldsymbol{\theta}_{n_{\boldsymbol{\theta}}}}^{k_{n_{\boldsymbol{\theta}}}}\right\}$ el vector que indica la componente de mezcla para cada estado y cada tiempo. Al expandirse la función $\mathscr{Q}$, los términos que hacen referencia a la actualización de la probabilidad de estado inicial y la probabilidad de transición de estados no cambian, debido a que no dependen de $\boldsymbol{z}$. Las relaciones de estimación recursiva para los coeficientes de las densidades de mezcla, c_{jm}, $\boldsymbol{\mu}_{jm}$, y $\boldsymbol{\Sigma}_{jm}$, para el caso de FDP gaussianas, empleando el algoritmo de máxima esperanza,

demuestran ser iguales a [8]

$$\widetilde{c}_{jm} = \frac{\sum_{n=1}^{n_\varphi} \gamma_n(j,m)}{\sum_{n=1}^{n_\varphi} \sum_{m=1}^{M} \gamma_n(j,m)}, \tag{6a}$$

$$\widetilde{\boldsymbol{\mu}}_{jm} = \frac{\sum_{n=1}^{n_\varphi} \gamma_n(j,m)\varphi_n}{\sum_{n=1}^{n_\varphi} \gamma_n(j,m)}, \tag{6b}$$

$$\widetilde{\boldsymbol{\Sigma}}_{jm} = \frac{\sum_{n=1}^{n_\varphi} \gamma_n(j,m)\ (\varphi_n - \boldsymbol{\mu}_{jm})(\varphi_n - \boldsymbol{\mu}_{jm})^{\mathsf{T}}}{\sum_{n=1}^{n_\varphi} \gamma_n(j,m)}, \tag{6c}$$

donde $\gamma_n(j,m)$ es la probabilidad de la observación $\boldsymbol{\varphi}_n$ dada por la componente de mezcla m del estado j, esto es:

$$\gamma_n(j,k) = \gamma_n(j) \left(\frac{c_{jm} \mathcal{N}_{\boldsymbol{\varphi}_n} (\boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm})}{\sum_{m=1}^{M} c_{jm} \mathcal{N}_{\boldsymbol{\varphi}_n} (\boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm})} \right). \tag{7}$$

El término $\gamma_n(j,m)$ generaliza la variable $\gamma_n(j)$ en la ecuación (2) que es el caso de una mezcla simple (una sola componente) o de una función de probabilidad discreta. En el algoritmo de máxima esperanza, en la estimación del promedio se calcula $\gamma_n(j,m)$ usando el conjunto de parámetros λ, mientras en la maximización se usan las ecuaciones (6a), (6b) y (6c) para estimar $\widetilde{\lambda}$.

La estimación recursiva de los coeficientes c_{jm} corresponde a la relación entre el número esperado de veces que el sistema está en el estado j empleando la componente de mezcla m, y el número esperado de veces que el sistema está en el estado j. De forma similar, la estimación recursiva para el vector medio $\boldsymbol{\mu}_{jm}$ pondera cada término del numerador de la ecuación (6b) por la observación, por lo que resulta un valor esperado de la porción del vector de observación que describe la componente de mezcla m. Así mismo, ocurre la interpretación para la matriz de covarianza $\boldsymbol{\Sigma}_{jm}$ obtenida.

4.2.2. Error de clasificación mínimo

El método de entrenamiento empleando el criterio MCE, introducido en [22] y extendido para HMM en [23], busca minimizar la probabilidad de error a través de una representación suavizada de la función de pérdida (*loss function*) que se asume igual a

cero en el caso de que la decisión de pertenencia una clase específica sea correcta, o uno de lo contrario. Esto se hace a través de la llamada medida del error de clasificación, que representa simplemente una medida de la distancia entre la probabilidad de una decisión correcta y otras decisiones. Existen varias definiciones de esta medida, una de las cuales está dada por:

$$d_\nu(\boldsymbol{\varphi}) = -g_\nu(\boldsymbol{\varphi}; \lambda^{(\nu)}) + \log \left\{ \frac{1}{V-1} \sum_{w, w \neq \nu} \exp\left(g_w(\boldsymbol{\varphi}; \lambda^{(w)}) \eta \right) \right\}^{1/\eta}, \tag{8}$$

donde η es un número positivo y $g_\nu = P(\boldsymbol{\varphi} \left| \lambda^{(\nu)} \right.)$ es la función de verosimilitud condicional de que la clase ν haya generado la observación $\boldsymbol{\varphi}$; y $\nu = \{1, \ldots, V\}$, con V el número total de clases. Para el caso de HMM puede ser utilizada la probabilidad conjunta estado-observación definida por:

$$g_\nu(\boldsymbol{\varphi}; \lambda^{(\nu)}) = \log \left\{ \max_{\boldsymbol{\theta}} g_\nu\left(\boldsymbol{\varphi}, \boldsymbol{\theta}; \lambda\right) \right\} = \log \left\{ g_\nu\left(\boldsymbol{\varphi}, \bar{\boldsymbol{\theta}}; \lambda\right) \right\},$$

donde $\bar{\boldsymbol{\theta}}$ corresponde a la secuencia de estados más probable en el modelo para una secuencia de observación dada, y puede ser calculada mediante el algoritmo de Viterbi [34]. La función de verosimilitud condicional está dada por:

$$g_\nu(\boldsymbol{\varphi}; \lambda^{(\nu)}) = \frac{1}{n_{\boldsymbol{\varphi}}} \left(\sum_{n=1}^{n_{\boldsymbol{\varphi}}} \left[\log \pi^{(\nu)}_{\bar{\boldsymbol{\theta}}_{n-1} \bar{\boldsymbol{\theta}}_n} + \log b^{(\nu)}_{\bar{\boldsymbol{\theta}}_n}\left(\varphi_n\right) \right] + \log P^{(\nu)}_{\bar{\boldsymbol{\theta}}_1} \right). \tag{9}$$

La escala $\frac{1}{n_{\boldsymbol{\varphi}}}$ en la ecuación (9) permite normalizar la función de verosimilitud con respecto a la duración de las secuencias [36].

La medida de error de clasificación es una función continua de los parámetros del decodificador e intenta emular la regla de decisión. Para una secuencia de observaciones $\boldsymbol{\varphi}$ que pertenezca a la clase ν, $d_\nu\left(\boldsymbol{\varphi}\right) > 0$ implica un error en la clasificación y $d_\nu\left(\boldsymbol{\varphi}\right) << 0$ implica una decisión correcta. Para finalizar la definición de la función objetivo [22], la medida de distancia definida en la ecuación (8) es embebida en una función zero-uno suavizada (que representa la función de pérdida)

$$\ell\left(d_\nu\left(\boldsymbol{\varphi}\right)\right) = \frac{1}{1 + \exp(-\rho d_\nu(\boldsymbol{\varphi}) + \omega)}, \tag{10}$$

en la función sigmoidal de la ecuación (10), ω es igual a 0 y $\rho \geq 1$. Para cualquier secuencia desconocida $\boldsymbol{\varphi}$, el desempeño del sistema se mide mediante el siguiente criterio

$$\ell\left(\boldsymbol{\varphi}; \lambda\right) = \sum_{\nu=1}^{V} \ell\left(d_\nu\left(\boldsymbol{\varphi}\right)\right) \mathbf{1}\left(\boldsymbol{\varphi} \in C_\nu\right),$$

donde $\mathbf{1}(\cdot)$ es una función de indicación, que es 1 si la observación evaluada pertenece a la clase ν o 0 de otra forma. Esta definición en tres etapas simula la operación de decodificación a la vez que la evaluación del desempeño en un forma funcional suavizada, adecuada para la optimización de los parámetros del clasificador/decodificador. Con base en el criterio anterior, se puede elegir minimizar una de dos cantidades para el cálculo de los parámetros del decodificador: la pérdida esperada o la pérdida empírica [23]. Esto se logra mediante técnicas de gradiente descendente; en particular el algoritmo descendente probabilístico generalizado GPD (*gradient probabilistic descent*) [22]. El algoritmo GPD puede ser generalizado de la siguiente forma

$$\lambda_{t+1} = \lambda_t - \epsilon_t U_t \nabla \ell \left(\boldsymbol{\varphi}_t, \lambda \right) |_{\lambda=\lambda_t}, \tag{11}$$

donde U_t es una matriz definida positiva y ϵ_t es la tasa de aprendizaje [23], el subindice t hace referencia al valor presente del parámetro λ y $t+1$ a la actualización. Una de las ventajas del algoritmo de minimización basado en GPD es que este no hace ninguna suposición explícita sobre las probabilidades desconocidas, propiedad importante para resolver problemas de reconocimiento y aprendizaje adaptativo. En particular, el algoritmo GPD es un esquema de minimización sin restricciones, por tanto requiere modificaciones para realizar el entrenamiento de parámetros para HMM. Se deben definir las siguientes transformaciones de parámetros que permiten mantener las restricciones probabilísticas de los parámetros de los HMMs durante la adaptación. Para los elementos que conforman la matriz de transición:

$$\pi_{ij} \rightarrow \tilde{\pi}_{ij} \;\text{ donde }\; \pi_{ij} = \frac{e^{\tilde{\pi}_{ij}}}{\sum_{j=1}^{n_\vartheta} e^{\tilde{\pi}_{ij}}},$$

y para cada elemento del vector de probabilidad inicial, la transformación está dada por

$$P_{\boldsymbol{\theta}_1}(j) \rightarrow \tilde{P}_{\boldsymbol{\theta}_1}(j) \;\text{ donde }\; P_{\boldsymbol{\theta}_1}(j) = \frac{e^{\tilde{P}_{\boldsymbol{\theta}_1}(j)}}{\sum_{j=1}^{n_\vartheta} e^{\tilde{P}_{\boldsymbol{\theta}_1}(j)}}.$$

En el caso de utilizar emisiones continuas, asumiendo que la matriz de covarianza es diagonal $\Sigma_{jm} = [\sigma^2_{jml}]_{l=1}^{L}$; siendo L la longitud de una observación, la adaptación de

parámetros de las componentes gaussianas del modelo es igual a:

$$\begin{aligned}
\mu_{jml} \rightarrow \tilde{\mu}_{jml} \quad & \text{donde,} \quad \tilde{\mu}_{jml} = \frac{\mu_{jml}}{\sigma_{jml}}, \\
\sigma_{jml} \rightarrow \tilde{\sigma}_{jml} \quad & \text{donde,} \quad \tilde{\sigma}_{jml} = \log \sigma_{jml}, \\
c_{jm} \rightarrow \tilde{c}_{jm} \quad & \text{donde,} \quad c_{jm} = \frac{e^{\tilde{c}_{jm}}}{\sum_{m=1}^{M} e^{\tilde{c}_{jm}}},
\end{aligned}$$

con $\mu_{jm} = [\mu_{jml}]_{l=1}^{L}$, con el mismo tratamiento por clase. Se puede mostrar que para una secuencia de observaciones $\boldsymbol{\varphi}_t \in C_\nu$ del conjunto de entrenamiento conformado por T secuencias, el ajuste discriminativo del parámetro $\tilde{\pi}$ obtenido a partir del algoritmo descendente probabilístico generalizado, está dado por:

$$\tilde{\pi}_{ij}^{(\nu)}(t+1) = \tilde{\pi}_{ij}^{(\nu)}(t) - \epsilon \frac{\partial \ell_\nu(\boldsymbol{\varphi}_t; \lambda)}{\partial \tilde{\pi}_{ij}^{(\nu)}}\bigg|_{\lambda=\lambda_n}. \tag{13}$$

La derivada parcial de la ecuación (13), puede ser resuelta utilizando la regla de la cadena

$$\frac{\partial \ell_\nu\left(\boldsymbol{\varphi}_t; \lambda\right)}{\partial \tilde{\pi}_{ij}^{(\nu)}} = \frac{\partial \ell_\nu\left(d_\nu\right)}{\partial d_\nu} \frac{\partial d_\nu}{\partial \tilde{\pi}_{ij}^{(\nu)}},$$

la derivada de la función de pérdida (10) con respecto a la medida de distancia definida en la ecuación (8) asociada a cada clase es igual a

$$\frac{\partial \ell_\nu\left(d_\nu\right)}{\partial d_\nu} = \frac{\rho \exp\left(-\rho d_\nu + \omega\right)}{\left(1 + \exp\left(-\rho d_\nu + \omega\right)\right)^2} = \rho \ell_\nu\left(d_\nu\right)\left[1 - \ell_\nu\left(d_\nu\right)\right], \tag{14}$$

luego, el segundo término que conforma la derivada (13) es evaluado como sigue

$$\frac{\partial d_\nu}{\partial \tilde{\pi}_{ij}^{(\nu)}} = -\sum_{n=1}^{n_\varphi} \delta\left(\bar{\theta}_{n-1}, i\right) \delta\left(\bar{\theta}_n, j\right) \frac{\partial \log \pi_{ij}^{(\nu)}}{\partial \tilde{\pi}_{ij}^{(\nu)}},$$

siendo $\delta(\cdot)$ la función delta de Kronecker, que es igual a 1 si los argumentos son iguales y 0 de lo contrario; para finalizar, se encuentra la derivada con respecto al parámetro transformado

$$\frac{\partial \log \pi_{ij}^{(\nu)}}{\partial \tilde{\pi}_{ij}^{(\nu)}} = \left(1 - \pi_{ij}^{(\nu)}\right).$$

Realizando un procedimiento similar se puede encontrar la actualización del vector de probabilidad de estado inicial:

$$\frac{\partial d_\nu}{\partial \tilde{P}_{\theta_1}^{(\nu)}(j)} = \frac{\partial \log P_{\theta_1}^{(\nu)}(\bar{\theta}_1)}{\partial \tilde{P}_{\theta_1}^{(\nu)}(j)} = \delta\left(\bar{\theta}_1, j\right) \frac{\partial \log P_{\theta_1}^{(\nu)}(j)}{\partial \tilde{P}_{\theta_1}^{(\nu)}(j)}, \tag{15}$$

$$\frac{\partial \log P_{\theta_1}^{(\nu)}(j)}{\tilde{P}_{\theta_1}^{(\nu)}(j)} = 1 - P_{\theta_1}^{(\nu)}(j).$$

Las derivadas asociadas a la actualización de los parámetros de las mezclas gaussianas del modelo, se presentan a continuación:

$$\frac{\partial \ell_\nu\left(\boldsymbol{\varphi}_t; \lambda\right)}{\partial \tilde{\mu}_{jml}^{(\nu)}} = \frac{\partial \ell_\nu\left(d_\nu\right)}{\partial d_\nu} \frac{\partial d_\nu}{\partial \tilde{\mu}_{jml}^{(\nu)}}.$$

Siendo $\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_t, \ldots, \boldsymbol{\varphi}_T\}$, el conjunto total de secuencias de observación, y cada secuencia $\boldsymbol{\varphi} = \left\{\varphi_1, \ldots, \varphi_n, \ldots, \varphi_{n_\varphi}\right\}$. El primer término que define la derivada anterior, es igual al obtenido en la ecuación (14) y el segundo se muestra a continuación

$$\frac{\partial d_\nu}{\partial \tilde{\mu}_{jml}^{(\nu)}} = \frac{\partial g_\nu(\boldsymbol{\varphi}; \lambda)}{\partial \tilde{\mu}_{jml}^{(\nu)}} = -\sum_{n=1}^{n_\varphi} \delta\left(\bar{\theta}_n, \vartheta_j\right) \frac{\partial \log\left(b_j^{(\nu)}\left(\varphi_n\right)\right)}{\partial \tilde{\mu}_{jml}^{(\nu)}}.$$

Realizando un estudio similar para los elementos que conforman la matriz de covarianza y los pesos para cada componente de mezcla, se obtienen las siguientes derivadas:

$$\frac{\partial \log\left(b_j^{(\nu)}\left(\varphi_n\right)\right)}{\partial \tilde{\mu}_{jml}^{(\nu)}} = \frac{1}{b_j^{(\nu)}\left(\varphi_n\right)} c_{jm}^{(\nu)} \mathcal{N}_{\varphi_n}\left(\boldsymbol{\mu}_{jm}^{(\nu)}, \boldsymbol{R}_{jm}^{(\nu)}\right) \left(\frac{\varphi_{n_l}}{\sigma_{jml}^{(\nu)}} - \tilde{\mu}_{jml}^{(\nu)}\right), \tag{16a}$$

$$\frac{\partial \log\left(b_j^{(\nu)}\left(\varphi_n\right)\right)}{\partial \tilde{\sigma}_{jml}^{(\nu)}} = \frac{1}{b_j^{(\nu)}\left(\varphi_n\right)} c_{jm}^{(\nu)} \mathcal{N}_{\varphi_n}\left(\boldsymbol{\mu}_{jm}^{(\nu)}, \boldsymbol{R}_{jm}^{(\nu)}\right) \left[\left(\frac{\varphi_{n_l} - \mu_{jml}^{(\nu)}}{\sigma_{jml}^{(\nu)}}\right)^2 - 1\right], \tag{16b}$$

$$\frac{\partial \log\left(b_j^{(\nu)}\left(\varphi_n\right)\right)}{\partial \tilde{c}_{jm}^{(\nu)}} = \frac{1}{b_j^{(\nu)}\left(\varphi_n\right)} \mathcal{N}_{\varphi_n}\left(\boldsymbol{\mu}_{jm}^{(\nu)}, \boldsymbol{R}_{jm}^{(\nu)}\right) c_{jm}^{(\nu)}\left(1 - c_{jm}^{(\nu)}\right). \tag{16c}$$

4.2.3. Criterio de margen máxima

El principio de margen máxima fue propuesto originalmente como un paradigma para la clasificación; su idea central es separar los resultados de clasificación correctos de los incorrectos, con una margen positiva. Los algoritmos de aprendizaje utilizados ajustan los parámetros de cada modelo con el fin de maximizar dicha margen. Este estudio está motivado por la teoría de aprendizaje estadístico, como una forma de equilibrar la complejidad del modelo y el error de generalización; y muchos algoritmos de aprendizaje

utilizados para su entrenamiento están enmarcados dentro de la optimización convexa (además de ser eficientes, no se ven afectados por óptimos locales).

En [40] se extiende su uso a mezclas de modelos gaussianos y a modelos ocultos de Markov con el fin de desarrollar algoritmos discriminativos de margen máxima con aplicación al reconocimiento de voz. Para un HMM con densidades de emisión continuas la siguiente función discriminante calcula la relación entre la secuencia de estados $\boldsymbol{\theta}$ y la secuencia de observación $\boldsymbol{\varphi}$:

$$\mathcal{D}\left(\boldsymbol{\varphi}, \boldsymbol{\theta}\right) = \sum_{n_{\boldsymbol{\varphi}}} \left[\varkappa\left(\theta_{n-1}, \theta_n\right) + \Upsilon\left(\varphi_n, \theta_n\right)\right],$$

en términos de la probabilidad de transición entre estados $\varkappa\left(\theta_{n-1}, \theta_n\right)$ y de la densidad de emisión de estados $\Upsilon\left(\varphi_n, \theta_n\right)$. Las densidades de emisión de salida están dadas por:

$$\Upsilon\left(\varphi_n, \theta_n\right) = \log \sum_{M} e^{\mathcal{Z}_n^T \Phi_{\theta_n m} \mathcal{Z}_n},$$

donde $\Phi_{\theta_n m}$ es una matriz semidefinida positiva y es equivalente a una reparametrización del modelo gaussiano por estado y componente de mezcla, como se muestra a continuación:

$$\Phi_{\theta_n m} = \begin{bmatrix} \boldsymbol{\Sigma}_{\theta_n m} & -\boldsymbol{\Sigma}_{\theta_n m} \boldsymbol{\mu}_{\theta_n m} \\ -\boldsymbol{\mu}_{\theta_n m}^T \boldsymbol{\Sigma}_{\theta_n m} & -\boldsymbol{\mu}_{\theta_n m}^T \boldsymbol{\Sigma}_{\theta_n m} + \boldsymbol{\mu}_{\theta_n m} + k_{\theta_n m} \end{bmatrix},$$

con $k_{\theta_n m}$, un escalar no negativo y siendo

$$\mathcal{Z}_n = \begin{bmatrix} \varphi_n \\ 1 \end{bmatrix}.$$

4.3. Reducción de dimensionalidad

PCA es una técnica ampliamente utilizada en la reducción de dimensionalidad aplicado al análisis multivariado. Sus aplicaciones incluyen áreas como compresión, análisis y visualización de imágenes, reconocimiento de patrones, regresión y predicción de series de tiempo. En forma general, se puede describir como una proyección lineal del espacio de entrada sujeto a la maximización de la varianza de los datos. Debido a que PCA no cuenta con un modelo generativo, se desarrolló PPCA [42]; que considera un modelo de variable latente para los vectores de observación y asume una matriz de covarianza isotrópica para el modelo del ruido. Los parámetros del modelo de PPCA se pueden estimar utilizando el criterio de máxima verosimilitud.

4.3.1. Análisis de componentes principales

Dado un conjunto de vectores observados $\{\boldsymbol{\xi}_n\}$ con $n = 1, \ldots, N$ y dimensión $p \times 1$. La matriz de covarianza muestral puede calcularse como [20]:

$$\Sigma_{\boldsymbol{\xi}} = \sum_n (\boldsymbol{\xi}_n - \boldsymbol{\mu}_{\boldsymbol{\xi}}) (\boldsymbol{\xi}_n - \boldsymbol{\mu}_{\boldsymbol{\xi}})^{\mathsf{T}}, \tag{17}$$

con $\boldsymbol{\mu}_\xi = E\{\boldsymbol{\xi}\}$ la media muestral de los datos. En general, los elementos que conforman los vectores están correlacionados, esto es, $\exists\, P(\boldsymbol{\xi}_m | \boldsymbol{\xi}_k) \neq 0, \forall\, m \neq k$.

El objetivo es proyectar los datos en un espacio con dimensionalidad $q < p$ mientras se máximiza la varianza de los datos proyectados; es decir, obtener una variable $\boldsymbol{\zeta}$, que represente de forma simple la información estadística considerada importante y contendida en $\boldsymbol{\xi}$. Los q ejes principales de la reducción $\boldsymbol{A}_j$ [9], con $j \in \{1, \ldots\ldots, q\}$ son ejes ortonormales representados por los q vectores propios dominantes (que están asociados directamente con los q valores propios más grandes) de la matriz de covarianza muestral de la ecuación (17), tal que $\Sigma_{\boldsymbol{\xi}} \boldsymbol{A}_j = \lambda_j \boldsymbol{A}_j$; siendo λ_j el $j-$ésimo valor propio de la matriz de covarianza muestral.

La representación reducida q-dimensional del vector de datos observados se obtiene como se muestra:

$$\boldsymbol{\zeta}_n = \mathbf{C}^T (\boldsymbol{\xi}_n - \boldsymbol{\mu}_\xi),$$

con $\mathbf{C} = (\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_q)$. La reconstrucción óptima de los datos observados se puede calcular como

$$\boldsymbol{\xi}_n = \mathbf{C}\boldsymbol{\zeta}_n + \boldsymbol{\mu}_\xi.$$

La estimación de los parámetros se realiza sobre una muestra $\mathbf{X}_{N \times p}$ compuesta por la matriz de las observaciones $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^{\mathsf{T}}$, donde cada observación es un vector con dimensión $\mathbf{x}_{p \times 1}$. La media se estima usando

$$\boldsymbol{\mu}_{\mathbf{X}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i.$$

En calidad de estimación de la función de covarianza se toma el valor

$$\tilde{\Sigma}_{\boldsymbol{\xi}} = \Sigma_{\mathbf{X}} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{X}})^{\mathsf{T}}.$$

4.3.2. Análisis de componentes principales probabilísticas

Se desea obtener una representación reducida del conjunto de datos $\boldsymbol{\xi}$, mediante la siguiente combinación lineal de vectores base y ruido:

$$\boldsymbol{\xi}_n = \mathbf{C}\boldsymbol{\zeta}_{\boldsymbol{n}} + \boldsymbol{\mu}_{\boldsymbol{\xi}} + \boldsymbol{\epsilon}, \tag{18}$$

donde, la matrix $\mathbf{C}$ de tamaño $(p \times q)$ es la base y $\boldsymbol{\mu_\xi}$ permite que el modelo tenga media diferente de cero. Convencionalmente $p(\boldsymbol{\zeta}) \sim \mathcal{N}(0, \mathbf{I})$ y $p(\boldsymbol{\epsilon}) \sim \mathcal{N}(0, \Psi)$, que corresponde a un modelo de error o ruido isotrópico con varianzas residuales $\Psi = \sigma_\epsilon^2 I$ [42].

El objetivo de la formulación probabilística de PCA es estimar la base $\mathbf{C}$ y la varianza del ruido σ_ϵ^2 a partir del conjunto $\boldsymbol{\xi}$. La ecuación (18) implica que la probabilidad de observar $\boldsymbol{\xi}$ dado $\boldsymbol{\zeta}$ es:

$$p(\boldsymbol{\xi}|\boldsymbol{\zeta}, \mathbf{C}, \boldsymbol{\mu_\xi}, \sigma_\epsilon^2) \sim \mathcal{N}(\mathbf{C}\boldsymbol{\zeta} + \boldsymbol{\mu_\xi}, \sigma_\epsilon^2 \mathbf{I}).$$

Integrando sobre $\boldsymbol{\zeta}$,

$$\begin{aligned} p(\boldsymbol{\xi}|\mathbf{C}, \boldsymbol{\mu_\xi}, \sigma_\epsilon^2) &= \int_{\boldsymbol{\zeta}} p(\boldsymbol{\xi}|\boldsymbol{\zeta}, \mathbf{C}, \boldsymbol{\mu_\xi}, \sigma_\epsilon^2) p(\boldsymbol{\zeta}) d\boldsymbol{\zeta} \\ &\sim \mathcal{N}(\boldsymbol{\mu_\xi}, \mathbf{C}\mathbf{C}^\mathsf{T} + \sigma_\epsilon^2 \mathbf{I}). \end{aligned}$$

La verosimilitud logarítmica de $p(\boldsymbol{\xi}|\mathbf{C}, \boldsymbol{\mu_\xi}, \sigma_\epsilon^2)$ está dada como

$$\Lambda = \log(p(\boldsymbol{\xi}|\mathbf{C}, \boldsymbol{\mu_\xi}, \sigma_\epsilon^2)) = -\frac{N}{2}\{\log|\mathbf{D}| + tr(\mathbf{D}^{-1}\Sigma_{\boldsymbol{\xi}})\}, \tag{19}$$

donde,

$$\begin{aligned} \mathbf{D} &= \mathbf{C}\mathbf{C}^\mathsf{T} + \sigma_\epsilon^2 \mathbf{I}, \\ \Sigma_{\boldsymbol{\xi}} &= E\left\{(\boldsymbol{\xi} - \boldsymbol{\mu_\xi})(\boldsymbol{\xi} - \boldsymbol{\mu_\xi})^\mathsf{T}\right\}, \end{aligned}$$

y se han omitido algunos términos constantes en la expresión (19).

Al igual que en el caso de PCA, la estimación de los parámetros necesarios en la representación (18) se realiza sobre una muestra $\mathbf{X}_{N\times p}$ compuesta por la matriz de las observaciones $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\mathsf{T}$, donde cada observación es un vector con dimensión $\mathbf{x}_{p\times 1}$. Haciendo $\boldsymbol{\mu_X} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i$ y $\Sigma_{\mathbf{X}} = \frac{1}{N}\sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu_x})(\mathbf{x}_i - \boldsymbol{\mu_x})^\mathsf{T}$, los parámetros en (18) se estiman a partir de la maximización iterativa de Λ [42],

$$\widetilde{\mathbf{C}}_{t+1} = \Sigma_{\mathbf{X}}\widetilde{\mathbf{C}}_t(\sigma_{\epsilon,n}^2 \mathbf{I} + \mathbf{M}_t^{-1}\widetilde{\mathbf{C}}_n^\mathsf{T}\sigma_{\mathbf{X}}\widetilde{\mathbf{C}}_t)^{-1}, \tag{20}$$

$$\widetilde{\sigma}_{\epsilon,t}^2 = \frac{1}{p}\mathrm{tr}(\sigma_{\mathbf{X}} - \Sigma_{\mathbf{X}}\mathbf{M}_n^{-1}\widetilde{\mathbf{C}}_t^\mathsf{T}), \tag{21}$$

donde

$$\mathbf{M}_t = \widetilde{\mathbf{C}}_t^\mathsf{T}\widetilde{\mathbf{C}}_t + \sigma_{\epsilon,t}^2 \mathbf{I}. \tag{22}$$

Las expresiones en la ecuaciones (20) y (21) se repiten hasta que se considere que el algoritmo converge. El máximo de (19) ocurre cuando [42]

$$\widetilde{\mathbf{C}}_{ML} = \mathbf{A}_{\mathbf{X}}(\mathbf{V}_q - \sigma_\epsilon^2 \mathbf{I})^{1/2}\mathbf{R},$$

donde los q vectores columna de la matriz $\mathbf{A_X}$ son los vectores propios $\widetilde{\mathbf{\Lambda}}_l$ de $\sigma_{\mathbf{x}}$ con los correspondientes $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_q$ en la matriz diagonal $\mathbf{V}_q$ y $\mathbf{R}$ es una matriz de rotación ortogonal arbitraria $q \times q$. Además, puede mostrarse que para $\widetilde{\mathbf{C}} = \widetilde{\mathbf{C}}_{ML}$, el estimador de máxima verosimilitud para $\widetilde{\sigma}_{\epsilon}^2$ está dado por

$$\widetilde{\sigma}_{\epsilon,ML}^2 = \frac{1}{p-q} \sum_{i=q+1}^{p} \widetilde{\lambda}_i,$$

donde $\widetilde{\lambda}_{q+1}, \ldots, \widetilde{\lambda}_p$ son los valores propios más pequeños de $\sigma_{\mathbf{X}}$ y $\sigma_{\epsilon,ML}^2$ se interpreta como la varianza promedio de las varianzas de las dimensiones que no se tienen en cuenta. Obtenidos los valores estimados, se puede encontrar la representación reducida de los datos como sigue:

$$\boldsymbol{\zeta}_n = \mathbf{M}^{-1}\mathbf{C}^T \left(\boldsymbol{\xi}_n - \boldsymbol{\mu}_{\boldsymbol{\xi}}\right), \tag{23}$$

siendo $\mathbf{M}$ la matriz expresada en la ecuación (22).

4.4. Densidades de observación en HMM

En general, un HMM asume que cada estado genera observaciones de acuerdo a un modelo gaussiano o mezclas de estos modelos. A continuación se muestran algunas variaciones en las observaciones que conforman el modelo; en cada uno de los casos presentados se puede aplicar el algoritmo EM para la estimación de parámetros.

4.4.1. Observaciones autoregresivas multivariadas

En este caso un modelo autoregresivo multivariado MAR representa las observaciones, es decir, se utiliza un predictor lineal para modelar series de tiempo múltiples [33]. Un modelo $MAR(p)$ predice el siguiente vector de valores en una serie de tiempo de dimensión d, φ_n es una combinación lineal de los p vectores previos de las series de tiempo

$$\varphi_n = -\sum_{p}^{P} \boldsymbol{F}_p \boldsymbol{y}_{n-p} + \boldsymbol{e}_n,$$

donde cada $\boldsymbol{F}_p$ es una matriz $d \times d$ de coeficientes AR y $\boldsymbol{e}_n$ es un vector de ruido gaussiano (iid) con media cero y covarianza $\mathbf{\Sigma}$. Así, el modelo MAR tiene el modelo gaussiano como un caso especial (donde se tiene en cuenta únicamente el término del ruido). La densidad de observación está representada por

$$P\left(\varphi_n|\theta_n, \widetilde{\lambda}\right) = \frac{1}{\left|\tilde{\mathbf{\Sigma}}_{\theta_n}\right|^{-1/2} 2\pi^{d/2}} \exp\left[-\frac{1}{2}\boldsymbol{x}_n\left(\theta_n\right)^T \tilde{\mathbf{\Sigma}}_{\theta_n} \boldsymbol{x}_n\left(\theta_n\right)\right],$$

$$\boldsymbol{x}_n(i) = \left(\varphi_n - \tilde{\boldsymbol{\mu}}_i - \sum_p^P \tilde{\boldsymbol{F}}_{i,p} \varphi_{n-p} \right),$$

y $\tilde{\boldsymbol{F}}_{i,p}$ es la matriz $d \times d$ de los coeficientes AR para el estado i en el retraso p. Los coeficientes MAR pueden ser encontrados por la minimización del problema de los mínimos cuadrados ponderado.

4.4.2. Observaciones basadas en componentes independientes

En el análisis de componentes independientes (ICA - *independient component analysis*) la variable observada es modelada como:

$$\varphi_n = \boldsymbol{W}_i \boldsymbol{x}_n + \boldsymbol{\mu}_i + \boldsymbol{e}_n.$$

siendo los factores ocultos, $\boldsymbol{x}_n$ independientes. La variable i hace referencia al estado del HMM. Para el caso mas sencillo, se considera el sistema un modelo con $(\boldsymbol{\mu}_i) = \mathbf{0}$ y libre de ruido $(\boldsymbol{e}_n) = \mathbf{0}$.

$$\log P\left(\varphi_n | \theta_n, \widetilde{\lambda}\right) = -\log |\boldsymbol{W}_{\theta_n}| + \sum_k \log \left[p_k \left(\left(\boldsymbol{W}_{\theta_n}^{-1} \right)_{kr} \right) \varphi_{n,r} \right],$$

con $\boldsymbol{W}_{\theta_n}$ la matriz de factores de carga por estado θ_n del HMM, $(\cdot)_{kr}$ se refiere al $k, r-$ésimo elemento de la matriz y $p_k(\boldsymbol{x})$ es la evaluación de la densidad univariada de la componente de fuente independiente [31].

4.4.3. Análisis de componentes principales ocultas de Markov

El análisis de componentes principales ocultas de Markov HMPCA (*hidden Markov principal component analysis*) [2] tiene como objetivo incorporar la dependencia entre observaciones dentro del análisis de componentes principales (PCA). Empleando la formulación de componentes principales probabilísticas (PPCA) [42] como el modelo de observación en una modelo oculto de Markov (HMM) es posible establecer un modelo dinámico de componentes principales, cuyas ventajas son su capacidad de representación de series de datos multivariadas, con variables que estén correlacionadas en el tiempo, además de posibilitar la reducción localizada de características dinámicas [3].

Dado un modelo oculto de Markov con densidades de emisión continuas y representado por los parámetros $\lambda = \{\boldsymbol{\Pi}, \mathbf{B}, \mathbf{p}_{\theta_1}\}$; se asume que, la variable $\boldsymbol{\varphi}$ está definida por una transformación lineal de la variable latente $\boldsymbol{\zeta}$ de dimensión q, más ruido gaussiano de forma que

$$\varphi_n = \mathbf{C}\boldsymbol{\zeta}_n + \boldsymbol{\mu} + \boldsymbol{\epsilon},$$

para un modelo de mezcla, siendo $\mathbf{\Omega} = \{\Omega_m\} = \{\mathbf{C}_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m^2\}$ la distribución marginal se modela como

$$p(\boldsymbol{\varphi}|\mathbf{\Omega}) = \sum_{m=1}^{M} c_m p(\boldsymbol{\varphi}|\Omega_m),$$

los coeficientes de mezcla cumplen las restricciones estocásticas $\sum_m c_m = 1$ y $0 \leq c_m \leq 1 \ \forall m$. Usando este modelo de mezcla como la densidad de observación en el contexto de un HMM, se obtiene para el estado ϑ_j

$$p(\varphi_n|\theta_n = \vartheta_j, \mathbf{\Omega}) = \sum_{m=1}^{M} c_{jm} p(\varphi_n|\Omega_{jm}) = \sum_{m=1}^{M} c_{jm} p(\varphi_n|\boldsymbol{\mu}_{jm}, \mathbf{C}_{jm}, \sigma_{jm}^2), \tag{24}$$

con

$$p(\varphi_n|\boldsymbol{\mu}_{jm}, \mathbf{C}_{jm}, \sigma_{jm}^2) = \mathcal{N}\left(\boldsymbol{\mu}_{jm}, \sigma_{jm}^2 I + \mathbf{C}_{jm}\mathbf{C}_{jm}^T\right). \tag{25}$$

Los parámetros del modelo completo pueden estimarse fácilmente [2], empleando el algoritmo EM y están dados por

$$\mathbf{C}_{jm} = \left[\sum_{n}^{n_\varphi} \gamma_n(j,m)\left(\varphi_n - \boldsymbol{\mu}_{jm}\right)\langle\boldsymbol{\zeta}_{njm}\rangle^T\right]\left[\sum_{n}^{n_\varphi} \gamma_n(j,m)\langle\boldsymbol{\zeta}_{njm}\boldsymbol{\zeta}_{njm}^T\rangle\right]^{-1},$$

$$\sigma_{jm}^2 = \frac{\sum_{n}^{n_\varphi} \gamma_n(j,m)\left[\|\varphi_n - \boldsymbol{\mu}_{jm}\|^2 - 2\langle\boldsymbol{\zeta}_{njm}\rangle^T\mathbf{C}_{jm}^T\left(\varphi_n - \boldsymbol{\mu}_{jm}\right) + \mathrm{tr}\left(\mathbf{C}_{jm}^T\mathbf{C}_{jm}\langle\boldsymbol{\zeta}_{njm}{\boldsymbol{\zeta}_{njm}}^T\rangle\right)\right]}{d\sum_{n}^{n_\varphi}\gamma_n(j,m)},$$

$$\boldsymbol{\mu}_{jm} = \left[\sum_{n}^{n_\varphi}\gamma_n(j,m)\right]^{-1}\left[\sum_{n}^{n_\varphi}\gamma_n(j,m)\left[\varphi_n - \mathbf{C}_{jm}\langle\boldsymbol{\zeta}_{njm}\rangle\right]\right],$$

$$c_{jm} = \frac{\sum_{n}^{n_\varphi}\gamma_n(j,m)}{\sum_{n}^{n_\varphi}\sum_{k=1}^{M}\gamma_n(j,m)}.$$

la variable $\gamma_n(j,m)$ está dada por la ecuación (7). En el caso de las variables Π y $\mathbf{p}_{\theta_1}$ se calculan de la misma forma que en HMM convencional [17] y las esperanzas asociadas a la variable latente de PPCA con respecto a las observaciones se muestran a continuación

$$\langle\boldsymbol{\zeta}_{njm}\rangle = \mathbf{M}_{jm}^{-1}\mathbf{C}_{jm}^T\left(\varphi_n - \boldsymbol{\mu}_{jm}\right),$$

$$\langle \boldsymbol{\zeta}_{njm} \boldsymbol{\zeta}_{njm}^T \rangle = \sigma_{jm}^2 \mathbf{M}_{jm}^{-1} + \langle \boldsymbol{\zeta}_{njm} \rangle \langle \boldsymbol{\zeta}_{njm} \rangle^T.$$

Siendo $\langle \cdot \rangle$ la evaluación de la esperanza del argumento, con

$$\mathbf{M}_{jm} = \sigma_{jm}^2 \mathbf{I} + \mathbf{C}_{jm}^T \mathbf{C}_{jm}.$$

4.4.4. Modelos ocultos de Markov y análisis de factores

Por lo general cuando se utilizan HMMs con densidades de emisión continuas, para el tratamiento de señales con dimensionalidad muy grande se utilizan matrices de covarianza diagonales debido a que su costo computacional se hace menos pesado. Así, la variabilidad de las características que son modeladas por las mezclas de gaussianas con matrices de covarianza diagonales son únicamente de tipo discreto. Con el fin de apuntar a un modelo acertado se deben tener en cuenta tanto las variaciones de tipo discreto como las continuas; es aquí donde se introduce el análisis de factores que permite modelar la variabilidad continua sin tener problemas de sobreentrenamiento.

Este modelo busca combinar análisis de factores (FA - factor analysis) al interior de un HMM, aplicando un modelo FA por componente en cada estado [38]; obteniendo como resultado una forma de reducción de dimensionalidad lineal adaptada a las propiedades locales de los datos. En donde el análisis de factores modela la estructura de la matriz de covarianza de los datos de forma compacta con un número reducido de parámetros.

Se obtiene una reducción de espacio a partir de un modelo de variable latente, desde un espacio inicial donde se encuentra la variable $\boldsymbol{x}$ con media muestral $\boldsymbol{\mu}$ a partir de un mapeo a un espacio de menor dimensión que contiene a $\boldsymbol{y}$ (factores), con el fin de capturar la mayor variación de la variable de entrada.

Donde $p(\boldsymbol{y}) \sim \mathcal{N}(0, I)$, $\boldsymbol{\Lambda}$ es una matriz de pesos y $\boldsymbol{\Psi}$ representa una matriz diagonal. Se asume que $\boldsymbol{x}$ se puede obtener a partir de muestreo de la variable $\boldsymbol{y}$, realizando el cálculo $\boldsymbol{\mu} + \boldsymbol{\Lambda y}$ y luego adicionando ruido gaussiano independiente con media 0 y matriz de covarianza $\boldsymbol{\Psi}$. La relación entre $\boldsymbol{x}$ y $\boldsymbol{y}$ se puede observar en la probabilidad condicional dada a continuación

$$p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}_{\boldsymbol{x}} \left(\boldsymbol{\mu} + \boldsymbol{\Lambda y}, \boldsymbol{\Psi} \right).$$

El cálculo de la distribución marginal de $\boldsymbol{x}$ es directo dando como resultado

$$p(\boldsymbol{x}) = \mathcal{N}_{\boldsymbol{x}} \left(\boldsymbol{\mu}, \boldsymbol{\Psi} + \boldsymbol{\Lambda\Lambda}^T \right).$$

Ahora considere un HMM con densidades de emisión continuas modelados por gaussianas, en donde los vectores que contienen las características están condicionados por los estados ocultos. A partir de el análisis de factores se pueden obtener modelos de distribución de salida por cada estado de la forma:

$$p(\varphi_n|\theta_n = \vartheta_j, \mathbf{\Omega}) = \sum_{m=1}^{M} c_{jm} p(\varphi_n|\Omega_{jm}) = \sum_{m=1}^{M} c_{jm} p(\varphi_n|\boldsymbol{\mu}_{jm}, \mathbf{\Lambda}_{jm}, \mathbf{\Psi}_{jm}),$$

donde, m representa cada componente de mezcla por estado y j hace referencia a cada estado del modelo HMM.

$$p(\varphi_n|\boldsymbol{\mu}_{jm}, \mathbf{\Lambda}_{jm}, \mathbf{\Psi}_{jm}) = \mathcal{N}_{\varphi_n}\left(\boldsymbol{\mu}_{jm}, \mathbf{\Psi}_{jm} + \mathbf{\Lambda}_{jm}\mathbf{\Lambda}_{jm}^T\right).$$

Para este modelo se pueden utilizar los mismos cómputos estadísticos utilizados para HMM que se encuentran en la literatura. La ventaja principal es que se requieren mucho menos operaciones que las requeridas para matrices de covarianza completas sin pérdida de generalidad.

Cabe aclarar, que como el análisis de factores solo afecta las densidades de emisión; la actualización de los demás parámetros que conforman un modelo HMM no se ve afectada.

5. Marco experimental

El reconocimiento de patrones se puede dividir en dos amplias tareas que son: el análisis de características y la clasificación de patrones. Para evaluar el rendimiento del sistema completo se evalúa el rendimiento de cada tarea, generalmente por su exactitud; por tanto, es de esperarse que la implementación de cada tarea esté enfocada hacia la minimización del error del proceso. Además, para evitar inconsistencias en el sistema ambas etapas deben estar relacionadas, es decir, deben utilizarse criterios de estimación con metodologías similares.

5.1. HMPCA con enfoque discriminativo

El método descrito en la sección 4.4.3, puede verse como una versión temporal del análisis de componentes principales probabilístico; en donde, cada estado está representado por una densidad de emisión $b_j(\varphi_n)$ asociada a un modelo PPCA como se muestra en la ecuación (24). La matriz de covarianza $\boldsymbol{R}_{jm}$ está relacionada con los parámetros de un modelo de componentes principales como se muestra en la ecuación (25) y es igual a

$$\boldsymbol{R}_{jm} = \left(\sigma_{jm}^2 \boldsymbol{I} + \mathbf{C}_{jm}\mathbf{C}_{jm}^T\right).$$

La actualización del modelo completo utilizando el método discriminativo de estimación MCE es en forma general similar a la mostrada en 4.2.2. La variación se centra únicamente en los parámetros que conforman las emisiones, que para el caso del modelo completo planteado y basados en la ecuación (11) se muestran a continuación. En el caso de evaluar las actualizaciones referentes a la clase ν, el vector de medias asociado al estado j y la componente de mezcla m se tiene que $\frac{\partial \ell_\nu(\boldsymbol{\varphi};\lambda)}{\partial \tilde{\boldsymbol{\mu}}_{jm}^{(\nu)}}$ es igual a

$$\frac{\partial \ell_\nu(d_\nu)}{\partial d_\nu}\frac{\partial d_\nu}{\partial \tilde{\boldsymbol{\mu}}_{jm}^{(\nu)}}.$$

El primer término se encuentra evaluado en la ecuación (14) y el segundo es igual a

$$\frac{\partial d_\nu}{\partial \tilde{\boldsymbol{\mu}}_{jm}^{(\nu)}} = \frac{\partial g_\nu(\boldsymbol{\varphi};\lambda)}{\partial \tilde{\boldsymbol{\mu}}_{jm}^{(\nu)}} = -\sum_{n=1}^{n_\varphi} \delta\left(\bar{\theta}_n, \vartheta_j\right) \frac{\partial \log\left(b_j^{(\nu)}(\varphi_n)\right)}{\partial \tilde{\boldsymbol{\mu}}_{jm}^{(\nu)}}.$$

siendo donde $\bar{\theta}_n$ el n-ésimo elemento de la secuencia de estados óptima $\bar{\boldsymbol{\theta}}$ como se muestra en la sección 4.2.2. La derivada del logaritmo de la densidad de emisión para el estado j con respecto a la media es

$$\frac{\partial \log\left(b_j^{(\nu)}(\varphi_n)\right)}{\partial \tilde{\boldsymbol{\mu}}_{jm}^{(\nu)}} = \frac{1}{b_j^{(\nu)}(\varphi_n)} c_{jm}^{(\nu)} \mathcal{N}_{\varphi_n}\left(\boldsymbol{\mu}_{jm}^{(\nu)}, \boldsymbol{R}_{jm}^{(\nu)}\right)\left[\left(\boldsymbol{R}_{jm}^{(\nu)}\right)^{-1}\left(\varphi_n - \boldsymbol{\mu}_{jm}^{(\nu)}\right)\right]. \tag{26}$$

Realizando un tratamiento similar, para el caso de la matriz de transformación para el estado j y la componente de mezcla m se tiene

$$\frac{\partial \log\left(b_j^{(\nu)}(\varphi_n)\right)}{\partial \tilde{\boldsymbol{C}}_{jm}^{(\nu)}} = \frac{c_{jm}^{(\nu)} \mathcal{N}_{\varphi_n}\left(\boldsymbol{\mu}_{jm}^{(\nu)}, \boldsymbol{R}_{jm}^{(\nu)}\right)}{b_j^{(\nu)}(\varphi_n)}\Bigg[\left(\boldsymbol{R}_{jm}^{(\nu)}\right)^{-1}\left(\varphi_n - \boldsymbol{\mu}_{jm}^{(\nu)}\right)\left(\varphi_n - \boldsymbol{\mu}_{jm}^{(\nu)}\right)^T - \boldsymbol{I}\Bigg]\left(\boldsymbol{R}_{jm}^{(\nu)}\right)^{-1}\boldsymbol{C}_{jm}^{(\nu)}. \tag{27}$$

Para finalizar la derivada asociada al término de la varianza está dada por

$$\frac{\partial \log\left(b_j^{(\nu)}(\varphi_n)\right)}{\partial \tilde{\sigma^2}_{jm}^{(\nu)}} = \frac{c_{jm}^{(\nu)} \mathcal{N}_{\varphi_n}\left(\boldsymbol{\mu}_{jm}^{(\nu)}, \boldsymbol{R}_{jm}^{(\nu)}\right)}{b_j^{(\nu)}(\varphi_n)}\Bigg[-\frac{1}{2} tr\left(\boldsymbol{R}_{jm}^{(\nu)}\right)^{-1} + \frac{1}{2}\left(\varphi_n - \boldsymbol{\mu}_{jm}^{(\nu)}\right)^T\left(\boldsymbol{R}_{jm}^{(\nu)}\right)^{-2}\left(\varphi_n - \boldsymbol{\mu}_{jm}^{(\nu)}\right)\Bigg]. \tag{28}$$

La actualización de las componentes de mezcla por estado es igual a la obtenida en la sección 4.2.2 y es igual a

$$\frac{\partial \log\left(b_j^{(\nu)}(\varphi_n)\right)}{\partial \tilde{c}_{jm}^{(\nu)}} = \frac{1}{b_j^{(\nu)}(\varphi_n)} \mathcal{N}_{\varphi_n}\left(\boldsymbol{\mu}_{jm}^{(\nu)}, \boldsymbol{R}_{jm}^{(\nu)}\right) c_{jm}^{(\nu)}\left(1 - c_{jm}^{(\nu)}\right). \tag{29}$$

Se debe recordar que, el gradiente descendente no garantiza las propiedades estocásticas de los parámetros del modelo; por tanto, se debe los parámetros deben ser transformados como se muestra:

$$\begin{aligned}
\boldsymbol{\mu}_{jm} \rightarrow \tilde{\boldsymbol{\mu}}_{jm} \quad & \text{donde,} \quad \tilde{\boldsymbol{\mu}}_{jm} = \frac{\boldsymbol{\mu}_{jm}}{\sigma_{jm}^2}, \\
\sigma_{jm}^2 \rightarrow \tilde{\sigma}_{jm}^2 \quad & \text{donde,} \quad \tilde{\sigma}_{jm}^2 = \log \sigma_{jm}^2, \\
\boldsymbol{C}_{jm} \rightarrow \tilde{\boldsymbol{C}}_{jm} \quad & \text{donde,} \quad \tilde{\boldsymbol{C}}_{jm} = \left(\sigma_{jm}^2 \boldsymbol{I}\right)^{-1} \boldsymbol{C}_{jm},
\end{aligned}$$

5.2. Metodología propuesta para la reducción de dimensionalidad dinámica

Esta metodología consta de dos etapas: la primera asociada a la reducción dinámica de características (HMPCA), conformada por un modelo de variable latente que utiliza un HMM con emisiones continuas para representar la variabilidad de los datos observados, asociado con el método de reducción lineal de características conocido como PPCA. Se busca mejorar el rendimiento de este sistema, empleando técnicas de entrenamiento discriminativo para la estimación de los parámetros del modelo completo. La segunda que es la clasificación, utiliza un HMM entrenado con el mismo criterio empleado para la etapa de reducción localizada. Su operación se explica a continuación en el algoritmo 1.

Algoritmo 1 Esquema propuesto para el reconocimiento de patrones compuesto por una etapa de reducción dinámica basada en el modelo HMPCA y por una etapa de clasificación basado en un HMM

Requiere: $\boldsymbol{\varphi}$, n_ϑ, M, η y ϵ
1: Modelo inicial
2: **hacer**
3: Ajuste de parámetros modelo reducción
Reestimar: (13), (15), (26), (27), (28) y (29)
4: **hasta** Convergencia
5: **retornar** λ % parámetros del modelo de reducción entrenado
% Obtener el espacio reducido
6: path=VITERBI($\boldsymbol{\varphi}$,λ) % selección de la mejor secuencia de estados
% Reducción dinámica localizada
7: **para** $n = 1$ to tamaño($\boldsymbol{\varphi}$) **hacer**
8: (j,m)=evalpath(n)
9: ζ_n =trans($\varphi_n, \Omega_{jm}, c_{jm}$), (23)
10: **fin para**
11: **retornar** $\boldsymbol{\zeta}$
Requiere: $\boldsymbol{\zeta}$, n_ϑ, M, η y ϵ
12: **hacer**
13: Ajuste de parámetros del clasificador basado en HMM
Reestimar: (13), (15), (16b), (16a) y (16c)
14: **hasta** Convergencia
15: **retornar** λ_c % parámetros del modelo de clasificación entrenado

En la primera etapa que se ha llamado modelo inicial, se genera un HMM con densidades de emisión Gaussianas como se muestra: se realiza en forma aleatoria la inicialización para cada clase de la matriz de transición, el vector de probabilidad inicial y los pesos de las componentes de mezcla asociadas a las densidades de estado; garantizando las propiedades estocásticas que restringen estos parámetros. Luego, como las

emisiones de estado están explicadas por modelos PPCA, se recurre al algoritmo de K-medias para inicializar los parámetros de cada modelo PPCA por estado y componente de mezcla.
En la etapa de estimación de parámetros, se toma el modelo completo HMPCA, conformado por el conjunto de modelos generados en la etapa anterior; utilizando para las densidades de emisión por estado, las ecuaciones del numeral 5.1 y para los demás variables que conforman el modelo las actualizaciones obtenidas en la sección 4.2.2.
Para poder realizar la reducción dinámica, se requiere encontrar cuál es la secuencia de estados que mejor representa a cada secuencia de observación, en esta etapa se utiliza el algoritmo de Viterbi. Obtenida la secuencia de estados óptima se realiza la reducción a cada observación en el tiempo utilizando el modelo correspondiente PPCA por estado y componente de mezcla. Al nuevo espacio, que ya ha sido transformado con el modelo de reducción dinámico basado en PPCA, se aplica un algoritmo de clasificación de series de tiempo, que para el caso de este estudio será de nuevo un modelo oculto de Markov con funciones de densidad de probabilidad por estado Gaussianas con estimación de parámetros basado en el algoritmo MCE, la estimación de parámetros para este modelo se muestra en el numeral 4.2.2.

5.3. Esquema propuesto para clasificación

Con este esquema se busca modelar de forma implícita y explícita la correlación que existe entre las características que definen las señales. Implícitamente, utilizando 2 o más componentes de mezcla por estado; otra forma es aprovechando el modelo PPCA asociado a cada estado y componente de mezcla que permite obtener matrices de covarianza compactas de los datos en donde los elementos fuera de la diagonal representan la correlación explícita entre las características.

El análisis PPCA utiliza un número pequeño de parámetros para modelar la estructura de la covarianza de los datos sin importar si presentan una alta dimensión. Para el esquema propuesto, el ajuste de los parámetros que conforman el modelo se realiza: minimizando el número de errores de clasificación ó maximizando la verosimilitud de los datos observados. A continuación se muestra el algoritmo 2 que explica el esquema propuesto para la clasificación.

El algoritmo propuesto para la clasificación consta de las siguientes etapas: la inicialización y la estimación de parámetros que funcionan de la misma forma como se explica en la sección 5.2. En la etapa de decisión, se evalúa la función de pérdida mostrada en la ecuación (10) que definirá a que clase pertenece cada muestra; utilizando la medida de distancia definida en la ecuación (8) como se muestra en el numeral 4.2.2.

Algoritmo 2 Clasificación dinámica de señales multivariadas utilizando el modelo HMPCA

Requiere: $\boldsymbol{\varphi}$, n_ϑ, M, η y ϵ

1: Modelo inicial
2: **hacer**
3: Ajuste de parámetros modelo HMPCA
Reestimar: (13), (15), (26), (27), (28) y (29)
4: **hasta** Convergencia
5: **retornar** λ % parámetros del modelo HMPCA entrenado

6. Resultados

El conjunto de pruebas está enfocado a realizar tareas de análisis de señales biológicas (ECG y voz), específicamente en la clasificación de señales con disfunciones, y está definido por tres esquemas diferentes como se muestra a continuación. El primero, está conformado por una etapa de reducción de dimensionalidad seguida de una etapa de clasificación basada en clasificadores Bayesianos; aquí, se busca comparar el desempeño del modelo HMPCA propuesto para reducción dinámica con la técnica clásica PCA utilizando clasificadores lineales y cuadráticos. En el segundo, el modelo HMPCA se utiliza de nuevo para la tarea de reducción trabajando la clasificación con un HMM. Se realiza la comparación de este esquema, cuando se trabaja como modelo generativo 4.4.3 y cuando se trabaja como función discriminante 5.2. Por último, aún cuando la aplicación del HMPCA es la reducción dinámica localizada, es posible utilizar el modelo como una función de clasificación directa, debido a que está conformado por un HMM con emisiones por estado definidas por modelos PPCA y por tanto se puede evaluar una señal de entrada en el conjunto de modelos que definen las clases para luego poder etiquetarla en una salida. En este esquema de clasificación, que es el tercero, de nuevo se realiza la comparación de resultados obtenidos mediante optimización basada en ML y en MCE 5.3.

Es importante mencionar que, con el fin de acelerar la convergencia en las optimización del modelo HMPCA basado en MCE, se realiza una actualización de la tasa de aprendizaje ϵ, utilizada en el ajuste de los parámetros con la técnica de gradiente descendente (11). Esta adaptación está definida como $\epsilon(t+1) = \epsilon_0 \Xi^{-t/T}$, que depende del número de muestras [43], con Ξ una constante que para todos los esquemas definidos en esta sección es igual a 100; para cada prueba se especificará el valor inicial de la tasa de aprendizaje ϵ_0.

6.1. Patologías en ECG

6.1.1. Base de datos y extracción de características

Para la clasificación de señales Electrocardiográficas (ECG) se utilizó la base de datos ST-T Europea, conformada por eventos isquémicos y eventos normales. Del total de

muestras se tomó un conjunto de 96 señales, 48 representaciones por clase. La caracterización se realizó utilizando la transformada wavelet, eligiendo los coeficientes de mayor energía por 13 niveles de descomposición [27]. Las wavelets utilizadas corresponden a la Daubechies4 (Db4) y Symlet2 (Sym2).

6.1.2. Clasificación

Para todas las pruebas realizadas, la evaluación del rendimiento del sistema se calcula utilizando validación cruzada con 4 particiones. En el caso del modelo HMPCA con enfoque discriminativo 5.1, los parámetros que definen la función de suavización mostrada en la ecuación (10) son iguales para todas las pruebas; sus valores son: $\eta = 2$, $\rho = 1$ y $\omega = 0$. Para la caracterización Sym2, el valor de ϵ_0 es $1e-2$ y para Db $5e-4$.

Los HMM utilizados en clasificación son ergódicos y las matrices de covarianza que definen sus modelos de emisión de estado son diagonales, de no especificarse el número de componentes de mezcla M para los modelos HMM y HMPCA se asume igual a uno.

Clasificación Bayesiana. Para este esquema de clasificación, inicialmente se transforma el espacio de entrada a un espacio de menor dimensión (q), utilizando tres esquemas de reducción: PCA, y el esquema de reducción dinámico HMPCA como función discriminante (MCE) y como modelo generativo (ML); luego, se evalúa cada comportamiento con dos modelos de clasificación Bayesiana, un clasificador discriminante lineal (LDC) y uno cuadrático (QDC). Se realizaron pruebas con 3, 5, y 7 estados (n_ϑ) por modelo. A continuación se muestran los mejores resultados:

TABLA 6.1. Precisión para la base de datos Sym2 utilizando clasificadores Bayesianos con $q = 9$

Esquema de clasificación	n_ϑ	Precisión %
PCA + LDC	–	$56,25 \pm 2,41$
PCA + QDC	–	$87,50 \pm 12,27$
HMPCA (ML) + LDC	7	$90,63 \pm 10,96$
HMPCA (MCE) + LDC	3	$91,76 \pm 7,61$
HMPCA (ML) + QDC	5	$94,79 \pm 6,25$
HMPCA (MCE) + QDC	7	$95,83 \pm 4,81$

En la tabla 6.1 se observa que el esquema de reducción basado en PCA y clasificado con LDC, no es óptimo para trabajar con la base de datos Sym2, dado que se asemeja a una clasificación aleatoria. Al probar con LDC el esquema de reducción localizada propuesto, es claro que los errores de clasificación se reducen para ML en más de un

30 % y para MCE en casi un 40 %. En el caso del clasificador QDC, es de esperarse que los resultados sean superiores a los obtenidos con LDC; y los errores se mejoran en un 7 % para ML y un 8 % para MCE.

TABLA 6.2. Precisión para la base de datos DB4 utilizando clasificadores Bayesianos con $q = 11$

Esquema de clasificación	n_ϑ	Precisión %
PCA + LDC	–	$52,08 \pm 5,38$
PCA + QDC	–	$82,29 \pm 12,44$
HMPCA (ML) + LDC	7	$92,71 \pm 5,24$
HMPCA (MCE) + LDC	3	$94,79 \pm 2,08$
HMPCA (ML) + QDC	5	$90,63 \pm 7,12$
HMPCA (MCE) + QDC	7	$97,92 \pm 4,17$

En la tabla 6.2 de forma similar que en la tabla 6.1, sucede que los porcentajes de error utilizando LDC disminuyen en más de 40 % para ambos casos de reducción dinámica; y en el caso de QDC, se mejora en un 8 % para ML y en un 15 % para MCE. Se puede afirmar para ambas caracterizaciones de las señales ECG, que los mejores porcentajes de acierto los presenta la reducción dinámica optimizada con el criterio del MCE, adicionalmente sus porcentajes de acierto son los que muestran las menores variaciones.

HMPCA aplicado a reducción. En este caso, se evalúa el sistema variando el número de dimensiones del espacio reducido q y el número de estados por modelo n_ϑ.

TABLA 6.3. Precisión utilizando la base de datos Db4 con el espacio reducido

		Precisión %		
HMPCA	q	$n_\vartheta = 3$	$n_\vartheta = 5$	$n_\vartheta = 7$
MCE	3	$85,42 \pm 15,77$	$87,50 \pm 9,00$	$85,42 \pm 7,22$
ML		$93,75 \pm 7,22$	$93,75 \pm 4,17$	$94,79 \pm 6,25$
MCE	5	$93,75 \pm 5,38$	$93,75 \pm 5,38$	$87,50 \pm 19,54$
ML		$97,92 \pm 2,41$	$98,96 \pm 2,08$	$96,88 \pm 2,08$
MCE	7	$96,88 \pm 2,08$	$95,83 \pm 4,81$	$97,92 \pm 2,41$
ML		$95,83 \pm 3,40$	$89,58 \pm 11,02$	$98,96 \pm 2,08$
MCE	9	$95,83 \pm 5,89$	$96,88 \pm 2,08$	$97,92 \pm 2,41$
ML		$96,88 \pm 3,99$	$98,96 \pm 2,08$	$99,99 \pm 0,10$
MCE	11	$97,92 \pm 4,17$	$99,99 \pm 0,10$	$99,99 \pm 0,10$
ML		$98,96 \pm 2,08$	$97,92 \pm 4,17$	$98,96 \pm 2,08$

La tabla 6.3 muestra los porcentajes de clasificación de la base de datos caracterizada usando Db4 con el espacio reducido. El desempeño del modelo, para ambos casos,

alcanza altos porcentajes y una estabilidad aceptable (desviación estándar). A medida que se aumenta el número de componentes, es decir, el espacio transformado se hace más grande, se obtienen mejores porcentajes de acierto, obteniendo con 11 componentes los más altos porcentajes de acierto.

TABLA 6.4. Precisión usando la base de datos Sym2 con el espacio reducido

		Precisión %		
HMPCA	q	$n_\vartheta = 3$	$n_\vartheta = 5$	$n_\vartheta = 7$
MCE	3	$95,83 \pm 3,40$	$90,63 \pm 3,99$	$88,54 \pm 8,59$
ML		$84,38 \pm 5,24$	$85,42 \pm 12,95$	$66,67 \pm 20,69$
MCE	5	$93,75 \pm 2,41$	$96,88 \pm 2,08$	$93,75 \pm 2,41$
ML		$62,50 \pm 12,27$	$73,96 \pm 25,32$	$92,71 \pm 7,12$
MCE	7	$96,88 \pm 3,99$	$96,88 \pm 3,99$	$96,88 \pm 3,99$
ML		$72,92 \pm 17,18$	$90,63 \pm 13,77$	$93,75 \pm 5,38$
MCE	9	$96,88 \pm 6,25$	$98,96 \pm 2,08$	$96,88 \pm 3,99$
ML		$96,88 \pm 6,25$	$97,92 \pm 2,41$	$93,75 \pm 12,50$

Resultados similares se obtienen para la base de datos caracterizada con Sym2, como se observa de la tabla 6.4. Sin embargo, para el esquema ML, los porcentajes de acierto usando la serie caracterizada con Db4, son en promedio mayores que los obtenidos con Sym2, como lo muestra una observación detallada de las tablas 6.4 y 6.3. Para este caso, el mayor porcentaje de acierto se logra con la optimización MCE para $q = 9$.

HMPCA como clasificador. La tabla 6.5 muestra porcentajes de acierto para la base de datos Db4 utilizando como modelo de clasificación HMPCA. Para cada valor de dimensión de reducción q, se muestran las características del modelo que presentó mejores resultados. En forma general, para el conjunto de pruebas realizadas, el rendimiento del modelo basado en MCE fue superior que para ML; el mejor porcentaje de acierto se obtiene para una dimensión igual a 7.

Al analizar los resultados con la base de datos Sym2, mostrado en 6.6, se puede observar que el rendimiento más alto se consigue con el criterio ML, Siendo el porcentaje más alto obtenido con una dimensión reducida igual a 7.

TABLA 6.5. Porcentajes de clasificación de la base de datos Db4 utilizando HMPCA

HMPCA	q	n_ϑ	M	Precisión %
MCE	3	7	2	$94,79 \pm 2,08$
ML			3	$93,75 \pm 2,41$
MCE	5	5	3	$94,79 \pm 6,25$
ML		7		$93,75 \pm 2,41$
MCE	7	7	3	$95,83 \pm 5,89$
ML		3	3	$94,79 \pm 3,99$
MCE	9	5	3	$92,71 \pm 6,25$
ML		7	1	$91,67 \pm 5,89$

TABLA 6.6. Porcentajes de clasificación de la base de datos Sym2 utilizando HMPCA

HMPCA	q	n_ϑ	M	Precisión %
MCE	3	7	2	$94,79 \pm 5,24$
ML			3	$94,79 \pm 5,24$
MCE	5	7	2	$95,83 \pm 3,40$
ML		5	3	$95,83 \pm 5,89$
MCE	7	5	2	$94,79 \pm 2,08$
ML			1	$96,88 \pm 2,08$
MCE	9	3	2	$89,58 \pm 2,41$
ML				$95,83 \pm 3,40$

6.2. Patologías de voz

6.2.1. Base de datos y Extracción de características

Las pruebas para este tipo de señales se realizan sobre dos bases de datos, sus principales características se enuncian a continuación. La base de datos nombrada en este documento como DB1, pertenece a la Universidad Nacional de Colombia sede Manizales, y contiene 80 muestras de la vocal sostenida $/a/$, pronunciadas por 40 pacientes con voz normal y 40 pacientes con voz disfónica. La frecuencia de muestreo de esta base de datos es de 22 kHz. Las señales de voz se segmentan usando ventanas de 30 milisegundos (ms) con un traslape de 10 ms. Por cada ventana, se extraen 12 coeficientes MFCC [16] y el coeficiente de energía. Se incluyen igualmente las derivadas de primer y segundo orden [35] para obtener un vector final de 39 variables por ventana.

La segunda, DB2, fue desarrollada por el Massachusetts Eye and Ear Infirmary, y corresponde a pronunciaciones de la vocal sostenida /ah/. Se utilizaron 173 registros de pacientes patológicos (patologías de tipo: vocal, orgánico, neurológico traumático

y psíquico) y 53 registros de pacientes normales. Los registros utilizados fueron remuestreados a 25 kHz con una resolución de 16 bits. Para su caracterización se emplean los coeficientes MFCC derivados del cálculo de la FFT; adicionalmente, la relación armónico ruido (Harmonic to Noise Ratio - HNR), la energía de ruido normalizada (Normalized Noise Energy - NNE), el parámetro de energía medida por trama de la señal y la relación excitación glottal ruido (Glottal to Noise Excitation Ratio), incluyendo sus primeras y segundas derivadas, debido a que estas medidas dan una idea de la calidad y grado de normalidad de la voz.

6.2.2. Clasificación

Para la clasificación de las señales de voz se utiliza validación cruzada con 5 particiones con el fin de evaluar el rendimiento del sistema. Con respecto a los parámetros que definen la función de suavización mostrada en la ecuación (10) se tomaron como valores para realizar las pruebas: $\eta = 2$, $\rho = 1$, $\omega = 0$. En el caso de DB1, $\epsilon_0 = 1e - 2$ y para DB2, es igual a $0,1$. Los HMM utilizados en clasificación son ergódicos y las matrices de covarianza que definen sus modelos de emisión de estado son diagonales, de no especificarse el número de componentes de mezcla M para los modelos HMM y HMPCA se asume igual a uno.

Clasificación Bayesiana. Se evalúa el desempeño de clasificación de la serie transformada, variando el número de estado y el tamaño de la dimensión final. La reducción se realiza con PCA, HMPCA basado en MCE y HMPCA basado en ML. Los mejores resultados se muestran en las tablas 6.7 y 6.8.

TABLA 6.7. Precisión para la base de datos DB1 utilizando clasificadores Bayesianos

Esquema de clasificación	q	n_ϑ	Precisión %
PCA + LDC	33	–	$52,50 \pm 3,42$
PCA + QDC	20	–	$53,75 \pm 8,39$
HMPCA (ML) + LDC	25	3	$62,50 \pm 7,65$
HMPCA (MCE) + LDC	33	7	$81,25 \pm 8,84$
HMPCA (ML) + QDC	20	3	$87,50 \pm 6,25$
HMPCA (MCE) + QDC	25	3	$91,25 \pm 7,13$

Los porcentajes de acierto obtenidos, para la base de datos DB1 al utilizar un esquema de reducción de dimensionalidad dinámico y localizado, permiten reducir los errores de clasificación hasta en un 38 %. Aquí de nuevo se observa que, la reducción

con la técnica PCA no ofrece resultados acertados. El mayor porcentaje de acierto se obtiene para una dimensión reducida igual a 25, que corresponde a trabajar con el 75 % del número de dimensiones que conforman el espacio original.

TABLA 6.8. Precisión para la base de datos DB2 utilizando clasificadores Bayesianos

Esquema de clasificación	q	n_ϑ	Precisión %
PCA + LDC	46	–	$76,62 \pm 1,46$
PCA + QDC	30	–	$93,69 \pm 6,35$
HMPCA (ML) + LDC	46	7	$80,65 \pm 3,45$
HMPCA (MCE) + LDC	46	3	$89,71 \pm 6,27$
HMPCA (ML) + QDC	20	3	$97,84 \pm 1,42$
HMPCA (MCE) + QDC	20	3	$97,49 \pm 3,27$

Para el caso de la base de datos DB2, tabla 6.8 el porcentaje de acierto más alto se obtiene para $q = 20$, esto indica una reducción de más del 50 % del espacio característico con una reducción en el porcentaje de error el orden de 4 %; si bien, el error de clasificación no disminuye en porcentajes tan altos comparados con los resultados de DB1, lograr un espacio reducido compacto beneficia de forma directa la etapa de clasificación.

HMPCA aplicado a reducción. Se evalúa el desempeño del modelo, realizando la transformación y reducción mediante HMPCA, luego se realiza la clasificación utilizando un HMM. Se usan diferentes valores de dimensión de espacio reducido q. Los resultados obtenidos se muestran en la tabla 6.9.

TABLA 6.9. Precisión usando la base de datos DB1 con el espacio reducido

		Precisión %		
HMPCA	q	$n_\vartheta = 3$	$n_\vartheta = 5$	$n_\vartheta = 7$
MCE	15	$67,50 \pm 19,47$	$78,75 \pm 9,48$	$66,25 \pm 11,35$
ML		$70,00 \pm 10,27$	$73,75 \pm 14,92$	$70,00 \pm 18,43$
MCE	20	$73,75 \pm 17,90$	$83,75 \pm 13,69$	$83,75 \pm 16,89$
ML		$71,25 \pm 14,39$	$63,75 \pm 2,80$	$76,25 \pm 8,15$
MCE	25	$82,50 \pm 6,85$	$83,75 \pm 14,39$	$78,75 \pm 11,35$
ML		$68,75 \pm 10,83$	$72,50 \pm 13,69$	$76,25 \pm 14,92$
MCE	30	$78,75 \pm 10,46$	$81,25 \pm 15,93$	$73,75 \pm 11,18$
ML		$67,50 \pm 11,18$	$48,75 \pm 16,18$	$68,75 \pm 15,31$
MCE	33	$85,00 \pm 10,46$	$81,25 \pm 22,96$	$86,25 \pm 5,23$
ML		$75,00 \pm 17,12$	$63,75 \pm 16,77$	$57,50 \pm 11,18$

TABLA 6.10. Precisión usando la base de datos DB2 con el espacio reducido

		Precisión %		
HMPCA	q	$n_\vartheta = 3$	$n_\vartheta = 5$	$n_\vartheta = 7$
MCE	10	$75,60 \pm 5,95$	$81,00 \pm 5,68$	$75,71 \pm 1,48$
ML		$59,00 \pm 16,55$	$75,09 \pm 6,71$	$69,00 \pm 7,79$
MCE	15	$81,64 \pm 6,72$	$80,71 \pm 5,04$	$76,16 \pm 5,32$
ML		$60,87 \pm 12,77$	$64,11 \pm 5,19$	$74,11 \pm 5,20$
MCE	20	$77,71 \pm 4,34$	$77,93 \pm 2,76$	$80,15 \pm 2,81$
ML		$66,84 \pm 14,95$	$67,36 \pm 8,30$	$70,42 \pm 7,20$
MCE	25	$79,29 \pm 2,09$	$84,64 \pm 7,00$	$80,55 \pm 1,18$
ML		$66,51 \pm 14,10$	$68,95 \pm 12,25$	$69,38 \pm 9,73$
MCE	30	$88,56 \pm 4,59$	$82,09 \pm 4,53$	$81,51 \pm 4,51$
ML		$57,56 \pm 11,68$	$69,38 \pm 9,86$	$66,35 \pm 15,51$

Para la base de datos DB1, que muestra sus resultados en la tabla 6.9, los porcentajes de acierto no son tan altos como los obtenidos para el esquema anterior;cabe resaltar que, los porcentajes de aciero son mayores para la optimización basada en MCE.

Observando los resultados obtenidos con la base de datos BD2 mostrados en 6.10, se nota que los porcentajes de acierto son más altos para el criterio MCE. Para ambos casos, ML y MCE, se necesita un espacio transformado de dimensión muy cercana al espacio original; y aun así no se obtienen porcentajes de acierto mayores a 90 %.

HMPCA aplicado a clasificación. Se evalúa el desempeño de clasificación de la serie multivariada transformada y reducida mediante el análisis PPCA a través del tiempo usando diferentes valores de q. A continuación se muestran los mejores resultados.

TABLA 6.11. Porcentajes de clasificación de la base de datos DB1 utilizando HMPCA

HMPCA	q	n_ϑ	M	Precisión %
MCE	5	3	2	$92,50 \pm 8,15$
ML		7	1	$93,75 \pm 0,00$
MCE	15	3	3	$88,75 \pm 10,27$
ML		5	2	$91,25 \pm 8,39$
MCE	25	7	2	$87,50 \pm 4,42$
ML		5	2	$96,25 \pm 5,59$
MCE	33	7	1	$83,75 \pm 7,13$
ML		5		$90,00 \pm 11,35$

Para DB1 y DB2, se muestra una ventaja significativa del criterio ML sobre MCE; además del espacio completo de pruebas de donde se extraen los resultados mostrados en las tablas 6.11 y 6.12, se nota claramente que a medida que se incrementan dimensiones al espacio reducido se reducen los porcentajes de acierto. Este fenómeno también sucede a medida que se aumenta el número de componentes de mezcla M por estado, y el número de estados por modelo.

TABLA 6.12. Porcentajes de clasificación de la base de datos DB2 utilizando HMPCA

HMPCA	q	n_ϑ	M	Precisión %
MCE	10	3	1	$93,80 \pm 1,89$
ML				$94,25 \pm 3,41$
MCE	15	3	1	$92,95 \pm 0,53$
ML				$95,22 \pm 2,61$
MCE	20	3	1	$90,27 \pm 2,48$
ML				$92,55 \pm 3,52$
MCE	25	3	1	$88,56 \pm 3,28$
ML				$92,95 \pm 0,53$
MCE	30	3	1	$89,20 \pm 4,57$
ML				$93,24 \pm 3,68$

7. Conclusiones

En este documento se proponen dos metodologías para la clasificación de señales variantes en el tiempo. La primera, que está conformada por un modelo de reducción dinámica localizada y posteriormente asociada a un HMM que tiene como finalidad la clasificación de las series reducidas. Esta metodología tiene en cuenta la variabilidad de las señales utilizando un modelo de reducción estático inmerso en un proceso doblemente estocástico.

La segunda, que aplicada el modelo HMPCA directamente como clasificador, permite modelar las correlaciones existentes entre las características que definen las señales con un número de variables reducido: de forma implícita, cuando utiliza múltiples Gaussianas por estado y de forma explícita, aprovechando los elementos fuera de la diagonal obtenidos de las matrices de covarianza compactas asociadas a cada modelo PPCA.

Luego de aplicar las metodologías propuestas se pudo observar que:

- El esquema de reducción demostró ser efectivo, al aplicarse sobre los conjuntos de datos de voz y ECG; dado que al evaluar su rendimiento utilizando un clasificador básico se obtuvieron porcentajes de acierto elevados, comparados con los obtenidos con el modelo clásico PCA. Los errores de clasificación se reducen en más del 30 %; y para el caso de DB2 que no presenta una mejora tan alta, el porcentaje de acierto más elevado se obtiene con una dimensión de reducción menor que la mitad de la dimensión del espacio original. En general, los resultados muestran mejor desempeño del modelo de reducción dinámica cuando se emplea como técnica de optimización el criterio MCE.
- La metodología propuesta para reducción dinámica de características, muestra que los porcentajes de acierto obtenidos para HMPCA basado en MCE son mayores en los 4 casos propuestos. Para las señales ECG, se obtienen porcentajes de acierto elevados, que aumentan a medida que se aumenta el dominio del espacio transformado. En el caso de las señales de voz, el rendimiento de este esquema no es tan bueno comparado con los demás esquemas del conjunto de pruebas.

– Con respecto al esquema de clasificación para las señales ECG, aun cuando sus resultados no son tan elevados como los obtenidos para el esquema de reducción; es importante tener en cuenta que, su modelo es computacionalmente menos costoso que el propuesto para la reducción dinámica de características, pues este último debe entrenar además del modelo de HMPCA, un modelo HMM que es el utilizado en la etapa de clasificación. Para las bases de datos de voz, en cambio, es notorio que este esquema tiene porcentajes de acierto mucho mayores que el propuesto para reducción. Cabe notar que para este esquema, una dimensión de espacio reducido alto lleva a disminuir los porcentajes de acierto obtenidos.

El ajuste simultáneo basado en MCE de la etapa de reducción de características y de la clasificación basada en HMM, permite reducir en forma directa el error de clasificación, esta característica no se puede garantizar cuando se emplea el criterio ML. Esto se debe a que el enfoque discriminativo permite maximizar la distancia entre los modelos de clases diferentes, ajustando sus parámetros sobre una función de costo que evalúa el error de clasificación en cada iteración; así cada modelo es estimado utilizando al igual que la información propia de la clase a la que pertenece, el modelo la información de las demás clases. Mientras que, el criterio de máxima verosimilitud apunta a encontrar la probabilidad máxima por clase y no se preocupa por los límites de decisión entre clases, esto hace que no se garantice que el error de clasificación tienda a minimizarse.

Los esquemas propuestos pueden ser comparados con:

– La metodología en [33, 31], que trabaja los HMM como modelos generativos, donde las observaciones generadas por diferentes modelos: auto-regresivo, de componentes principales y de componentes independientes. En [32] aplican el modelo de componentes independientes ocultas de Markov (*hidden Markov independent components*) al análisis de señales biomédicas , debido a que son altamente no estacionarias. Específicamente se aplica a señales ECG (señales electroencefalográficas) para detectar ciertos estados motores.
– El enfoque en [38], que propone utilizar el modelo de reducción de dimensionalidad conocido como análisis de factores (FA - *factor analysis*), al interior de un HMM; cuyos parámetros son elegidos maximizando la verosimilitud de los datos observados (ML) ó minimizando el error de clasificación (MCE); para utilizarlo como modelo de reconocimiento de patrones en señales de voz.
– En [4], se muestra una propuesta metodológica basada en la extracción de características y entrenamiento simultaneo de HMM basada en MCE; que realiza

en las densidades de observación, transformaciones lineales asociadas a la reducción de dimensionalidad basada en MCE. Se aplica el enfoque propuesto al reconocimiento de disfunciones en señales de voz.

Si bien los primeros dos esquemas no tiene como fin la reducción dinámica de características; pero si tienen forma similar al esquema propuesto para clasificación directa basada en HMPCA.

Las contribuciones de este trabajo consisten principalmente en haber desarrollado el algoritmo HMPCA como una función discriminativa, que permita que la estimación del modelo esté directamente relacionada con la disminución del error de clasificación que la hace apropiada para trabajar en tareas de reconocimiento de patrones. Además, de poder obtener un modelo de reducción dinámica que utiliza la información de matrices de covarianza localizadas dadas por un HMM que permita obtener las relaciones entre los datos del espacio original en un espacio compacto.

Aunque las metodología de reducción y clasificación propuestas en este trabajo han permitido obtener porcentajes altos de acierto, es importante realizar un estudio detallado de ciertas tareas:

La sintonización de los parámetros que definen la función de suavización en MCE, que están directamente relacionados con el tipo de señales que se desean analizar; estos parámetros requieren un estudio que permita encontrar un conjunto óptimo que apunte a una mejor definición de los limites de decisión entre clases. La técnica asociada a la optimización de parámetros en el criterio MCE, que puede mejorarse, utilizando gradiente conjugado o escalado. O en su defecto, mejorar la adaptación de la tasa de aprendizaje, en el gradiente de tal forma que permita acelerar la convergencia.

Hay que tener en cuenta que los criterios de estimación de parámetros, ML y MCE, no garantizan mínimos globales cuando ajustan los parámetros del modelo; una forma de dar solución a este problema es utilizar técnicas de optimización convexa. Una propuesta se muestra en [39], en donde se extiende el uso del criterio de margen máxima (*large margin*) a la estimación de parámetros del HMM. Esta propuesta se utiliza para reconocer fonemas dentro de una frase, aquí cada frase está representada por un estado dentro del HMM, así cada estado hace referencia a una clase. Para aplicar el criterio de margen máxima al modelo HMPCA, debe entonces estudiarse como extender el criterio para que sea aplicado ya no entre estados de un mismo HMM, sino entre HMMs que representen clases diferentes.

La selección del modelo, puesto que el esquema utilizado en esta tesis es computacionalmente ineficiente (validación cruzada). Esto quiere decir, encontrar el número de óptimo que mejor explique el comportamiento dinámico de los datos; debido a que

a mayor número de estados el modelo se hace más grande y mínimo por cada estado existe un modelo PPCA asociado.

El número de componentes de mezcla por estado, que en el caso del esquema de clasificación fue 1 con el fin de trabajar con modelos lo menos complejos posibles.

Adicionalmente, como encontrar la dimensión de espacio reducido que brinde resultados óptimos a medida que se transforma el espacio inicial de los datos.

Bibliografía

[1] M. Alvarez, G. Castellanos, J. F. Suárez, and J. I. Godino, "Identificación de Patologías de Voz usando HMM," in *Libro de Actas del XXII Congreso Anual de la Sociedad Española de Ingeniería Biomédica*, Noviembre 2004, pp. 217–220.

[2] M. Alvarez and R. Henao, "Hidden markov bayesian principal component analysis," in *Proceedings of the 14th International Conference on Neural Information Processing (ICONIP2007)*, 2007.

[3] M. A. Alvarez, "Reducción de dimensión de características dinámicas empleando procesos markovianos aplicados al reconocimiento de disfunciones en bioseñales," Master's thesis, Universidad Tecnológica de Pereira, Colombia, 2006.

[4] J. D. Arias, "Reducción de espacios de entrenamiento empleando modelos ocultos de markov basados en entrenamiento discriminativo," Master's thesis, Universidad Nacional de Colombia Sede Manizales, 2007.

[5] L. R. Bahl, P. F. Brwon, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, 1986.

[6] ——, "Estimating hidden markov model parameters so as to maximize speech recognition accuracy," *IEEE Transaction on Speech and Audio Processing*, vol. 1, pp. 77–83, 1993.

[7] L. Bahl, F. Jelinek, and R. Mercer, "A maximum likelyhood approach to continuous speech recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. PAMI-5, pp. 179–190, 1983.

[8] J. Bilmes, "A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," Dept. of EECS, CS Division, U.C. Berkeley, ICSI-TR-97-021, 1997.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[10] A. P. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of The Royal Statistica Society B*, vol. 39, pp. 1–38, 1977.

[11] A. Dibazar and S.Ñarayanan, "A system for automatic detection of pathological speech," in *Proceedings of the 36th Asilomar Conf. Signals, Systems & Computers*, 2002.

[12] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, segunda edición ed. John Wiley & Sons, INC, 2001.

[13] R. Durbin, *Biological Sequence Analysis: Probabilistics Models of Proteins and nucleic acids*. Cambridge University Press, 1998, ch. Markov Chains and Hidden Markov Models.

[14] Y. Ephraim and L. Rabiner, "On the relations between modeling approaches for speech recognition," *IEEE Transaction on Information Theory*, vol. 36, pp. 372–380, 1990.

[15] O. D. C. Gómez, "Método para la Discriminación entre Voces Normales y Disfuncionales, basado en la Selección Efectiva de Parámetros Acústicos de la Voz. aplicación en la población de la zona centro de colombia," Ph.D. dissertation, Universidad Nacional de Colombia, 2004.

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

[17] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing*. Upper Saddle River, New Jersey: Prentice Hall, 2001.

[18] A. Jain, R. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, Enero 2000.

[19] C. R. Jankowski, H.-D. H. Do, and R. Lippmann, "A comparison of signal processing front-ends for automatic word recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, Julio 1995.

[20] I. Jolliffe, *Principal Component Analysis*, Spinger, Ed. Springer Verlag, 2002.

[21] B. H. Juang, *Pattern Recognition in Speech and Language Processing*, W. Chou, Ed., 2003.

[22] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, vol. 40, no. 12, Diciembre 1992.

[23] B.-H. Juang, C. Wu, and C. H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, Mayo 1997.

[24] J. Kaiser, B. Horvat, and Z. Kacic, "A novel loss function for the overall risk criterion based discriminative training of hmm models," in *ICSLP*, 2000.

[25] A. Lojle, Y. Ephraim, and L. R. Rabiner, "Estimation of hidden markov model parameters by minimizing empirical error rate," in *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, 1990, pp. 709–712.

[26] P. G. M.A. Mohamed, "Generalized hidden markov models-part ii," *IEEE Transactions on fuzzy systems*, vol. 8, no. 1, p. 82, 2000.

[27] S. Mallat, *A Wavelet Tour of Signal Processing.* San Diego, California: Academic Press, 1998.

[28] L. T. Niles, H. F. Silverman, and M. A. Bush, "Neural networks, maximum mutual information training, and maximum likelihoood training," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 1990, pp. 493–496.

[29] Y.Ñormandin, *Automatic Speech and Speaker Recognition.* Kluwer Academic Publishers, 1996, ch. Maximum Mutual Information Estimation of Hidden Markov Models.

[30] J.Ñouza, "Feature selection methods for hidden markov model based speech recognition," in *Proceedings of the 13th International Conference on Pattern Recognition*, 1996.

[31] W. Penny, R. Everson, and S. Roberts, *Advances in Independent Component Analysis.* Springer, 2000, ch. Hidden Markov Independent Component Analysis.

[32] W. Penny, S. Roberts, and R. Everson, "Hidden markov independent components for biosignal analysis," *First International Conference on Advances in Medical Signal and Information Processing*, pp. 244–250, 2000.

[33] W. D. Penny and S. J. Roberts, "Hidden markov models with extended observation densities," Imperial College of Science, Technology and Medicine, Tech. Rep., 1998.

[34] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of The IEEE*, vol. 77, no. 2, Febrero 1989.

[35] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition.* Prentice Hall., 1993.

[36] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Comunications*, vol. 17, pp. 91–108, 1995.

[37] S. K. Riis, "Hidden Markov Models and Neural Networks for Speech Recognition," Ph.D. dissertation, Technical University of Denmark, 1998.

[38] L. K. Saul and M. G. Rahim, "Maximum likelihood and minimum classification error factor analysis for automatic speech recognition." *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 115–125, 2000.

[39] F. Sha, "Larger margin training of acoustic models for speech recognition," Ph.D. dissertation, University of Pennsylvania, 2007.

[40] F. Sha and L. K. Saul, "Large margin hidden markov models for automatic speech recognition," pp. 1249–1256, 2007.

[41] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Academic Press, 1999.

[42] M. Tipping and C. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society, Series B*, vol. 21, no. 3, pp. 611–622, 1999.

[43] Y. Tsuruoka, J. Tsujii, and S. Ananiadou, "Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty," in *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, 2009, p. 477–485.