Universidad de las Ciencias Informáticas
Facultad 5

Título: Trayectoria del modelo del habla para el reconocimiento automático del locutor independiente del texto.

Trabajo de Diploma para optar por el título de
Ingeniero en ciencias Informáticas

Autor: Flavio Jorge Reyes Díaz

Tutor: Lic. Gabriel Hernández Sierra

Asesor: Dr. José Ramón Calvo de Lara

Ciudad de la Habana, junio 2009
"Año del 50 aniversario del triunfo de la Revolución"

Las ciencias aumentan la capacidad de juzgar que posee el hombre, y le nutren de datos seguros.

José Martí

DATOS DE CONTACTO

Tutor:

Lic. Gabriel Hernández Sierra

CENATAV – Centro de Aplicaciones de Tecnologías de Avanzada.

Ave. 7ma # 21812 % 218 y 222, Rpto. Siboney, Playa, Ciudad Habana, Cuba. CP: 12200.

Teléfonos: 272-0422, 272-1670, 272-1676 Fax: 273-0045. E-mail: gsierra@cenatav.co.cu

Web: http://www.cenatav.co.cu

Asesor:

Dr. C. José Ramón Calvo de Lara

CENATAV – Centro de Aplicaciones de Tecnologías de Avanzada.

Ave. 7ma # 21812 % 218 y 222, Rpto. Siboney, Playa, Ciudad Habana, Cuba. CP: 12200.

Teléfonos: 272-0422, 272-1670, 272-1676 Fax: 273-0045. E-mail: Jcalvo@cenatav.co.cu

Web: http://www.cenatav.co.cu

Agradecimientos:

Agradecerles a todos los que de una forma u otra me han brindado su apoyo, a Maria de los Ángeles (Mary), por sus grandiosos consejos y sabiduría, a todos los trabajadores del Cenatav, a todas mis amistades de universidad, a mis amigos Amed, Villanueva y Arianna por siempre estar presente a la hora cero, a mi tutor, por soportarme y dedicarme todo su tiempo y a mi novia (Maday) que siempre estuvo presente y me ha brindado todo su amor y dedicación.

Dedicatoria

A mis abuelas aunque no disfruten junto a mí este momento, verme graduando era su mayor sueño.

A mi madre y mi padre, por ser mis mayores consejeros en mi camino hasta aquí.

RESUMEN

El aumento de los niveles de seguridad en el acceso a departamentos, instalaciones, etc. conlleva a un crecimiento del interés de nuestra sociedad en la utilización de los patrones biométricos para el reconocimiento de la persona. Dentro de los patrones podemos encontrar las huellas dactilares, el iris, la escritura, la voz y otros.

El Reconocimiento automático de una persona mediante la voz es el proceso mediante el cual se verifica o identifica a un individuo mediante una señal desconocida. Este proceso presenta tres fases principales como son: la extracción de rasgos, la clasificación o entrenamiento y la identificación o prueba. Hasta la actualidad se han venido desarrollando algoritmos de enfoques estadísticos y discriminativos, todos trabajando sobre un mismo modelo. No obstante se mantienen los problemas ante la variabilidad del canal y el estado del hablante, motivando al desarrollo de investigaciones en la búsqueda de nueva información con el objetivo de robustecer los sistemas de reconocimiento automático del locutor.

Este trabajo analiza la información implícita en la representación de los rasgos del locutor en el espacio acústico y la dinámica de la señal en el tiempo, con el propósito de encontrar nuevos datos para la creación del modelo que mejoren los resultados logrados hasta la actualidad.

Se desarrollaron algoritmo utilizando estas dos variantes, fusionadas con el método “Modelo de mezclas gaussianas” (GMM) obteniendo como resultados 60% de mejoras, en un experimento controlado; y 24% en otro experimento no controlado en comparación con las GMM clásicas.

PALABRAS CLAVE

Reconocimiento de patrones, identificación, verificación, clasificación, coeficientes Mel, modelos, valores y vectores propios.

INDICE GENERAL

INTRODUCCIÓN

El interés de la sociedad por utilizar patrones biométricos para identificar o verificar la autenticidad de las personas ha sufrido un aumento drástico, que se refleja no solamente en novelas, películas y series de TV, sino también en la aparición de diversas aplicaciones prácticas como sistemas de acceso a instalaciones, ordenadores y teléfonos móviles, mediante las huellas dactilares, el iris, la escritura, la voz y otros.

El reconocimiento automático de una persona por su voz, es actualmente un área de investigación y de desarrollo de aplicaciones de gran importancia. Tan antigua en el establecimiento de sus principios teóricos como el reconocimiento automático del habla, ha tenido sin embargo, poca atención y como consecuencia, un desarrollo menor. Ha sido en estos últimos años, cuando a la luz de los nuevos e importantes campos de aplicación surgidos, su desarrollo ha sido mayor.

Los sistemas automáticos de reconocimiento del locutor, conllevan las siguientes fases, propias de los métodos de Reconocimiento de patrones:

- *Extracción de Rasgos:* Extracción de los rasgos acústicos de la señal manteniendo la información temporal de cuando se van produciendo.
- *Entrenamiento:* Obtención de los modelos y umbrales correspondientes a cada uno de los integrantes de la población de locutores
- *Prueba:* Comparación entre los modelos de la población y las muestras de voz del sospechoso, que permitirá tomar decisiones acerca de la identidad del mismo.

Descripción de las fases de un sistema automático de reconocimiento del locutor (RAL).

Extracción de Rasgos:

1. Elegir una frecuencia de muestreo apropiada.
2. Seleccionar la trama con la que vamos a trabajar
3. Cálculo de los coeficientes Mel.

Entrenamiento:

1. Adquisición adecuada de las muestras de voces conocidas.
2. Mejoramiento de la calidad de las muestras, sin afectar su inteligibilidad ni sus rasgos para el procesamiento posterior.
3. Extracción automática de rasgos acústicos de las muestras.
4. Aplicación de métodos de Reconocimiento de Patrones a los rasgos extraídos de las muestras, para entrenar al sistema**,** creando los modelos y clases necesarios para la posterior clasificación.

Prueba:

1. Adquisición adecuada de la muestra de voz desconocida.
2. Mejoramiento de la calidad de la muestra, sin afectar su inteligibilidad ni sus rasgos para el procesamiento posterior.
3. Extracción automática de rasgos acústicos de la muestra.
4. Identificación de los rasgos extraídos con los modelos creados en la fase de entrenamiento.
5. Establecimiento de conclusiones dentro de un marco Bayesiano.

Dentro del RAL podemos distinguir dos tareas diferenciadas:

- *Verificación Automática del Locutor (VAL).* El objetivo es verificar la identidad reclamada por el locutor, o sea, tenemos un individuo que dice ser alguien y una muestra de su voz, la tarea a realizar es ver si ambas coinciden o no; la respuesta del sistema será, por lo tanto, binaria: identidad aceptada o rechazada.
- *Identificación Automática del Locutor (IAL).* Aquí el objetivo es, dada una muestra de voz, señalar, dentro de un grupo de personas, los propietario más probables de la muestra.

Sadaoki Furui, ingeniero de NTT Human Interface Laboratories, en Tokio, y una de las autoridades del momento actual en sistemas automáticos de reconocimiento de locutores, define perfectamente los anteriores conceptos:

- Reconocimiento de locutores: Todo proceso automático de reconocimiento de hablantes basado en la información individual incluida en la señal de habla. Dicho proceso se divide en Identificación y Verificación de hablantes.

- Identificación de hablantes: Proceso por el que se determina a quien pertenece la muestra anónima aportada, de entre un número de muestras registradas pertenecientes a distintos hablantes.

- Verificación de hablantes: Proceso de aceptación o rechazo de identidad a través de voz, solicitado por un hablante.

En relación con dichos conceptos, Furui [1] puntualiza:
"*...la diferencia fundamental entre identificación y verificación es el número de decisiones alternativas. En identificación, el número de decisiones alternativas es igual al número de sujetos de la población que conforma la base de datos, mientras que en verificación sólo existen dos decisiones alternativas, aceptar o rechazar, con independencia de la talla de la población…*"

Se han venido desarrollando a lo largo de la historia una serie de métodos estadísticos, discriminativos y la fusión de ambos, todos trabajando sobre el mismo modelo estadístico de la expresión de voz. Aunque los nuevos algoritmos han sido capaces de mejorar los resultados del reconocimiento del locutor aun persisten problemas en cuanto a la variabilidad del canal y el estado del locutor, lo que trajo consigo la creación de un nuevo proyecto en el Centro de Aplicaciones de Tecnología de Avanzada (CENATAV), el cual aborda el proceso de reconocimiento automático del locutor independiente del texto, con el objetivo de buscar nuevos caminos o nueva información que sea capaz de robustecer los resultados obtenidos hasta hoy. De ahí que los resultados de nuestro trabajo tributan a este proyecto.

Es por todo lo planteado anteriormente que se precisó el siguiente *problema científico:*

¿Cómo encontrar nueva información que caracterice al locutor para hacer más robustos los sistemas de reconocimiento de locutor ante la variabilidad del canal y el estado del hablante?

Por lo tanto el *objeto de estudio* de este trabajo es el Reconocimiento Automático del Locutor independiente del texto.

De ahí que nuestro *campo de acción* es: la creación del modelo del locutor y la clasificación dos de las principales etapas en los sistemas de reconocimiento del locutor.

Los seres humanos sólo son capaces de generar una gama limitada de sonidos que ocupan una región confinada del espacio acústico. En este caso, podemos asumir que los rasgos de voz yacen sobre una variedad incrustada en el espacio acústico de altas dimensiones y nos proponemos como *objetivo general* obtener información en la trayectoria de la voz que represente nueva información del locutor para elevar la efectividad del reconocimiento.

De ahí que surgen las siguientes *tareas investigativas* a resolver:

- Establecer el estado del arte sobre los métodos que modelan la información estática y dinámica de la voz.
- Identificar los puntos principales para establecer un soporte adecuado que permita el desarrollo efectivo de soluciones nuevas.
- Desarrollar algoritmos para obtener información en la trayectoria de la voz y su clasificación.
- Diseñar e implementar un experimento para probar la efectividad del modelo-dinámico con bases de datos con gran variabilidad de la voz.

Para el desarrollo de la investigación se utilizaron métodos del nivel teórico, empírico y matemáticos-estadísticos.

Entre los *métodos teóricos* utilizados se encuentran:

Análisis-Síntesis, el cual proporcionó la información necesaria del estado actual del objeto de investigación, al considerarse diversos autores que han trabajado el tema y sus resultados.

Además permitió profundizar en los estudios existentes del reconocimiento del locutor, la estructura de los modelos acústicos.

Histórico-Lógico, el cual se empleó para conocer el desarrollo de la evolución de los métodos o algoritmos de reconocimiento del locutor, con el estudio del estado del arte y una serie de artículos del tema.

*Inducción-deducción***,** permitió el estudio de elementos particulares para lograr la elaboración de conclusiones generales y viceversa durante el proceso de estructuración de la propuesta.

*Modelación***,** posibilitó representar las diferentes clases de rasgos acústicos existentes en el espacio universal de sonidos de un locutor, lo que facilitó que se descubrieran las relaciones y los principales rasgos del locutor.

Como *métodos empíricos* se utilizaron los siguientes:
Observación, se aplicó en el monitoreo del entrenamiento de los modelos obtenidos de los rasgos de la señal de la voz.

Métodos Matemáticos-Estadísticos fueron utilizados para el procesamiento de la información obtenida a través de los instrumentos y técnicas del nivel empírico, el uso de las propiedades de la matemática funcional y probabilística, para la creación de los modelos del habla y el entrenamiento. Se utilizaron las posibilidades de las aplicaciones informáticas para procesar la información cuantitativa, realización de pruebas y demostraciones.

Se utilizo como *herramienta* en el proceso de implementación:

- El editor matemático *Matlab*.

Para la culminación de este trabajo se deberá obtener una serie de *aportes* de ayuda para el desarrollo de esta rama y para trabajos en un futuro como son:

- Nueva información que caracterice el habla para el reconocimiento del locutor independiente del texto.
- Un demo o prueba que demuestre lo antes investigado.

- Un aumento del conocimiento acerca del tema.

Desde el punto de vista científico, los resultados teóricos a obtener permitirán hacer más efectivas las técnicas actuales de modelación del habla, elevando su robustez, lo que permitirá elevar la efectividad de los clasificadores en el reconocimiento del locutor. Los nuevos métodos serán incorporados a los proyectos aplicados relacionados con las tecnologías del habla que se desarrollan en el Centro de Aplicaciones de Tecnología de Avanzada, lo cual posibilitará la obtención de aplicaciones autóctonas de tecnologías del habla más eficientes y efectivas.

CAPÍTULO 1: ESTUDIO DE LOS MODELOS UTILIZADOS PARA REPRESENTAR AL LOCUTOR

Las necesidades de la sociedad del perfeccionamiento de los métodos y algoritmos para el reconocimiento del locutor han llevado a un profundo estudio de esta temática. El reconocimiento del locutor se divide en tres etapas fundamentales, la extracción de rasgos, el entrenamiento y las pruebas. Nuestro trabajo abordará la etapa de entrenamiento del locutor, para la investigación de esta etapa, se realizó un estudio del estado del arte desde los años 60 hasta la actualidad.

1.1 Algoritmos utilizados en el Reconocimiento Automático del locutor

Desde la década de los 70 han estado en evolución los métodos de reconocimiento del locutor, donde predominaba la clasificación comparando plantillas de palabras. Ya en los 80 se clasificó aplicando la Cuantización Vectorial (VQ) y la Distorsión Dinámica en el Tiempo (DTW). A mediados de los 90 se comenzaron a obtener grandes logros en los resultados del reconocimiento del locutor con enfoques estadísticos como los Modelos Ocultos de Markov (HMM) y los Modelos de Mezclas Gaussianas (GMM). Ya a principios del 2000 comienzan a aparecer diferentes adaptaciones de los GMM con el objetivo de mejorar los resultados de estos, donde aparecen los Modelos Universales de Background (UBM) y la Adaptación Máximo a Posteriori (MAP), la cual resaltó, ya que constituyó un paso clave para las mejoras de los clasificadores RAL.

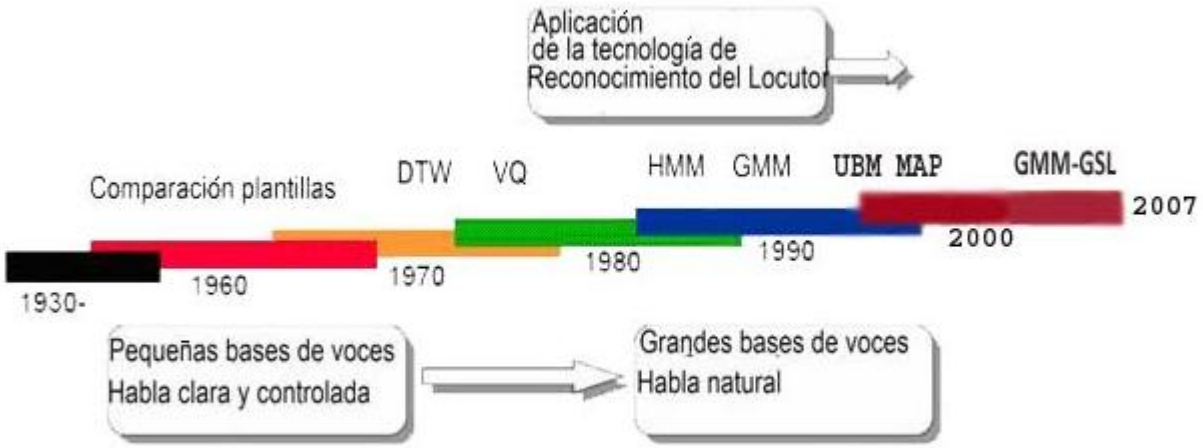

Fig. 1. 1 Evolución de la utilización de herramientas de Reconocimiento de Patrones en el reconocimiento automatizado del locutor hasta el 2001.

A partir de aquí se comienza a desarrollar un enfoque de clasificación discriminativa utilizando los resultados de los GMM, como los GMM Supervector Linear Kernel (GSL) que utilizan supervectores construidos a partir de las medias de los modelos GMM adaptados MAP, para clasificar los locutores con máquinas de soportes vectoriales (SVM). El estudio del estado del arte y las investigaciones nos han llevado a la siguiente comparación entre los algoritmos de clasificación del locutor.

A continuación les presentamos una tabla que puntualiza las ventajas y las desventajas de estos algoritmos:

Algoritmos	Ventajas	Desventajas
Distorsión Dinámica en el Tiempo (DTW)	Detectan y comparan tramos fonéticos de alta estabilidad (vocales abiertas, consonantes nasales) aplicando técnicas de correlación cruzada, coherencia, entre otras, para la medida de distancias. Los sistemas DTW	Los principales inconvenientes de estos sistemas se relacionan con la enajenación de la información a nivel suprasegmental y la necesidad de supervisión en las tareas de segmentación.

	han sido utilizados en algunas metodologías forenses como un complemento a otros análisis clásicos.	
Cuantización Vectorial(VQ)	Reducción sensible de la capacidad de almacenamiento en el cálculo del análisis espectral y una reducción de la complejidad computacional en el cálculo de distancias (se puede usar cálculos tan simples como la distancia Euclidiana o la de Mahalanobis).	Sus inconvenientes más significativos están relacionados con la distorsión espectral por el error de cuantificación (al representar cada vector por un representante).
Redes Neuronales (ANN)	Las redes son robustas al ruido y permite tener en cuenta el contexto de la señal, pueden crearse redes que tengan un funcionamiento similar a las VQ, a las GMM, HMM y otros algoritmos en el reconocimiento del locutor.	Presentan como limitantes que la mayoría de las redes requieren almacenar todos los datos del entrenamiento durante la clasificación, requiriendo en algunos casos un volumen apreciable de memoria y poder de cálculo.
Modelos Ocultos de Markov (HMM)	Su gran versatilidad, tanto en lo que se refiere a los procesos de entrenamiento como a ciertas características variables de la muestra: duración, contenido fonético o lingüístico, contexto, etc. A todo ello, hemos de añadir su gran adaptabilidad a la variación de las condiciones de voz o del canal de transmisión y, lógicamente, su funcionalidad en condiciones dependientes de texto.	Alto costo computacional, y sus mejores resultados se encuentran en el reconocimiento del locutor dependiente del texto.
Modelo de mezclas de Gaussianas (GMM)	Las GMM pueden representar con un alto grado de fidelidad un amplio margen de distribuciones muestrales, como es el caso con los diferentes coeficientes cepstrales que puede generar	Presentan un alto costo computacional implicando un tiempo considerable al crear los modelos.

	una locución. Además de las ventajas citadas, interesantes estudios comparativos sobre el rendimiento de diferentes técnicas de reconocimiento automático ante distintas circunstancias (procesos de entrenamiento, factores de degradación, etc.) han contribuido tomar como mejor sistema básico a las GMM.	

Nota: Las GMM a diferencia de los HMM no necesitan en la fase de entrenamiento segmentar en estados, ni entrenar la matriz de probabilidades de transiciones, además en la etapa de reconocimiento del locutor, no será necesario buscar la secuencia de estados de máxima verosimilitud, sino que bastará con acumular las probabilidades que asocia el modelo con cada uno de los vectores de entrada.

A partir de la investigación realizada se ha seleccionado como línea base para desarrollar las necesidades de este trabajo, a los Modelos de Mezclas Gaussianas (GMM), de ahí que el resultado de este trabajo será comparado con la eficiencia de estos modelos. Algunos de los motivos principales para usar GMM como línea base son:

- La idea intuitiva de que las componentes individuales de una densidad multi-modal son capaces de modelar las clases acústicas subyacentes en el proceso de identificación [2]; esto es, que el espacio acústico que caracteriza la voz de un individuo se puede aproximar mediante un conjunto de clases acústicas (que representan conjuntos amplios de eventos acústicos) como pueden ser las vocales, las consonantes nasales o fricativas. Estas clases acústicas denotan dependencia respecto a las configuraciones del tracto vocal específicas de cada locutor, siendo de gran utilidad a la hora de caracterizar a un hablante.
- Una agrupación es una estrategia de elección, por ejemplo el algoritmo de máxima expectancia empleado en los algoritmos de mezcla gaussiana (EMGMM) [3].

Es de interés de este capítulo precisar la definición de las GMM y la descripción de métodos de suma importancia para la comprensión de la investigación y la realización de todo el trabajo.

1.2 Modelos de Mezclas Gaussianas

Las GMM [1, 4-7] modelan los distintos vectores de rasgos de una locución, realizando una suma ponderada o mezcla de funciones de densidad de probabilidad Gaussiana.

Descripción del modelo

El modelo de densidades de mezclas Gaussianas es una suma pesada de *M* (número de mezclas) componentes de densidad descrita por:

$$p(\vec{x}/\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \qquad (1)$$

En la siguiente figura se muestra gráficamente la suma de las componentes de densidad:

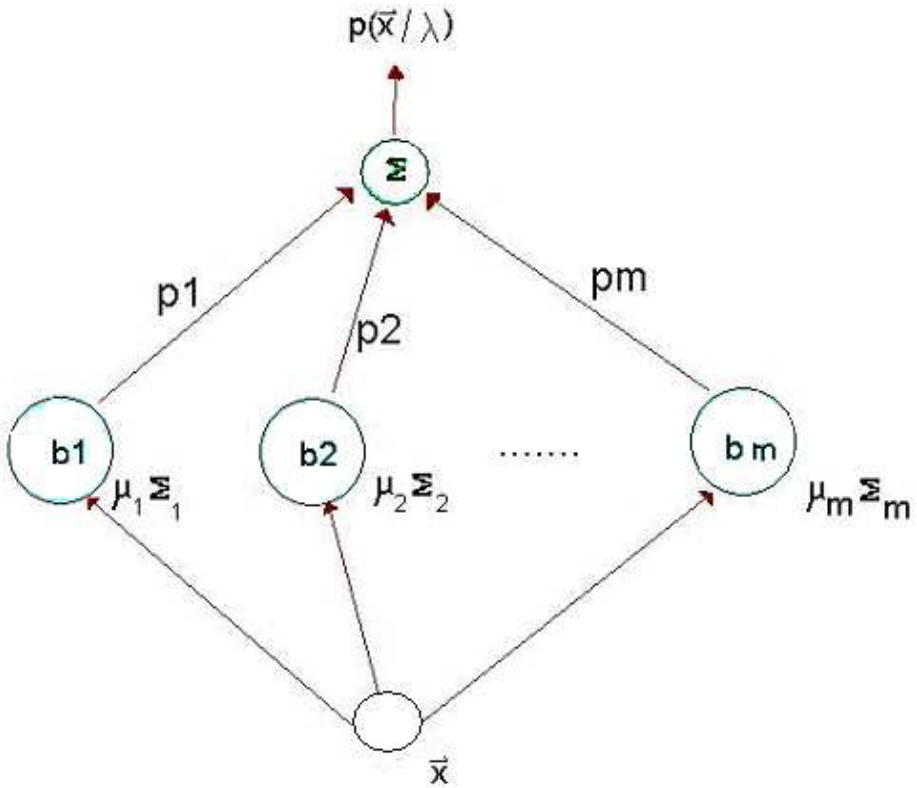

Fig. 1. 2 Descripción de un modelo de mezclas Gaussianas de M componentes.

donde:

- $\bar{x}$ es un vector *D-dimensional*

$$X = \begin{Bmatrix} \underset{\downarrow}{\vec{x}_1} & \underset{\downarrow}{\vec{x}_2} & \cdots & \underset{\downarrow}{\vec{x}_T} \\ c_{1,1} & c_{1,2} & \cdots & c_{1,T} \\ c_{2,1} & \cdots & & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ c_{Z,1} & \cdots & & c_{Z,T} \end{Bmatrix}$$ **Matriz de rasgos Mel**

La matriz X es una sucesión de variables aleatorias indexadas por una variable discreta, el tiempo (t = 1,…, T). Cada una de las variables aleatorias del proceso tiene su propia función de distribución de probabilidad y asumiremos que son independientes. X matriz de rasgos Mel (MFCC-Delta) [8, 9] obtenida de una expresión de voz de un locutor según la Fig. 1.3

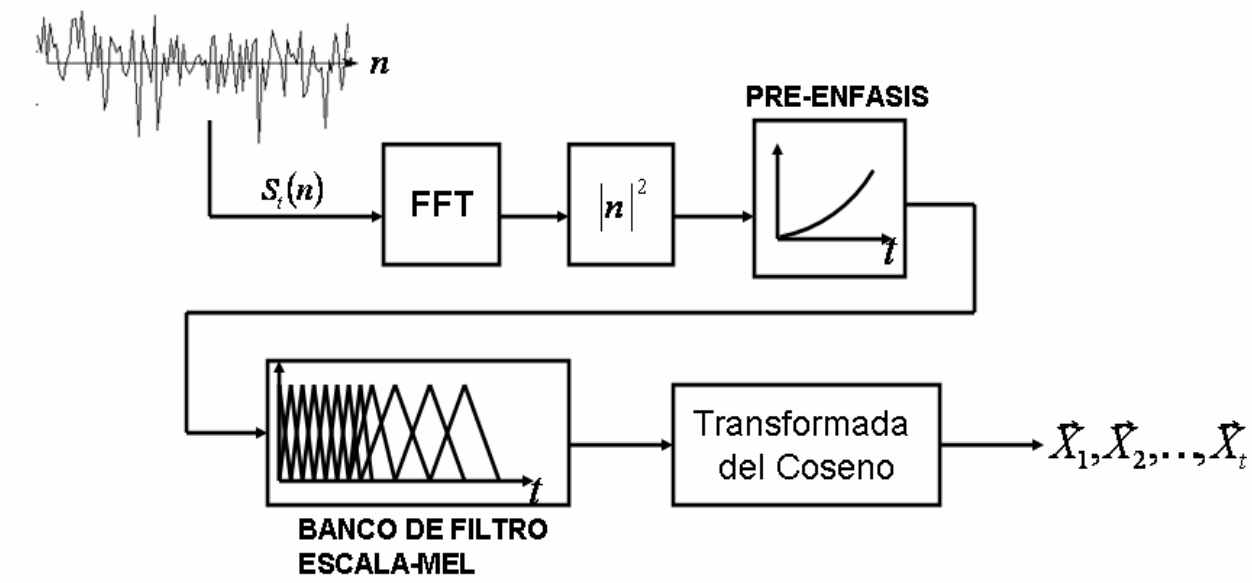

Fig. 1. 3 Rasgos cepstrales en escala Mel.

- $b_i(x)$ son las componentes de densidad, con $i = 1,2,\ldots,M$. Cada componente es una función Gaussiana de la forma:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}\left|\sum_i^{1/2}\right|} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)' \sum_i^{-1}(\vec{x}-\vec{\mu}_i)\right\} \quad (2)$$

Donde μ_i y $\sum_i$ son el vector de medias y el vector de covarianza correspondientes a la i-esima mezcla, los cuales son inicializados desde la matriz de rasgos *X.*

- p_i son los pesos de las mezclas con $i = 1, 2, \ldots, M$ y satisfacen la condición $\sum_{i=1}^{M} p_i = 1$.
- $\lambda = \{\bar{p}, \mu, \sum\}$ representa el modelo de mezclas Gaussianas, donde $\bar{p}$ es el vector de pesos, μ es la matriz de media y $\sum$ es la matriz de varianzas. Cada locutor es representado por un modelo λ obtenido utilizando el algoritmo EM (para más detalle [10]).

1.2.1 **Componentes de densidad Gaussiana**

Un factor crítico en el entrenamiento de las GMM es el número de las mezclas a seleccionar. Decidir el número de mezclas M para el modelo del locutor es un problema importante, ya que el objetivo es elegir el número mínimo de componentes de densidades Gaussianas necesarias para modelar adecuadamente un locutor.

- Elegir pocos componentes de la mezcla puede producir un modelo del locutor que no modele exactamente las características que lo distinguen entre el grupo de locutores.
- Elegir demasiados componentes puede llevar a que el costo computacional sea excesivamente alto a la hora del entrenamiento o la clasificación.

En la Fig. 1.4 se muestra el comportamiento de la identificación del locutor para varios órdenes del modelo (8, 16, 32, 64 y 128 componentes de densidad) y diferente duración de señal de prueba (5, 10 y 15 seg.). Para el entrenamiento se tomaron muestras de 1 minuto. Hay varias observaciones que se harán de estos resultados. Primero, la clara disminución del EER de la identificación de 8-32 componentes de la mezcla, la estabilidad de 32 - 64 y el aumento de este

de 64-128. La estabilidad del EER en el intervalo de 32-64 componentes de las mezclas indica que este es el número de componentes necesarios para modelar los locutores de la base de datos a utilizar y con un tiempo aproximado de un minuto de duración de las muestras, o sea los modelos deben contener por lo menos este número mínimo de componentes y esta duración para mantener un buen funcionamiento en la identificación del locutor.

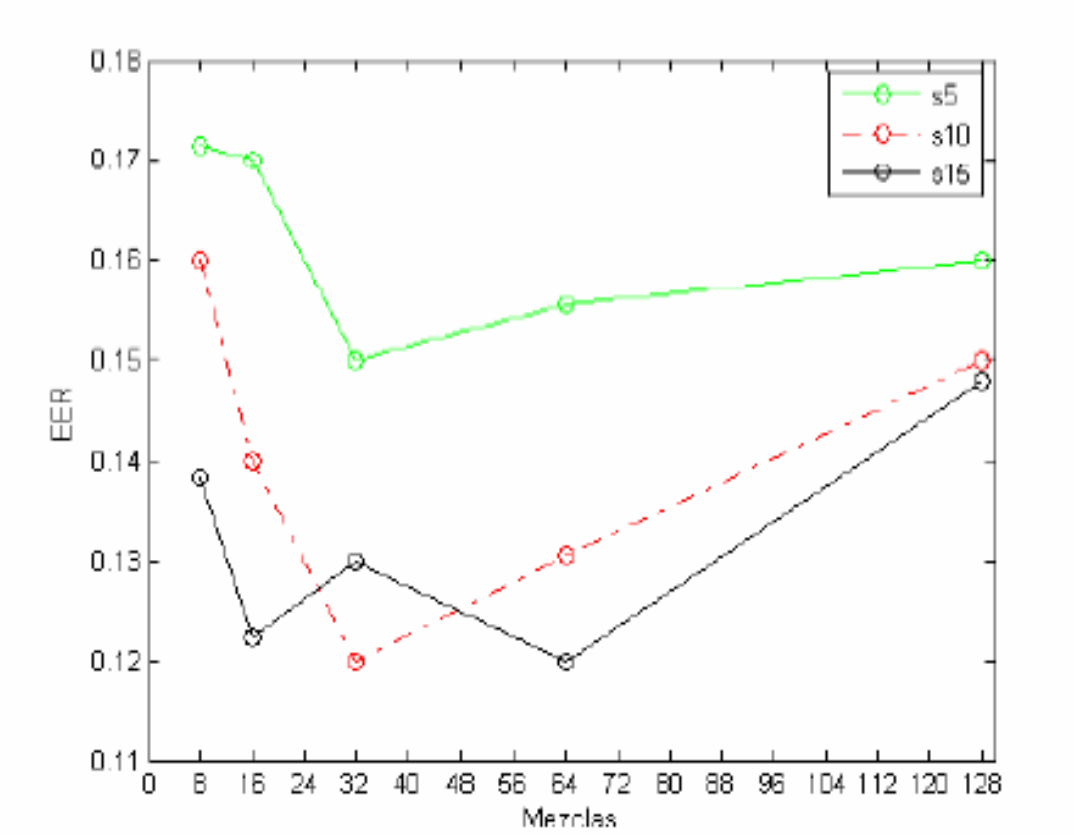

Fig. 1. 4 Error de identificación del locutor para diversos ordenes del modelo (8, 16, 32, 64, 128) y tiempo de la expresión de prueba (5, 10 y 15 segundos).

1.3 Equal Error Rate

En el mundo existen diversas formas de medir la efectividad de los sistemas RAL, una de las más utilizadas es la denominada tasa de EER: que consiste en medir el error del sistema cuando el umbral decisión es tal que el porcentaje de falsas aceptaciones es igual al de falsos rechazos (fig. 1.5) entonces si la salida del sistema es menor que el umbral de decisión la

identidad reclamada por el cliente es rechazada, en caso contrario será aceptada. Hasta el momento estos algoritmos y sus modificaciones han provocado grandes pasos de avance basados en los índices del EER que es el criterio utilizado para evaluar tanto el funcionamiento de los algoritmos como el progreso en los sistemas de reconocimiento del locutor, (línea base en el 2004, EER = 9.92%; sistemas fusionados en el 2005, EER = 7.20%; NAP [11], FA [12] y GSL-NAP [13], EER = 5.02%; GSL-FA-no supervisado en el 2007, EER = 2.27%), a pesar de esto sigue siendo un desafío clave para el reconocimiento del locutor el problema de la clasificación ante la variabilidad del canal y la sesión.

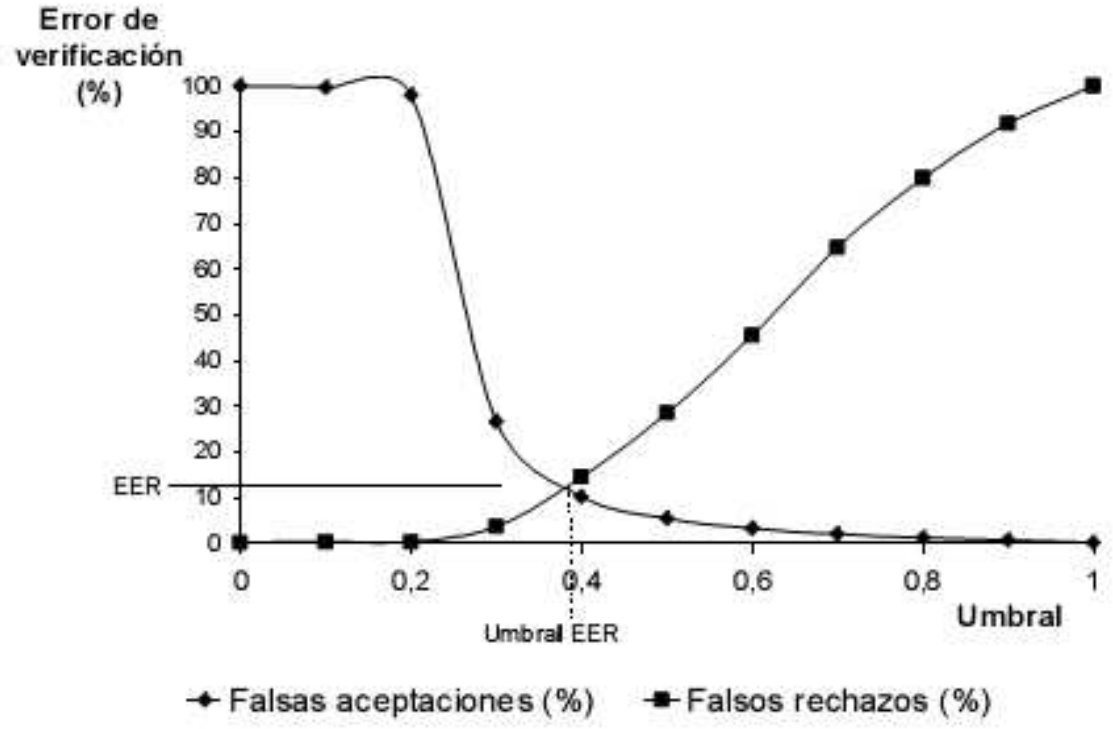

Fig. 1. 5 Curvas de error. Falsas aceptaciones (FA) y Falsos rechazos (FR). FA: la entrada pertenece a un impostor y es falsamente aceptada. FR: la entrada pertenece al cliente y es falsamente rechazada.

Teniendo en cuenta que todos estos algoritmos son utilizados en el reconocimiento automático del locutor trabajando sobre un mismo modelo, nos a llevado a la búsqueda de nuevos caminos o nueva información que sea capaz de mejorar los resultados obtenidos hasta hoy, ejemplo de esta nueva información podría ser la trayectoria de los rasgos de la voz o del modelo, la cual

puede aportar más datos del locutor. Por lo que hace necesario la investigación del estado del arte donde se estudian los patrones dinámicos de la voz.

1.4 Estudio y análisis de los Patrones Dinámicos

En el estudio sobre los patrones dinámicos el Dr. Ke Chen en [14] los resume en tres aspectos fundamentales: la exploración-explotación de las representaciones eficaces, la explotación de la información contextual intrínseca y el uso del principio divide y vencerás.

- Exploración y Explotación de representaciones: están basados principalmente en dos problemas de la representación de patrones dinámicos. El primero es la exploración de una representación adecuada de los patrones dinámicos de aprendizaje para filtrar la información poco relevante. El segundo es el estudio y explotación de las diferentes representaciones de los patrones dinámicos que no son bien representados por una sola vía.
- Explotación de la información contextual intrínseca: Los patrones dinámicos frecuentemente tiene información mezclada, un criterio genérico de similaridad puede fallar para trabajos con una cierta precisión en la medida de similitud entre los patrones dinámicos para una tarea específica.
- El principio Divide y Vencerás: Propone dividir un problema de gran tamaño en varios más pequeños y posteriormente la fusión de estos.

El autor en [15] propone para el modelado del locutor un camino alternativo usando simultáneamente varias representaciones de la voz, para explotar o utilizar cada dato específico que brinde cada una de ellas. Desarrollando un algoritmo denominado Modelo de Mezcla Gaussiana Generalizada (GGMM).

Este método propone obtener un Modelo de Mezclas Gaussianas para cada expresión de la voz y posteriormente mezclar los modelos obtenidos para de esta forma llegar a un modelo generar de todas las expresiones.

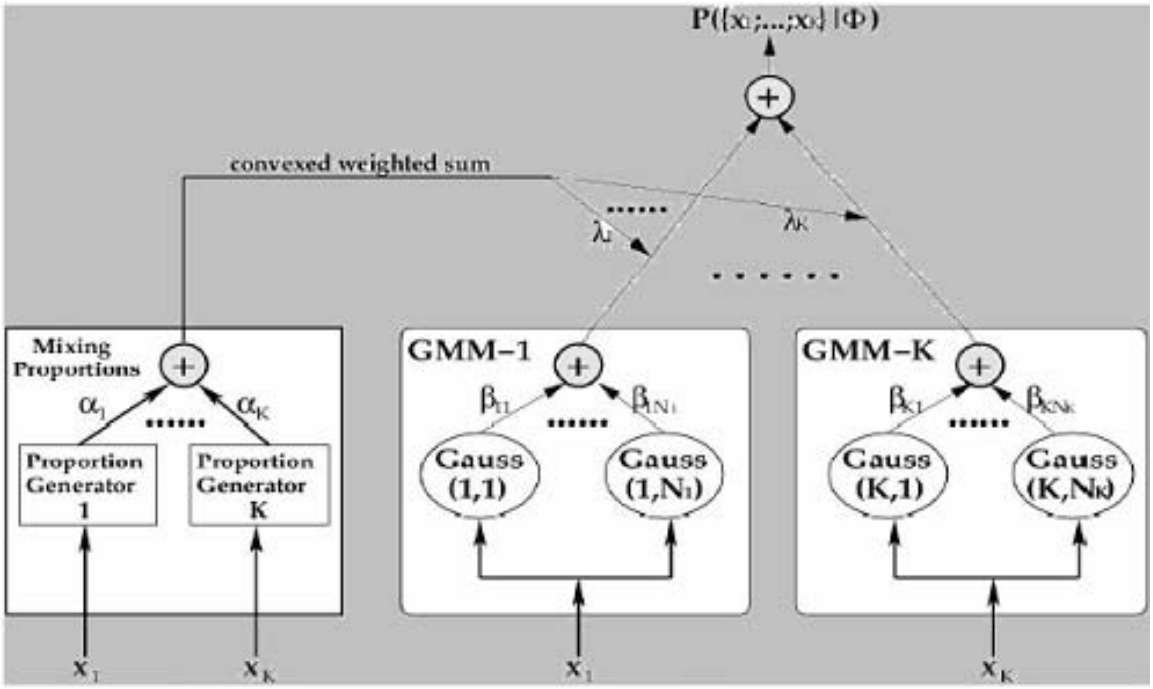

Fig. 1. 6 Descripción de los Modelos de Mezclas Gaussianas Generales.

Este algoritmo es un modelo general de mezclas finitas, que propone la adaptación del algoritmo de Máxima Expectancia (EM), lo que trae consigo que al inicializarse aleatoriamente exista un fallo en el alcance del máximo local, a diferencia de los GMM, que se comporta bien al inicializarlo de forma aleatoria.

En comparación con diversas técnicas de modelación del locutor basadas en la representación estática y simple de la voz las GGMM presentan un alto costo computacional.

A pesar de estos problemas el GGMM demostró que una representación simple y estática de la voz (GMM) no siempre produce la mejor representación en diferentes expresiones en términos de reconocimiento del locutor, como si la fusión de estos modelos, probando que existe información en la trayectoria de la voz que es importante para el reconocimiento de locutor.

Otro enfoque para la identificación del locutor independiente del texto realizado por Bing Xiang en [16] demostró que los segmentos en la trayectoria tienen información específica del locutor, analizando la entropía del locutor en el espacio de trayectoria, la cual puede usarse para

caracterizar a la persona. Para reflejar la dinámica de los rasgos en el espacio de trayectoria Xiang utiliza cadenas de componentes de densidad Gaussiana. Los componentes que conforman las cadenas tienen que ser consecutivos o sea tiene que pertenecer a tramas continuas.

Dada la poca información existente sobre los algoritmos que trabajan con la dinámica de la voz se decidió estudiar una nueva herramienta matemática: la Topología, la que se interesa por conceptos como proximidad, número de agujeros, el tipo de consistencia (o textura) que presenta un objeto entre otros múltiples atributos donde destacan conectividad, compacidad, metricidad, etcétera; para buscar nueva información en la trayectoria del modelo.

Para obtener una nueva representación se utilizaran los espacios topológicos (Ver anexo 1); donde se estudian aquellas propiedades de los cuerpos geométricos que permanecen inalteradas por transformaciones continuas.

Para esto se realizo una búsqueda de artículos que abordaran las representaciones dinámicas de la voz tanto en el campo del reconocimiento del locutor como el del habla.

Andrew y John en [17] asumen que los sonidos acústicos yacen sobre una variedad de gran dimensión, por lo que el espacio de sonidos acústicos de la voz humana, el cual es un subconjunto de todos los sonidos existentes, yace en una sub-variedad de baja dimensión que esta incrustada en el espacio de altas dimensiones de todos los posibles sonidos. En este artículo Andrew y John evalúan las habilidades de los algoritmos de aprendizaje con variedades de separar las vocales en espacios de dimensiones bajas comparándolos con métodos clásicos para separar vocales. Dando como resultado que los algoritmos que utilizan variedades funcionan mejor que los métodos clásicos para separar vocales en espacios de dimensiones bajas.

Aren Jansen y Partha Niyogi en [18-20] motivan y demuestran la existencia de una variedad curva de dimensión baja para sonidos sonoros de la voz. También presentan una nueva técnica de espectrograma basada en la estructura de las variedades y llegan a la conclusión de que esta representación permite una mejor distinción fonética en espacios de bajas

dimensiones. A su vez tienen como objetivo explotar las implicaciones desde el punto de vista geométrico en el habla humana y muestran que la estructura variedad de los sonidos de la voz puede ser explotada por reducción de la dimensionalidad, aprendizaje semi-supervisado y representación del habla con algunas mejoras tanto el voz artificial como en los datos de voz real.

Realizan una comparación a partir del uso de la variedad para representar los fonemas /a/ y /ae/ utilizando una variedad basada en funciones propias Laplacianas y el algoritmo tradicional de Análisis de componentes principales.

De estos tres artículos nos centraremos en el estudio de la existencia de una estructura variedad para sonidos sonoros de la voz.

Teniendo en cuenta lo anterior y que los sonidos sonoros de la voz forman una variedad, entonces la representación y clasificación de la voz requieren estimar mapas funcionales donde el dominio seria esta variedad. Esto incita al uso se ciertos algoritmos que se pueden utilizar para la reducción de dimensionalidad y la explotación de la estructura geométrica.

Algoritmos de reducción de dimensionalidad que se basan en la estructura de los patrones dinámicos

Locally Linear Embedding

El algoritmo no supervisado LLE [21] procesa datos de dimensiones bajas incrustadas en datos de dimensiones altas. Este algoritmo mantiene los puntos que son vecinos en el espacio de altas dimensiones como vecinos en el de bajas dimensiones, por lo que preserva las características locales, de las vecindades de los puntos.

Este algoritmo consiste en:

Primer paso — Para cada punto de datos X_i calcular sus **K** vecinos más cercanos, basado en la distancia Euclidiana.

Segundo paso — Calcular los pesos W_{ij} para una mejor reconstrucción a partir de sus

vecinos, minimizando la reconstrucción del error E:

$$E(W) = \sum_i \left| X_i - \sum_j W_{ij} X_j \right|^2 \quad (3)$$

Tercer paso Calcular la dimensión baja incrustada de y_i, según los pesos, minimizando el costo funcional Ω **:**

$$\Omega(W) = \sum_i \left| Y_i - \sum_j W_{ij} Y_j \right|^2 \quad (4)$$

En el paso 2 la reconstrucción del error es minimizado según dos restricciones: primero, si cada entrada se reconstruye solo por el vecino más cercano entonces $w_{ij} = 0$ si solo si x_i no es vecino de x_j, segundo, en la reconstrucción de los peso, cada uno de los puntos suma *1* $\sum_j W_{ij} = 1, \forall i$.

Isomap

El algoritmo Isomap [22] propone una motivación diferente para el aprendizaje de la variedad. Este algoritmo es una generalización no linear de Multidimensional Scaling (MDS), que busca un mapeo del espacio X de altas dimensiones preservando las distancia geodésicas entre los pares de puntos. Isomap y LLE presentan similares metas, en particular este algoritmo intenta preservar las propiedades geométricas globales de una variedad mientras que el algoritmo anterior preserva las propiedades geométricas locales.

Este algoritmo consta de tres pasos:

Primer paso: Construye un grafo de vecindad y determina cuales puntos son vecinos en la variedad basado en la distancia Euclidiana *d (i, j)* entre el par de puntos *i, j* en el espacio de entrada. Esa relación de vecindad es representada por un grafo pesado donde cada arista contiene un peso que consiste en la

distancia entre los nodos correspondientes *d (i, j)*.

Segundo paso: Buscar el camino mas corto entre todos los nodos utilizando técnicas como Dijkstra.

Tercer paso: Aplicar el algoritmo Multidimensional Scaling clásico [23] para incrustar los datos en el espacio de dimensión euclidiano deseado preservando la distancia geodésica.

Laplacian Eigenmaps

El principio de este algoritmo [24, 25] es calcular la representación de baja dimensión de los datos de alta dimensión preservando la relación de proximidad. Dado *K* puntos $x_1, \cdots, x_k$ en un espacio de *Z*-dimensiones, se construye un grafo pesado de *K* nodos. El mapa de baja dimensión incrustado en el espacio de altas dimensiones proviene del cálculo de los vectores propios de la matriz Laplaciana creada de los *K* puntos.

Primer paso Construir un grafo de vecindades como en el Isomap.

Segundo paso Asignar los pesos W_{ij} a las aristas del grafo. Existen tres variantes para la asignación de pesos a las aristas, si los nodos *i* y *j* están conectados podemos asignar constantes como $W_{ij} = 1/k$, un exponencial decadente como $W_{ij} = e^{\left(-\|X_i - X_j\|^2/\sigma\right)}$ donde σ es el valor de escalado, o simplemente W_{ij} =1 si y solo si los vértices están conectado por una arista.

Tercer paso Calcular los valores y vectores propios mediante el problema generalizado de vectores propios:

$$Lf = \delta Df \tag{5}$$

donde *L*= *D* – *W* es la matriz Laplaciana *y D* es la matriz diagonal de los

pesos, que se construye con la suma de cada columna(o fila si la matriz *W* es simétrica) de *W*, $D_{ii} = \sum_{j} W_{ij}$. Posteriormente se ordenan los vectores propios de menor a mayor según los valores propios correspondiente, se obvia el vector f_0 y se utilizan los primeros *N, N* << *Z* siendo *N* la nueva dimensión.

CAPÍTULO 2: DESARROLLO DE ALGORITMOS PARA LA CLASIFICACIÓN UTILIZANDO LA DINÁMICA DE LOS RASGOS

Esta investigación tiene como objetivo principal robustecer los sistemas del RAL ante la variabilidad del canal y la sesión, brindando como resultado el desarrollo de una serie de algoritmos con el objetivo de demostrar la existencia de información implícita en la trayectoria de la señal, útil para la optimización de los sistemas del reconocimiento del locutor.

2.1 Aplicación de una nueva representación de la señal en el reconocimiento del locutor

Con el análisis desarrollado en el capítulo anterior sobre los patrones dinámicos, al igual que los algoritmos que trabajan con variedades y partiendo que Aren Jansen y Partha Niyogi plantean que los sonidos acústicos de la voz humana yacen sobre una variedad de baja dimensión incrustada en la variedad de alta dimensión de todos los sonidos existentes, decidimos apoyarnos en los algoritmos que trabajan con variedades descritos en el capítulo 1, para obtener una mejor representación de los rasgos de la voz buscando robustecer la variabilidad del canal y la sesión. Estos algoritmos tienen como premisa calcular la baja dimensión de los rasgos iniciales que yacen en una variedad no lineal. Para obtener una nueva representación de los rasgos de la voz se utilizó el Laplacian Eigenmaps.

En la e*ntrenamiento* tenemos como datos iniciales a los rasgos acústicos que caracterizan al locutor, los cuales asumiremos que están en óptimas condiciones.

Partiendo de una matriz de rasgos Mel $X_{Z,T}$ (ver Fig. 1.3), se construye un grafo de vecindades determinándose los puntos vecinos en función de la distancia Euclidiana entre el par de puntos i, j en el espacio de entrada, tomando $d(i, j)$ como el peso de la arista entre ellos. Cada punto tendrá como vecino los *K* puntos más cercanos entre todos. A partir de la matriz de adyacencia obtenemos la matriz de peso W, asignándole a $W(i, j)=1$ siempre que exista una arista entre los puntos i y j, en caso contrario $W(i, j)=0$.

Se calcula la matriz Laplaciana $L = D - W$, siendo D una matriz diagonal donde se encuentra la suma de cada columna(o fila si la matriz *W* es simétrica) de *W*,

$$D_{ii} = \sum_{j} W_{ij} \tag{2.1}$$

Luego se calculan los vectores y valores propios [26] de la matriz Laplaciana según su problema generalizado:

$$Lf = \delta Df \tag{2.2}$$

Se ordenan de forma creciente los valores propios (δ) juntos a los vectores propios correspondientes y se toman los *N*-primeros obviando f_0, *N* << *Z,* (Z es la dimensión de X y *N* la nueva dimensión). Como resultado tenemos el espacio propio $\bar{f}_{Z,N}$ del locutor. Solo nos resta proyectar en el espacio propio los rasgos, obteniendo una nueva representación.

$$\overline{X}_{T,N} = X'_{Z,T} * \bar{f}_{Z,N} \tag{2.3}$$

Calculamos la transpuesta $\overline{X}_{N,T} = \overline{X}'_{N,T}$ de la proyección obteniendo una nueva matriz de rasgos de dimensión *N* y *T* tramas. Esta nueva representación la utilizaremos como rasgos iniciales para el entrenamiento del locutor.

Para el entrenamiento del locutor se utilizó las GMM básicas, que como se explicó en el capítulo anterior consiste en modelar las clases acústicas estimando los parámetros del modelo que maximicen la probabilidad. Para esto se utilizó el método llamado maximización de la verosimilitud (ML), que estima los parámetros del modelo para una secuencia de T vectores de entrenamiento.

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(x_t \mid \lambda). \tag{2.4}$$

Desafortunadamente esta expresión es no linear, debido a los parámetros del modelo y maximizarla directamente no es posible, sin embargo estimar los parámetros ML puede realizarse iteradamente utilizando el caso especial de la maximización de la expectancia (EM).

La idea básica del algoritmo de EM es estimar a partir de un modelo inicial λ un nuevo modelo $\overline{\lambda}$ que cumpla con $p(X \mid \overline{\lambda}) \geq p(X \mid \lambda)$. En la próxima iteración el nuevo modelo será el inicial y el proceso se repetirá hasta que converja, $p(X \mid \overline{\lambda}) - p(X \mid \lambda) \leq \varepsilon$ o cuando la cantidad de iteraciones sea mayor que un cierto número. En cada iteración del método se utilizan un grupo de fórmulas de reestimación que garantizan el crecimiento de la monotonía de la verosimilitud:

Peso de las mezclas:

$$\overline{p}_i = \frac{1}{T} \sum_{t=1}^{T} p(i \mid x_t, \lambda) \tag{2.5}$$

Medias:

$$\overline{\mu}_i = \frac{\sum_{t=1}^{T} p(i \mid x_t, \lambda) x_t}{\sum_{t=1}^{T} p(i \mid x_t, \lambda)}. \tag{2.6}$$

Varianzas:

$$\overline{\sigma}_i = \frac{\sum_{t=1}^{T} p(i \mid x_t, \lambda) x_t^2}{\sum_{t=1}^{T} p(i \mid x_t, \lambda)} - \overline{\mu}_i^2. \tag{2.7}$$

Donde la probabilidad a posteriori para la i-esima mezcla está dada por:

$$p(i \mid x_t, \lambda) = \frac{p_i b_i(x_t)}{\sum_{k=1}^{M} p_k b_k(x_t)}. \tag{2.8}$$

Para la creación del modelo Gaussiano se utilizaron parámetros como: número de mezclas, condición de parada por iteraciones, condición de parada por convergencia. Para esto se desarrolló un algoritmo en Matlab, del cual se muestra el seudo código a continuación:

Algoritmo Clasificación utilizando una nueva representación del locutor

Entrada: $(X_1, X_2, \ldots, X_S)$ **{Base de datos de locutores}**

Variables:
covar_type ← **'diag'**
Mezclas: entero potencia de 2
Opción: arreglo de enteros

Salida: $(\lambda_1, \lambda_2, \ldots, \lambda_S)$ **{S cantidad de modelos de los locutores}**

```
Inicio
    Desde i=1 hasta S
        X ← Rasgos de los Locutores(i)
        E ← Laplaciano(X) {Calcula vectores propios del locutor}
        espacio propio ← X' * Vectores Propios {espacio propio del Locutor}
        [Tramas, dimensión] ← tamaño (espacio propio)
        mix ← Función GMM (dimensión, mezclas, covar_type) {inicializa el modelo}
    {Inicialización de los parámetros para aplicar el algoritmo Máxima Expectancia}
        Opción ←    ceros (1, 18)              {arreglo de ceros}
        Opción (14) ←   200                    {Máximo número de iteraciones}
        Opción (3)  ←   0.001                  {Error mínimo}
        [Mix, opción, errlog] ← Función EM (mix, espacio propio, opción)
        Se almacenan los Vectores Propios del locutor en Mix
        Modelos (i) ← Mix {Se almacena el modelo del locutor  en la variable Modelos}
    Fin hasta
Fin
```

Ya entrenados los rasgos acústicos y partiendo de un grupo de locutores representados por $(\lambda_1, \lambda_2, \ldots, \lambda_S)$ se procede a la *identificación,* que consiste en encontrar los modelos que tenga la máxima probabilidad a posteriori para una señal de prueba desconocida $Y_{Z,T}$ y ordenarlos descendentemente. Donde $Y_{Z,T}$ se proyecta en su espacio propio (este es un estudio controlado de la nueva representación) utilizando la ecuación 2.3. Para la obtención de la máxima probabilidad entre todos los modelos nos apoyamos en la siguiente ecuación:

$$\hat{s} = \arg \max_{1 < k < S} \Pr(\lambda_k \mid X) = \arg \max_{1 < k < S} \frac{p(X \mid \lambda_k) \Pr(\lambda_k)}{p(X)} \qquad (2.9)$$

donde el miembro extremo derecho es la regla de Bayes.

$p(X \mid \lambda_k) \rightarrow$ Probabilidad de que la matriz de rasgos pertenezca al modelo λ_k , o sea que la expresión de voz de donde se extrae la matriz de rasgos X pertenezca al locutor *k*.

$\Pr(\lambda_k) \rightarrow$ Probabilidad del modelo k.

$p(X) \rightarrow$ La suma de las probabilidades de que la matriz de rasgos X pertenezca a todos los modelos.

Asumimos igualmente para todos los modelos que $\Pr(\lambda_k) = 1/S$ y note que $p(X)$ es la misma para todos también. Por lo tanto:

$$\hat{s} = \arg \max_{1 < k < S} \Pr(X \mid \lambda_k) \tag{2.10}$$

Sustituyendo (2.4) en (2.10) y aplicando el logaritmo dada la independencia entre observaciones, la identificación de los locutores queda:

$$\hat{s} = \arg \max_{1 < k < S} \sum_{t=1}^{T} \log p(x_t \mid \lambda_k) \tag{2.11}$$

, donde $p(x_t \mid \lambda_k)$ fue definido en (1.1).

A continuación se muestra el seudo código de la identificación:

Algoritmo Identificación utilizando una nueva representación

Entrada: $(\lambda_1, \lambda_2, \ldots, \lambda_S)$ **y** $Y_{Z,T}$

Salida: Probabilidad

```
Inicio
    Matriz proyectada ← matriz de los rasgos acústicos Y_{Z,T} en su espacio propio según la
    ecuación 2.3
    Desde k=1 hasta S
        Mix ← λ_k
        Probabilidades ← Función Calcular probabilidades (Mix | Matriz proyectada)
        Probabilidad ← mean (Probabilidades) {Calcular media}
        Log Likelihood (k) ← exp.(Probabilidad) {Se calcula el exponencial}
    Fin hasta
    Se ordena Log Likelihood de forma descendente
Fin
```

Al finalizar este algoritmo obtendremos la probabilidad ordenada descendentemente de todos modelos de los locutores.

2.2 Algoritmo utilizando la segmentación de la señal en la fase de entrenamiento

Teniendo en cuenta lo expuesto en la referencia [16], que nos plantea que la trayectoria de la señal de la voz presenta información implícita que caracteriza al locutor, nos dimos a la tarea de buscar un nuevo atributo incrustado en la dinámica de la señal acústica para el mejoramiento de los sistemas RAL.

Para buscar la información en la trayectoria de la voz dividimos la matriz de rasgos como se muestra continuación:

$X_{Z,T} = X_1, X_2, \ldots, X_N$, donde N es la cantidad de segmentos en que se puede dividir la matriz inicial, se puede calcular N de la forma:

$$N = T \diagup seg \tag{2.12}$$

Donde X es la matriz de rasgos, *seg* tamaño del segmento y T el número máximo de tramas. De esta forma garantizamos que los modelos gaussianos de cada segmento converjan a un resultado óptimo. A continuación se observa de forma gráfica la segmentación de una señal de voz:

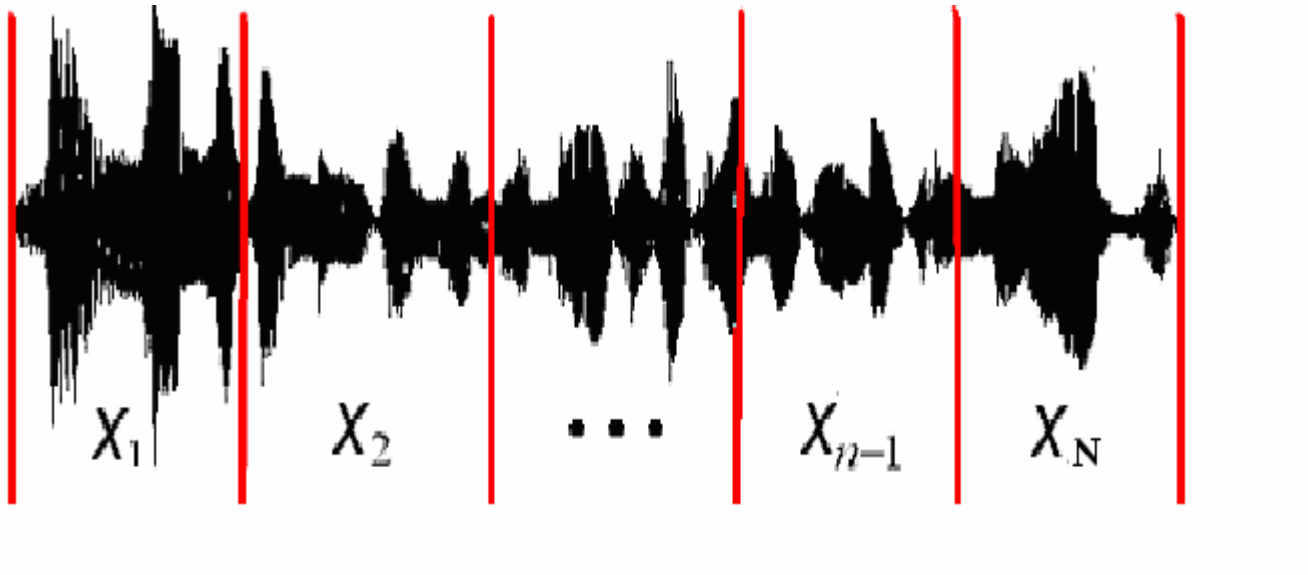

Fig. 2. 1 Señal dividida en segmentos

A partir de este momento por cada segmento X_n, con $n = 1,2,\ldots,N$; entrenamos un modelo utilizando el algoritmo de Modelos de Mezclas Gaussianas, obteniendo un $\overline{\lambda}_n = \{\overline{p}_n, \overline{\mu}_n, \overline{\Sigma}_n\}$ por cada segmento de la matriz que caracteriza al locutor, con los parámetros: número de mezclas, condición de parada por iteraciones, condición de parada por convergencia. Para esto se desarrolló el algoritmo denominado '*Clasificación utilizando la señal segmentada*' que se muestra a continuación:

Algoritmo Clasificación utilizando la señal segmentada

Entrada: $X_1, X_2, \ldots, X_S$

Variables:
covar_type $\leftarrow$ **'diag'**
Mezclas, seg: entero{ seg número de tramas a tomar como segmento}
Opción: arreglo de enteros

Salida: $\overline{\lambda}_1$, $\overline{\lambda}_2$, … , $\overline{\lambda}_N$ **{Modelos que cumplen con** $p(X \mid \overline{\lambda}) - p(X \mid \lambda) \leq \varepsilon$ **}**

Inicio
  Desde s=1 hasta S
    Se segmenta X_s **en función de seg,** $X_s = X_1, X_2, \ldots, X_N$
    dimensión $\leftarrow$ **seg**
    Desde *n*=1 hasta *N*
      $X_n \leftarrow$ **numero de tramas correspondientes al intervalo *n***
      mix $\leftarrow$ **Función GMM (dimensión, mezclas, covar_type)**
      {Inicialización de los parámetros para aplicar el algoritmo Máxima Expectancia}
      Opción $\leftarrow$ **ceros (1, 18)** **{arreglo de ceros}**
      Opción (14) $\leftarrow$ **200** **{Máximo número de iteraciones}**
      Opción (3) $\leftarrow$ **0.001** **{Error mínimo}**
      [Mix, opción, errlog] $\leftarrow$ **Función EM (mix, X, opción)**
      Modelos Locutor (n) $\leftarrow$ **Mix**
    Fin hasta
    Modelos Locutores(s) $\leftarrow$ **Modelos Locutor**
  Fin hasta
Fin

Para iniciar la fase de *identificación* se toman como datos iniciales los modelos obtenidos en el entrenamiento, donde cada locutor presenta un grupo de modelos; y la señal $Y_{Z,T}$ que vamos a identificar. A la hora de calcular la probabilidad nos apoyamos en la ecuación 2.9 y le hallamos la probabilidad a todos los modelos del locutor quedándonos con la mayor probabilidad por locutor. Como se observa en el siguiente algoritmo:

Algoritmo Identificación de los locutor a partir de los modelos de la señal segmentada de cada uno

Entrada: $\left(\overline{\lambda}_1, \overline{\lambda}_2, \ldots, \overline{\lambda}_S\right)$ **donde** $\overline{\lambda}_S$ **es un conjunto del locutor S y** $Y_{Z,T}$ **señal a identificar**
LLH: entero cualquiera
Salida: Probabilidad

```
Inicio
    Muestra ← X_{Z,T}
    Desde s=1 hasta S
        LLH ← El menor valor que nunca tomaría
        Desde n =1 hasta longitud (λ_(s))
                Mix ← λ̄_{s,n}
                Probabilidades ← Función Calcular probabilidades (Mix | Muestra)
                Probabilidad ← mean (Probabilidades) {Calcular media}
                Si LLH < Probabilidad
                    LLH ← Probabilidad
                Fin Si
        Fin hasta
        Log Likelihood (s) ← exp (LLH) {Se calcula el exponencial}
    Fin hasta
    Se ordenan los Log Likelihood descendentemente
Fin
```

El resultado de la identificación nos brindará una lista ordenada descendentemente de los locutores según el valor de probabilidad de cada cual.

CAPÍTULO 3: RESULTADOS Y APORTES

Este capítulo presenta como objetivo fundamental exponer una serie de experimentos y sus resultados durante la realización de este trabajo. Para medir el error del sistema se utilizó la tasa del EER y se tomó como datos para el desarrollo de los mismos, ficheros de diversos canales de la base de datos, Ahumada.

3.1 Base de Datos Ahumada

La información utilizada fue tomada de la base de datos Ahumada, la cual presenta señales tomadas por varios canales y con presencia de voces espontáneas, con el objetivo de trabajar en función a la variabilidad del canal y el estado del locutor. De esta se tomaron señales de distintos canales como son “T1” y “T2”, los cuales presentan señales pertenecientes a 100 locutores, en nuestro caso solo nos centramos en vías telefónicas. Las bases utilizadas presentan una serie de características:

- "T1".- Grabación telefónica, 1ª sesión.
- "T2".- Grabación telefónica, 2ª sesión.

Estos ficheros presentan un número de características técnicas dentro de las cuales, “T1”, “T2” presentan una frecuencia de muestreo de 8.000 Hz, que se corresponden a las sesiones donde ha estado implicado el teléfono.

3.2 Experimento y Resultados utilizando una nueva representación del locutor

Para la realización del experimento se tomó de la base de datos “Ahumada” al canal “T1”, para el entrenamiento; el cual presenta 100 matrices de rasgos pertenecientes a locutores distintos. Se *entrenó* mediante el algoritmo desarrollado ‘Clasificación utilizando una nueva representación del locutor’ con los siguientes parámetros para la creación del modelo Gaussiano:

- Número de mezclas: 16
- Con condición de parada por iteraciones: 200

- Con condición de parada por convergencia: $p\left(X \mid \overline{\lambda}\right) - p\left(X \mid \lambda\right) \leq 0.001$

Del cual se obtuvo un arreglo de modelos de la forma $\left(\overline{\lambda}_1, \overline{\lambda}_2, \ldots, \overline{\lambda}_{100}\right)$ donde cada $\overline{\lambda}$ pertenece a un locutor diferente y esta formado por una matriz de media (μ), una matriz de covarianza $(\sum)$, un vector de peso $(\vec{p})$ y una matriz conformada por los vectores propios pertenecientes al locutor (vp).

Para *identificar* se utilizaron los resultados de la fase de clasificación como atributos de entrada $\left(\overline{\lambda}_1, \overline{\lambda}_2, \ldots, \overline{\lambda}_{100}\right)$; para el algoritmo 'Identificación utilizando una nueva representación' además del fichero ("T2") como señales a identificar. Con los vectores propios (vp) incluidos en el modelo a la hora de identificar ya tenemos los datos necesarios para poder representar la Test ("T2") en el espacio propio correspondiente, teniendo en cuenta que este experimento es controlado dado que se tiene el conocimiento a que locutor pertenece la señal a identificar y por tanto los vectores que le corresponden. Se representó la señal en el espacio propio correspondiente apoyándonos en la ecuación 2.3, se calculó la probabilidad de los rasgos acústicos dado el modelo y se ordenaron descendentemente los locutores en función de a probabilidad.

Para medir el error del sistema se utilizó la tasa del EER y se comparó con una línea base clasificada mediante las GMM, tomando como datos para el entrenamiento a T1 y como Test a T2 con los siguientes parámetros:

- Número de mezclas: 64
- Con condición de parada por iteraciones: 200
- Con condición de parada por convergencia: $p\left(X \mid \overline{\lambda}\right) - p\left(X \mid \lambda\right) \leq 0.001$

Obteniendo los resultados mostrados en la figura 3.1.

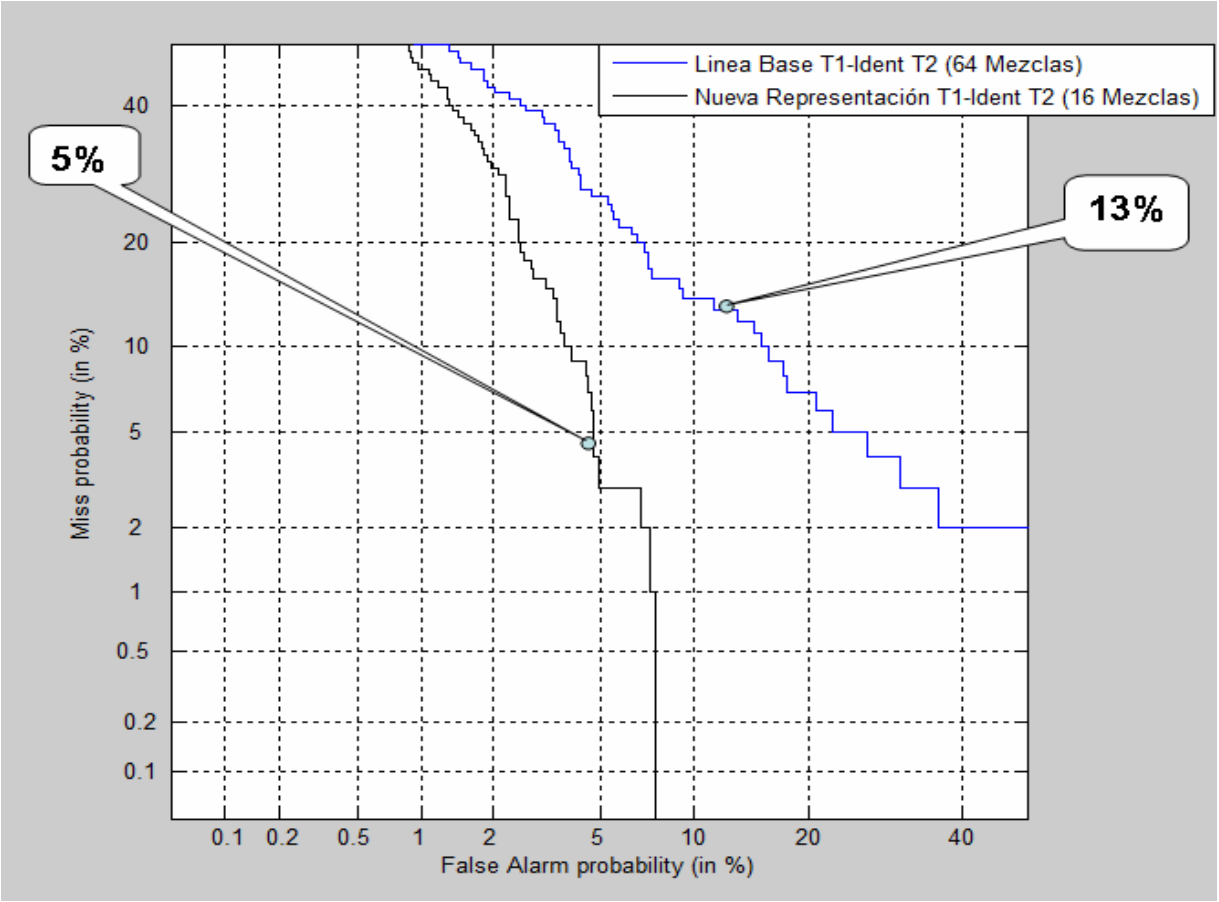

Fig. 3.1 Línea Base con 64 mezclas "azul" y experimento con 16 números diferentes de componentes de densidad Gaussiana "negro".

La línea azul es la línea base que presenta un 13% de probabilidad en fallos y la roja es el experimento controlado que se desarrolló utilizando la técnica de reducción de dimensionalidad para representar al locutor en un nuevo espacio, dando como resultado un 5% de probabilidad en fallos.

Con este experimento quedó demostrado que la nueva representación de los rasgos del locutor mejoran ampliamente los resultados obtenidos hasta el momento, teniendo en cuenta que con menor número de componentes de densidad gaussiana (16) existe una gran disminución del error del sistema en base al EER, por lo que se puede afirmar: que conociendo el espacio propio en el cual se representará el Test, se marcaría una pauta en el desarrollo de sistemas de reconocimiento automático del locutor independiente del texto.

3.3 Experimento y Resultados utilizando la técnica de Segmentar la Señal

Para desarrollar este experimento se escogió como rasgos iniciales en el entrenamiento la sesión T1 de la base de datos Ahumada, la cual presenta matrices de rasgos pertenecientes a 100 locutores, cada matriz de rasgos equivale a un minuto de señal acústica; para la ejecución del algoritmo denominado 'Clasificación utilizando la señal segmentada', con el cual se desea explotar la información implícita en la trayectoria de la voz, utilizando la señal dividida en segmentos. Se entrena cada segmento obtenido de la matriz de rasgos, mediante las GMM y teniendo como parámetros para el modelo a:

- Número de mezclas: 16
- Con condición de parada por iteraciones: 200
- Con condición de parada por convergencia: $p\left(X \mid \overline{\lambda}\right) - p\left(X \mid \lambda\right) \leq 0.001$

Del entrenamiento se obtuvo un conjunto de modelos por cada locutor de la forma $\left(\lambda_1, \lambda_2, \ldots, \lambda_N\right)$, donde n es el número de modelos que van a caracterizar la información en la trayectoria de la señal del locutor. Se tomaron los resultados del entrenamiento y las señales existentes en el fichero ("T2"), que de igual forma contienen la misma cantidad de matrices de rasgos que caracterizan a los locutores; como datos iniciales para el algoritmo 'Identificación de los locutores a partir de los modelos de la señal segmentada de cada uno'. Dicho método calcula la probabilidad de los conjuntos de modelos ante la secuencia de T2, escogiendo por cada locutor la mayor probabilidad del conjunto de modelos que caracterice la señal, luego estas probabilidades son ordenadas descendentemente. Obteniéndose como resultado una matriz de probabilidades de dimensión 100x100, dado que se calcula la probabilidad de todos los modelos por cada señal existente en T2.

Para medir el error del sistema se tomó la tasa del EER y se comparó el experimento contra la línea base GMM, entrenada con T1, test T2 y como parámetros de entrada:

- Número de mezclas: 64
- Con condición de parada por iteraciones: 200

- Con condición de parada por convergencia: $p\left(X \mid \overline{\lambda}\right) - p\left(X \mid \lambda\right) \leq 0.001$

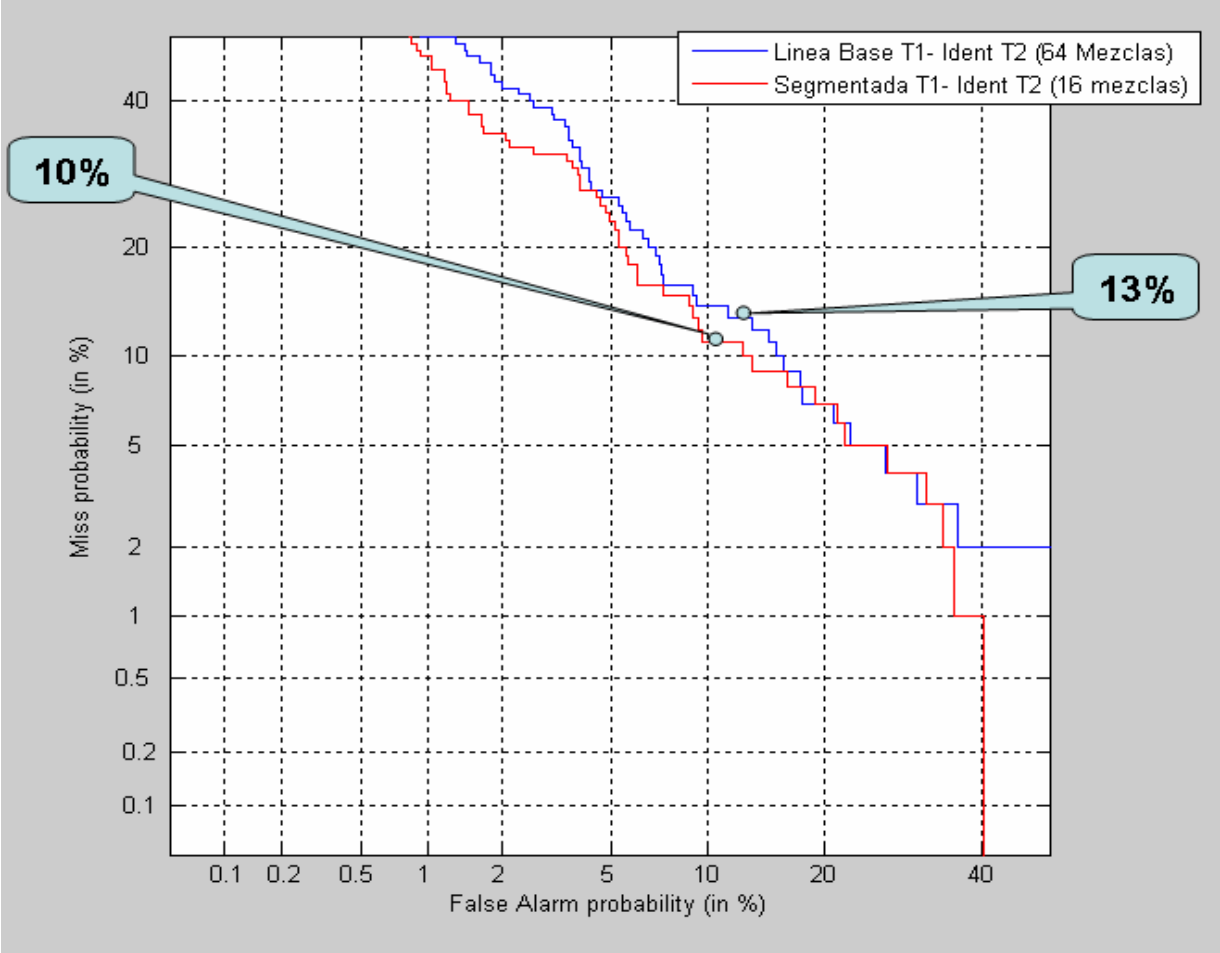

Fig. 3.2 Línea Base con 64 mezclas "azul" y el experimento diseñado con 16 componentes de densidad Gaussiana, "rojo".

La figura 3.2 nos muestra los resultados obtenidos con la línea base en azul y el experimento utilizando la técnica de segmentado de la señal en rojo. Se observa que el experimento realizado obtiene un error menor que la línea base. Lo que demuestra que la explotación de la información dinámica de la voz mejoraría los resultados obtenidos en los sistemas RAL, ya que utilizando solo 20 segundos de la trayectoria de una señal de un minuto y menor número de componentes Gaussianos se obtienen mejores resultados. Este experimento demuestra que es posible disminuir el costo computacional utilizando la dinámica de señal.

CONCLUSIONES

El desarrollo de los algoritmos utilizando diferentes enfoques en la trayectoria del modelo como son: la utilización de una nueva representación del locutor y la segmentación de la señal, con el objetivo de robustecer los sistemas de RAL, lograron una mejora de un 60%, en un experimento controlado; y 24% en el no controlado, comparados con los resultados obtenido a partir de la línea base escogida, GMM clásicas; y apoyándonos en la tasa del EER para medir el error del sistema, como se muestra en la siguiente tabla.

Algoritmos	Error EER	Por ciento de mejoras
Línea Base GMM T1-T2---64 Mezclas	13%	-------
Exp. Nueva representación T1-T2---16 Mezclas	5%	60%
Exp. Segmentando la señal T1-T2---16 Mezclas	10%	24%

Teniendo en cuenta los resultados obtenidos, podemos plantear, que con solo la utilización de 20 segundos de la señal de un minuto, se obtiene un mejor resultado y entrenando con menor número de componentes de densidad gaussiana, también se diminuye el costo computacional empleado, un aspecto de gran importancia en la rama del reconocimiento del locutor.

Tomando como trayectoria para la creación de los modelos, la representación del locutor en su espacio propio; trae consigo un significante mejoramiento de eficiencia de los sistemas RAL, no obstante conociendo como proyectar una señal desconocida en su espacio propio obtendríamos una pauta a nivel internacional en el tema de reconocimiento del locutor.

RECOMENDACIONES

A partir de los resultados obtenidos en el trabajo proponemos:

- Continuar con la profundización del estudio de las representaciones del locutor, con el objetivo de lograr encontrar el espacio propio a representar una señal desconocida.
- Continuar con el análisis de la dinámica de la señal de voz utilizando segmentos de corto tiempo, para encontrar más información ya que se demostró que con 20 segundos de una señal de un minuto se obtuvieron mejores resultados.
- Representar todas las medias de los modelos de la segmentación en un espacio para buscar el centro de cada grupo de medias, obteniendo una media generalizada y proyectándola en el espacio del locutor.
- Llevar los algoritmos al lenguaje C++, para una futura utilización en algún sistema de reconocimiento del locutor.

REFERENCIAS BIBLIOGRÁFICAS

1. Furui S. An overview of Speaker Recognition Technology, ESCA Workshop on Automatic Speaker Recognition, 1994, pp. 1-10.
2. Douglas A. Reynolds y Richard C. Rose. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Transactions on speech and audio processing, 1995, Vol. 3 (No. 1).
3. P. Moreno y S. Agarwal. An experimental study of em-based algorithms for semi-supervised learning in audio classification, Proceedings of ICML, 2003.
4. Doddington et al. The NIST speaker recognition evaluation: overview, methodology, systems, results, perspective. Speech Communication, 2000, Vol. 31(54), pp.225-254.
5. Ortega J. Técnicas de mejora de voz aplicadas a sistemas de reconocimiento de locutores. Tesis doctoral, ETSI. Telecomunicación, Universidad Politécnica de Madrid, 1996.
6. Ortega J. y González J. Estudio comparativo de técnicas de identificación automática de locutores. X Simposio nacional de la Unión Científica Internacional de Radio URSI, Valladolid, 1995.
7. Vivaracho C. E., Ortega J. y Romero L. A., Perceptrón Multicapa frente a Modelos de Mezcla de Gaussianas en verificación automática de locutores. Actas del I Congreso de la SEAF, 2000, pp.85-90.
8. Kinnunen T., Spectral Features for Automatic Text-Independent Speaker Recognition. Department of Computer Science, University of Joensuu, Finland, 2003.
9. Campbell J.P. Speaker Recognition: A Tutorial, Proceedings of the IEEE, 1997, Vol. 85 (No. 9): pp. 1437- 1462.
10. Gauvain J. L. y Lee C. H., Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. 1994, 2, pp. 291-298.
11. Solomonoff A., W. M. Campbell, and I. Boardman, “Advances in channel compensation for SVM speaker recognition,” in Proc. ICASSP, 2005, pp. 629–632.

12. P. Kenny y P. Demouchel, “Eigenvoice modeling with sparse training data,” IEEE Trans. Speech Audio Process., vol. 13, no. 3, pp. 345–354, May 2005.
13. Campbell W. M., Sturim D., Reynolds D., y Solomonoff A., SVM based speaker verification using a GMM Supervector kernel and NAP variability compensation. In Proc. ICASSP, 2006, Vol. 1.
14. Ke Chen. On the Dynamic Pattern Analysis, Discovery and Recognition, IEEE Systems, Man & Cybernetics Society E-Newsletter, 2005, Issue 12.
15. Ke Chen. On the Use of Different Speech Representations for Speaker Modeling, IEEE Transactions on Systems, Man, and Cybernetics-Part C: Aplications and Reviews, 2005, Vol. 35 (No. 35).
16. Bing Xiang. SPEAKER VERIFICATION USING GAUSSIAN COMPONENT STRINGS IN DYNAMIC TRAJECTORY SPACE. ISCA Archive, Denver, Colorado, USA, 2002, pp 16-20.
17. Errity A. y McKenna J., An Investigation of Manifold Learning for Speech Analysis. INTERSPEECH, 2006.
18. Aren Jansen, The Manifold Nature of Vowel Sounds, Master’s Paper, Dept. of Computer Science, Univ. of Chicago, 2007.
19. Aren Jansen y Partha Niyogi, A Geometric Perspective on Speech Sounds. Report TR-2005-08. Computer Science Dept., Univ. of Chicago. 2005.
20. Aren Jansen y Partha Niyogi. INTRINSIC FOURIER ANALYSIS ON THE MANIFOLD OF SPEECH SOUNDS, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 2006.
21. L. K. Saul y S. T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds, Journal of Machine Learning Research, 2003, vol. 4, pp. 119–155.
22. J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction, Science, 2000, vol. 290, pp. 2319–2323.
23. T. Cox y M. Cox, Multidimensional Scaling. Chapman and Hall, Boca Raton, 2001.

24. Mikhail Belkin y Partha Niyogi. Laplacian Eigenmaps and spectral techniques for embedding and clustering, in Advances in Neural Information Processing Systems, 2002, vol. 14, pp. 585–591.
25. Mikhail Belkin y Partha Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Technical Report TR-2002-01, 2001.
26. Mora F. W. Matrices con entradas enteras e inversa con entradas enteras. Revista Virtual Matemática, Educación e Internet, 2004.

BIBLIOGRAFÍA

1. Aren Jansen, The Manifold Nature of Vowel Sounds, Master's Paper, Dept. of Computer Science, Univ. of Chicago, 2007.
2. Aren Jansen y Partha Niyogi, A Geometric Perspective on Speech Sounds. Report TR-2005-08. Computer Science Dept., Univ. of Chicago. 2005.
3. Aren Jansen y Partha Niyogi. INTRINSIC FOURIER ANALYSIS ON THE MANIFOLD OF SPEECH SOUNDS, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 2006.
4. Bing Xiang. SPEAKER VERIFICATION USING GAUSSIAN COMPONENT STRINGS IN DYNAMIC TRAJECTORY SPACE. ISCA Archive, Denver, Colorado, USA, 2002, pp 16-20.
5. Bennani Y. y Gallinari P., Neural networks for discrimination and modelization of speakers. Speech Communication 17, 1995, pp. 159-175.
6. Campbell J.P. Speaker Recognition: A Tutorial, Proceedings of the IEEE, 1997, Vol. 85 (No. 9): pp. 1437- 1462.
7. Cole R., Mariani J., Uszkoreit H., Batista G., Zaenen A., Zampolli A. y Zue V. Survey of the State of the Art in Human Language Technology. Cambridge University Press and Giardini, 1997.
8. Campbell W. M., Sturim D. y Reynolds D. A., Support vector machines using GMM supervectores for speaker verification. IEEE Signal Process, 2006, Vol. 13 (No. 5): pp. 308-311.
9. Campbell W. M., Sturim D., Reynolds D., y Solomonoff A., SVM based speaker verification using a GMM Supervector kernel and NAP variability compensation. In Proc. ICASSP, 2006, Vol. 1.
10. Campbell W., Campbell J., Reynolds D., Singer E., y Torres Carrasquillo P. Support

vector machines for speaker and language recognition. Compute. Speech Lang, 2006, Vol. 20, pp. 210-229.

11. Douglas A. Reynolds y Richard C. Rose. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, 1995, VOL. 3 (NO. 1).
12. Doddington et al. The NIST speaker recognition evaluation: overview, methodology, systems, results, perspective. Speech Communication, Vol. 31, pp.225-254, 2000. 54.
13. Deller J. R., Proakis J. G., y Hansen J. H. Discrete-Time Processing of Speech Signals, Prentice Hall, 1993.
14. Errity A. y McKenna J., An Investigation of Manifold Learning for Speech Analysis. INTERSPEECH, 2006.
15. Furui S. An overview of Speaker Recognition Technology, ESCA Workshop on Automatic Speaker Recognition, 1994, pp. 1-10.
16. Gauvain J. L. y Lee C. H., Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. 1994, 2, pp. 291-298.
17. Huang X., Acero A. y Hon H.-W. Spoken Language Processing: a Guide to Theory, Algorithm, and System Development, Prentice-Hall, 2001.
18. Javier Ortega-Garcia, Joaquin Gonzalez-Rodriguez y Victoria Marrero Aguiar, AHUMADA: A large speech corpus in Spanish for speaker characterization and identification, in Speech Communication, 2000, pp. 255-264.
19. J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction, Science, 2000, vol. 290, pp. 2319–2323.
20. Ke Chen. On the Dynamic Pattern Analysis, Discovery and Recognition, IEEE Systems, Man & Cybernetics Society E-Newsletter, 2005, Issue 12.

21. Ke Chen. On the Use of Different Speech Representations for Speaker Modeling, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, 2005, VOL. 35 (NO. 3).

22. Kinnunen T., Spectral Features for Automatic Text-Independent Speaker Recognition. Department of Computer Science, University of Joensuu, Finland, 2003.

23. Lee J. M., Introduction to Topological Manifolds, Springer, 2000.

24. L. K. Saul y S. T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds, Journal of Machine Learning Research, 2003, vol. 4, pp. 119–155.

25. Mikhail Belkin y Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering, in Advances in Neural Information Processing Systems, 2002, vol. 14, pp. 585–591.

26. Mikhail Belkin y Partha Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. Technical Report TR-2002-01, 2001.

27. Morgan D.B. y Scofield C.L. Neural Networks and Speech Processing, Kluwer Academic Publishers, 1991.

28. Mora F. W. Matrices con entradas enteras e inversa con entradas enteras. Revista Virtual Matemática, Educación e Internet, 2004, 5(2).

29. Ortega J. Técnicas de mejora de voz aplicadas a sistemas de reconocimiento de locutores. Tesis doctoral, ETSI. Telecomunicación, Universidad Politécnica de Madrid, 1996.

30. Ortega J. y González J. Estudio comparativo de técnicas de identificación automática de locutores. X Simposio nacional de la Unión Científica Internacional de Radio URSI, Valladolid, 1995.

31. P. Kenny y P. Demouchel, “Eigenvoice modeling with sparse training data,” IEEE Trans.

Speech Audio Process., vol. 13, no. 3, pp. 345–354, May 2005.

32. P. Moreno y S. Agarwal. An experimental study of em-based algorithms for semi-supervised learning in audio classification, Proceedings of ICML, 2003.
33. Rabiner L., A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of IEEE, 1989.
34. Rabiner L. y Juang B. H. Fundamentals of Speech Recognition, Prentice Hall, 1993.
35. Solomonoff A., W. M. Campbell, and I. Boardman, “Advances in channel compensation for SVM speaker recognition,” in Proc. ICASSP, 2005, pp. 629–632.
36. S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding, Science, 2000, vol. 290 (no. 5500), pp. 2323 – 2326.
37. T. Cox y M. Cox, Multidimensional Scaling. Chapman and Hall, Boca Raton, 2001.
38. Vivaracho C. E., Ortega J. y Romero L. A., Perceptrón Multicapa frente a Modelos de Mezcla de Gaussianas en verificación automática de locutores. Actas del I Congreso de la SEAF, 2000, pp.85-90.

Anexos: 1

Conceptos y definiciones matemáticas

Continuidad y Convergencia en el espacio Euclidiano: Una función $f : X \to Y$ entre dos subconjuntos del espacio Euclidiano es continua si y solo si para todo $x \in X$ y para cada $\varepsilon > 0$ existe un $\delta > 0$ tal que $|x - y| < \delta \Rightarrow |f(x) - f(y)| < \varepsilon$. Una sucesión $\{x_i\}$ de puntos en $\Re^n$ converge a $x \in \Re^n$ si para un $\varepsilon > 0$ existe un N tal que para los $i \geq N$ implica que $|x_i - x < \varepsilon|$.

Bola abierta: Dado un punto $a \in \Re^n$ y un número real r > 0, llamamos bola abierta con centro a y radio r al conjunto de $B(a,r) = \{x \in \Re^n \mid d(a,x) < r\}$. Dado un espacio métrico M, se dice que un conjunto $A \subset M$ es abierto si contiene bolas abiertas alrededor de todos sus puntos y la unión de todas las bolas es el conjunto A.

Criterio de conjunto abierto para continuidad: Un mapa $f : X \to Y$ entre espacios métricos es continuo si y solo si la imagen inversa de todo conjunto abierto es un abierto, o sea, cualquiera sea U subconjunto abierto en Y, entonces la inversa f^{-1} (U) es un abierto en X.

Continuidad en espacios métricos: La continuidad en un punto $x \in M_1$ de un mapa $f : M_1 \to M_2$ en un espacio métrico esta definida de la siguiente forma: si $\forall \varepsilon > 0$ existe $\delta > 0$ tal que para cada $y \in M_1$, $d_1(y,x) < \delta$ entonces $d_2(f(y), f(x)) < \varepsilon$, entonces un mapa es continuo si y solo si es continuo en todos los puntos del conjunto.

En los espacios topológicos la continuidad en un punto no es generalmente usada, dado que estos espacios no son continuos de forma general, sin embargo es muy importante el hecho de que tiene un comportamiento continuo de forma local, por lo tanto un mapa es continuo si y solo si es continuo en la vecindad de cada punto del espacio

Criterio local para continuidad: Un mapa $f : X \to Y$ entre espacios topológicos es continuo si y solo si en cada punto de X tiene una vecindad en la cual f es continua.

Espacio topológico: Una topología en un conjunto X es una colección T de subconjuntos de X, llamados conjuntos abiertos, que satisfacen las siguientes propiedades:

I. X y Φ son elementos de T.

II. $(O_1 \in T, O_2 \in T) \Rightarrow (O_1 \cap O_2 \in T)$ (es cerrado bajo intersecciones finitas)

III. $\forall i \in A, O_i \in T \Rightarrow (\bigcup_{i \in A} O_i \in T)$ (es cerrado bajo uniones arbitrarias)

Entonces el par (X, T) consiste en un conjunto X de puntos, provisto de una topología T y se le

llama espacio topológico.

Topologías equivalentes: Dos espacios topológicos, X y Y son topológicamente equivalentes si existe un mapa $f : X \to Y$ tal que:

I. El mapa $f : X \to Y$ sea biyectiva.

II. El mapa $f : X \to Y$ sea continua.

III. Exista la inversa $f^{-1} : Y \to X$ y sea continua.

Se dice entonces que f es un homeomorfismo entre X y Y o que X y Y tienen el mismo tipo de topología y se escribe como $X \approx Y$.

Definición de variedad: Un espacio topológico M se dice localmente Euclidiano de dimensión n si para todo punto $q \in M$ tiene una vecindad que es homeomorfica a un subconjunto abierto de $\Re^n$. Estas vecindades son llamadas vecindades Euclidianas de q.

Lema 1: Un espacio topológico M es localmente Euclidiano de dimensión n si y solo si tiene cualquiera de las siguientes propiedades:

I. Todo punto de M tiene una vecindad homeomorfica a una bola abierta en $\Re^n$.

II. Todo punto de M tiene una vecindad homeomorfica a $\Re^n$.

Si un espacio topológico M es localmente Euclidiano de dimensión n, un homeomorfismo de un subconjunto abierto $U \subset M$, a un subconjunto abierto de $\Re^n$, es llamado mapa o carta en U.

Entorno abierto: En un espacio topológico, un entorno abierto de un punto es un conjunto abierto que contiene al conjunto abierto que contiene al punto. Es decir si (X, T) es un espacio topológico, $a \in X$ y $U \in X$, diremos que V es un entorno de a si existe una bola abierta V, de forma que $a \in V \subset U$.

Espacios de Hausdorff: Dados dos puntos de un espacio topológico X, $q_1, q_2 \in X$ se dice que ambos puntos gozan de la propiedad de Hausdorff si existen dos entornos U_1 de q_1 y U_2 de q_2 tales que su $U_1 \bigcap U_2 = \phi$.

Un espacio topológico se dice que es un espacio de Hausdorff, si todo par de puntos del espacio verifican la propiedad de Hausdorff.

Algunas propiedades de los espacios de Hausdorff se expresan en el siguiente lema.

Lema 2: Dado X un espacio de Hausdorff.

Todo conjunto de un punto en X es cerrado.

Si una sucesión $\{x_i\}$ en X converge a un límite $x \in X$, el límite es único.
Variedad Topológica: Una Variedad Topológica de dimensión n, es un espacio topológico de Hausdorff en el que todo punto tiene un entorno abierto homeomorfico a una bola abierta en $\Re^n$.

GLOSARIO

1. Patrones biométricos: Características del cuerpo humano utilizadas para el reconocimiento de la persona.
2. Locutor: Un hablante cualquiera.
3. RAL: Reconocimiento Automático del Locutor.
4. Coeficientes Mel: Variables aleatorias indexadas por una variable discreta, el tiempo $(t = 1, \ldots, T)$ y distorsionado en frecuencia con escala Mel.
5. VAL: Verificación automática del locutor.
6. IAL: Identificación automática del locutor.
7. CENATAV: Centro de Aplicaciones de Tecnologías de Avanzada.
8. Matlab: Herramienta matemática utilizada para procesamiento de datos.
9. VQ: Algoritmo Cuantización Vectorial utilizado para la creación de los modelos.
10. DTW: Algoritmo Distorsión Dinámica en el Tiempo utilizado también en la creación de los modelos.
11. HMM: Algoritmo Modelos Ocultos de Markov se usa para la creación de los modelos.
12. GMM: Algoritmo Modelos de Mezclas Gaussianas utilizado en el entrenamiento para la obtención de los modelos.
13. UBM: Modelos Universales de Background.
14. MAP: Adaptación máximo a posteriori.
15. GSL: GMM Supervector Lineal Kernel.
16. EER: Equal Error Rate. Índice erróneo equivalentes.
17. Entropía: Medida del grado de desorden o de desconocimiento de un sistema.
18. Topología: Ciencia que estudia de las propiedades de los cuerpos geométricos que permanecen inalteradas por transformaciones continuas.
19. Sesión: Estado del Locutor.
20. Trama: Un vector de la matriz de rasgos.
21. Test: Señal del locutor que se desea identificar.

22. Homeomorfismo: Un homeomorfismo, en topología, es un isomorfismo entre dos espacios topológicos: una función matemática continua de uno en otro, cuyo recíproco es continuo. En este caso, los dos espacios topológicos se llaman homeomorfos.