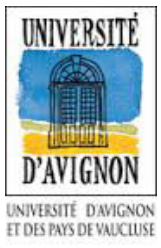
Métodos de representación y verificación del locutor con independencia del texto

Gabriel Hernandez Sierra

Centro de Aplicaciones de Tecnologías de Avanzada
Departamento de Imágenes y Señales



Universidad de Avignon

Laboratorio de Informática de Avignon

Métodos de representación y verificación del locutor con independencia del texto

---

Tesis presentada en opción al grado científico de
Doctor en Ciencias Técnicas

Autor:

Lic. Gabriel Hernández Sierra

Tutores:

Dr. Jean-François Bonastre (Universidad de Avignon, Francia)
Dr. José R. Calvo de Lara (CENATAV, Cuba)

Instituto Superior Politécnico José Antonio Echeverría
Ciudad de La Habana
2014

A mi mamá, a mi papá, a mis hijos y a mi esposa, los pilares de esta travesía

Agradecimientos

A todos los que contribuyeron a que el largo camino fuera más corto:

A mi tutor José Ramón Calvo, por todo su aporte a mi formación profesional. Por su esfuerzo y dedicación, por convertirse en un amigo.

A Bonastre, por su paciencia y comprensión. Por sus importantes aportes al desarrollo de este trabajo y la confianza que depositó en mí.

A Pierre-Michel, por su colaboración y deseos de ayudar. Por el tiempo que pasamos frente a una pizarra comprendiendo y proponiendo nuevas ideas.

A Juanma y Pati por ofrecerme el calor del hogar que nunca olvidaré y tratarme como a un hijo.

A Eitan y Tania por el trato especial que me ofrecieron y el cariño que me dieron.

A Julio y Sulan que compartían mi música cubana y hacían mi estadía más alegre.

A los amigos de Francia que me acogieron y me brindaron su apoyo incondicional estando lejos de la tierra. Los amigos del LIA, que hicieron más placentera mi estancia. A Hugo y Enki por acogerme como un hermano.

A los profesores Julian y Carlos por sus observaciones y sugerencias en la predefensa de la tesis, que ayudaron a mejorar la calidad la misma.

A mis padres por el apoyo, la confianza y el amor que siempre han depositado en mí, por ser mis guías en la vida.

A mi esposa, por ser la sangre que le da oxígeno a mi cuerpo, a mi chacha por ser responsable hacer que la ame todos los días más

A mis hijos por llenarme de alegrías, entregarme su cariño y hacerme la vida más feliz.

A mi equipo de trabajo por soportarme diariamente, gracias por la colaboración de todos: Ana, Dayana, Claudia, Flavio y Serguei.

Al Gare por siempre estar dispuesto a resolver los problemas que fueron surgiendo.

Al Artur por ofrecerme su amistad y conocimiento de idioma.

A Heydi, Ricardo, Chang, Edel, Dina, Noslen, Javier, Palancar, Cuba, Medina, mis compañeros del antiguo departamento que vivieron el despegue de este proyecto.

A mi coterráneo Shul que me apoyó y me brindó la oportunidad de convertir este sueño realidad.

A Isneris por su ejemplo y espíritu, haciendo posible lo imposible.

A los choferes, las pantristas, los compañeros de la administración, a todo el CENATAV que por una vía u otra han aportado su granito de arena para el éxito de este trabajo.

A toda mi familia que siempre ha estado preocupada y atenta a los sucesos que han transcurrido.

A la familia de Santa Fé por acogerme, quererme y estar pendientes de mí.

A mis incondicionales amigos de Madruga por estar a mi lado y alimentarme de su amistad: Flavio, Amed, Yoel, Javier y Beraldo.

A todos aquellos que puedo estar olvidando ahora, pero que me han guiado y apoyado en el transcurso de este andar.

A la Revolución cubana, por ser la madre de mis sueños e inculcarme los principios que guían mi vida.

Gabriel Hernández Sierra.

SÍNTESIS

El reconocimiento automático del locutor independiente del texto, es un método de reciente incorporación en los sistemas biométricos. El desarrollo y auge del mismo se refleja en las competencias internacionales, pero aun la eficacia de los métodos de reconocimiento se encuentra afectada por la cantidad de información discriminatoria del locutor que esta presente en las representaciones actuales de las expresiones de voz. En esta tesis se realizó un estudio donde se identificaron dos principales debilidades presentes en las representaciones actuales del locutor. En primer lugar, no se tiene en cuenta el comportamiento temporal de la voz, siendo este un rasgo discriminatorio del locutor y en segundo lugar los eventos pocos frecuentes dentro de una población de locutores pero frecuentes en un locutor dado, apenas son tenidos en cuenta por estos enfoques, lo cual es contradictorio cuando el objetivo es discriminar los locutores. Motivado por la solución de estos problemas, se confirmó la redundancia de información existente en las representaciones actuales y la necesidad de emplear nuevas representaciones de las expresiones de voz. Se propuso un nuevo enfoque con el desarrollo de un método para la obtención de un modelo generador capaz de transformar la representación actual del espacio acústico a una representación en un espacio binario, donde se propuso una medida de similitud asociada con una representación global (vector acumulativo) que contiene tanto los eventos frecuentes como los pocos frecuentes en una expresión de voz. Para la compensación de la variabilidad de sesión se incorporó en la matriz de dispersión intra-clase, la información común de la población de locutores, lo que implicó la modificación de tres algoritmos de la literatura que mejoraron su desempeño respecto a la eficacia en el reconocimiento del locutor, tanto utilizando el nuevo enfoque propuesto como el enfoque actual de referencia. La información temporal existente en las expresiones de voz fue capturada e incorporada en una nueva representación, mejorando aun más la eficacia del enfoque propuesto. Finalmente se propuso y evaluó una fusión lineal entre los dos enfoques que demostró la información complementaria existente entre ellos, obteniéndose los mejores resultados de eficacia en el reconocimiento del locutor.

ÍNDICE

INTRODUCCIÓN

La biometría (del griego bios vida y metron medida) consiste en desarrollar métodos automatizados que analizan determinadas características humanas para identificar o autentificar personas, usando técnicas de reconocimiento de patrones. Estas características son difíciles de perder, transferir u olvidar y pueden ser físicas o de comportamiento. Las huellas dactilares, la retina, el iris, los patrones faciales o la geometría de la palma de la mano, representan ejemplos de características físicas, mientras que las del comportamiento pueden incluir la firma, la forma de caminar y la forma de teclear [43]. Existe además la voz, que se considera una mezcla de características físicas y del comportamiento.

Los sistemas biométricos aportan una solución efectiva al problema de la identificación. No obstante, cada característica biométrica es diferente, y no siempre todas son adecuadas en las distintas aplicaciones. Al diseñar un sistema biométrico se evalúan diversos parámetros, como el poder distintivo de la característica biométrica que se emplee y la facilidad que se tenga para su uso, entre otros, que determinan su utilidad en las aplicaciones reales. La Tabla 1, cuyos datos fueron extraídos del Manual de Biometría [43] en el año 2007, muestra el comportamiento de algunos de los sistemas basados en diferentes características biométricas.

Tabla 1: Tabla comparativa de sistemas biométricos

	Huella Dactilar	Rostro	Iris	Geometría de la Mano	*VOZ*
Capacidad distintiva	Muy Alto	Alto	Muy Alto	Alto	*Medio*
Facilidad de uso	Alto	Alto	Bajo	Medio	*Alto*
Seguridad ante ataques	Media	Media	Muy Alta	Alta	*Media*
Aceptación	Media	Muy alta	Bajo	Medio	*Muy alta*
Estabilidad de los datos	Alta	Media	Alta	Alta	*Media*

Como se muestra en la tabla los sistemas basados en la voz presentan *muy alta* aceptación, provocado porque el lenguaje hablado es, sin duda, el método de comuni-

cación más natural, intuitivo y eficiente para los seres humanos. Además el intercambio de información mediante el habla juega un papel fundamental en nuestras vidas, provocando que las estructuras lingüísticas y acústicas de la voz sean reconocidas desde antaño como estrechamente relacionadas con nuestra capacidad intelectual y de comunicación social. Por ello, no es de extrañar que durante décadas la idea de interaccionar oralmente con máquinas como si de personas se tratase ha fascinado a ingenieros y científicos.

El reconocimiento automático de una persona por su voz, o reconocimiento automático del locutor, es actualmente un área de investigación y desarrollo de aplicaciones de gran importancia.

Las primeras aproximaciones documentadas del reconocimiento automático del locutor se producen a principios de los años setenta. Los ingenieros de los laboratorios Bell [33] y Aaron Rosenberg [77] publican sus primeros estudios utilizando ya, como base de extracción para los rasgos, coeficientes cepstrales y coeficientes de predicción linear [77]. En esta misma época, son probados diversos parámetros acústicos del habla para su utilización en sistemas de reconocimiento automático: en 1972 Wolf [87] analiza combinaciones de hasta veintisiete referencias extraídas de consonantes nasales, espectros de vocales, frecuencia fundamental. Su y Fu en 1973 [83] utilizan como informaciones eficientes los espectros de consonantes nasales. Li y Hughes en 1974 [58] toman como referencia matrices de correlación referidas a fragmentos de habla continua. La mayoría de estos primeros intentos estaban estrechamente ligados a comparaciones dependientes del texto (expresiones de habla con textos similares), aunque igualmente se reportan estudios forenses de reconocimiento automático independiente del texto (expresiones de habla con textos espontáneos[1]) [16].

El reconocimiento automático del locutor, contemporáneo en el establecimiento de sus principios teóricos con el reconocimiento automático del habla, ha tenido sin embargo, menor atención y como consecuencia, un desarrollo inferior. Ha sido en estas últimas décadas cuando, a la luz de los nuevos e importantes campos de aplicación surgidos –como la biometría– que su desarrollo se ha acelerado.

Dentro del Reconocimiento Automático del Locutor (ASR, del inglés Automatic Speaker Recognition) [66] podemos distinguir dos tareas diferenciadas:

1. *Verificación Automática del Locutor* (ASV, del inglés Automatic Speaker Verification) [29]: El objetivo es verificar la identidad

[1]No se conoce a priori el contenido del habla, lo cual permite mayor flexibilidad, pero también incrementa la dificultad del problema.

reclamada por el locutor, o sea, se cuenta con el modelo de la muestra de voz del cliente y se recibe una solicitud de un individuo que dice ser el cliente y una muestra de su voz, la tarea a realizar consiste en comparar si la muestra y el modelo del cliente coinciden o no; la respuesta del sistema será, por lo tanto, binaria: identidad aceptada o rechazada. Hay diversas formas de medir la efectividad de este tipo de sistemas, una de las más utilizadas es la denominada tasa de igual error (EER, del inglés Equal Error Rate) [61]: es el error del sistema cuando el umbral de decisión es tal que el porcentaje de falsas aceptaciones es igual al de falsos rechazos (fig. 1) 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.

2. Identificación Automática del Locutor (ASI, del inglés Automatic Speaker Identification) [76]. Aquí el objetivo es, dada una muestra de voz y un conjunto de modelos de muestras de voces de clientes, señalar, dentro del grupo de modelos, cuales son los más posibles propietarios de la muestra, de forma ordenada.

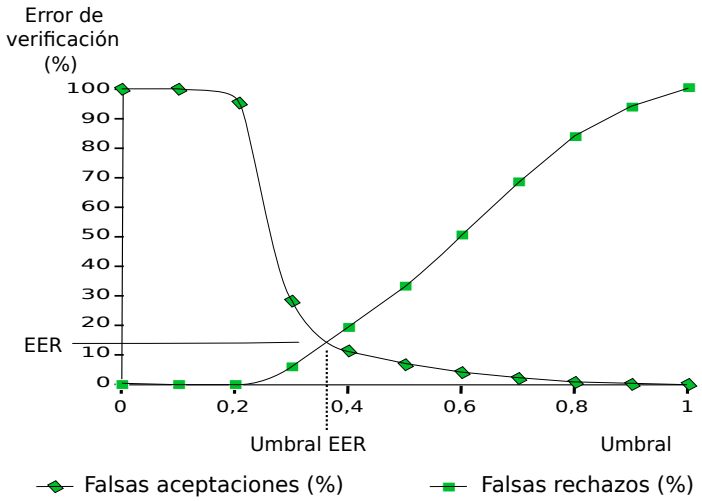


Figura 1: Curvas de error. Falsas aceptaciones: la entrada pertenece a un impostor y es falsamente aceptada. Falsos rechazos: la entrada pertenece al cliente y es falsamente rechazada.

En relación con dichos conceptos, Furui en 1994 [30] 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."

En las señales de voz existen varios niveles de información relacionados con la identidad del locutor, los rasgos de nivel superior están relacionados con: el tiempo de las pausas, patrones del tono y del tiempo, la semántica, la dicción, la pronunciación, la idiosincrasia, etc. Los *rasgos de bajo nivel* están basados en los rasgos acústicos extraídos de las características espectrales de las señales de voz. La fig. 2 resume los diferentes niveles de información adecuados para los sistemas de reconocimiento del locutor.

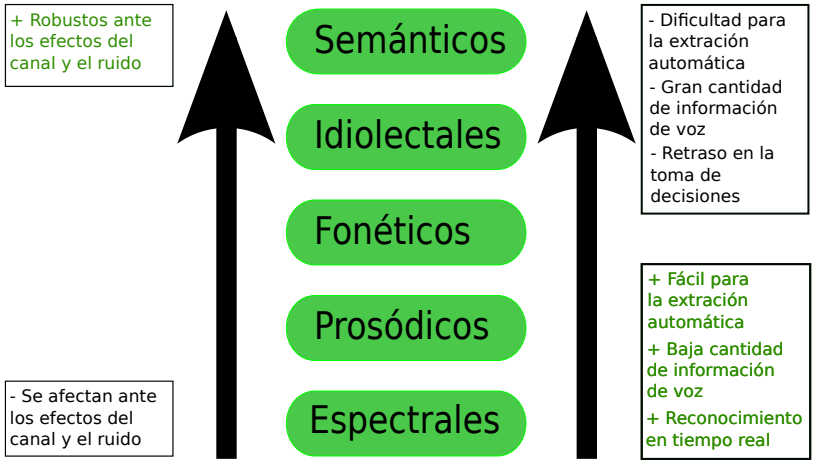


Figura 2: Niveles de información para los sistemas de reconocimiento del locutor.

El factor crítico para la aplicabilidad de los rasgos de alto nivel es la cantidad de información necesaria para obtener modelos estables y robustos del locutor, lo cual limita la posible aplicación de ellos en los sistemas reales. Este problema ha provocado que hasta la fecha, los rasgos de bajo nivel, que son características derivadas del espectro de voz, han demostrado ser los más eficaces y eficientes en los sistemas automáticos de reconocimiento del locutor, debido a que el espectro de voz refleja la geometría del sistema (tracto vocal) que genera la señal. En el caso del reconocimiento del locutor en específico, la variabilidad en la dimensión del tracto vocal, que puede considerarse un rasgo físico discriminatorio de cada persona, se refleja en la variabilidad del espectro de la voz entre locutores.

Es de señalar un grupo de factores que afectan los rasgos de bajo nivel y en especial a los espectrales, estos son: la variabilidad del canal de transmisión, el ruido en la señal de voz, el estado emocional o de salud del hablante, etc. En general, cualquier variación provocada por los factores anteriores, entre dos expresiones de voz de la

misma persona, se conoce como *variabilidad de sesión* [51]. La variabilidad de sesión se describe como la diferencia de condiciones entre el objeto de entrenamiento y el objeto de prueba y constituye el problema más difícil a enfrentar en el reconocimiento del locutor. Estos factores son los principales retos a que se enfrentan los algoritmos en el estado del arte de los ASR.

La motivación de esta tesis lo constituye la solución de problemas para robustecer los modelos en la verificación del locutor *independiente del texto* a partir de rasgos espectrales pertenecientes a señales de voz. Las investigaciones a realizar y los resultados que se esperan obtener deben tributar a la profundización en los métodos y algoritmos utilizados para el reconocimiento del locutor en general.

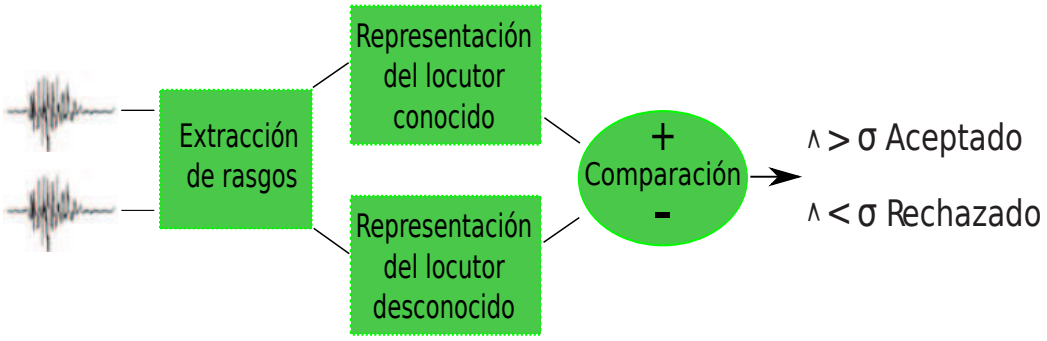


Figura 3: Componentes básicos de los Sistemas de Verificación del Locutor.

La fig. 3 muestra un diagrama en bloques de las etapas básicas de los sistema de ASV. En el módulo de extracción de rasgos se transforma la señal de voz en rasgos acústicos, donde se enfatizan las propiedades específicas del locutor. A partir de los rasgos del locutor conocido y desconocido, se entrenan sus respectivas representaciones y por último se comparan, obteniendo un valor de similitud que se confronta con un umbral de decisión resultando aceptado o rechazado.

En lo adelante, el desarrollo de esta tesis estará dirigido a llevar a cabo investigaciones en las áreas de "Representación del Locutor" y "Comparación".

Los métodos del estado del arte de los sistemas de ASR se basan fundamentalmente en la representación del locutor en el contexto de los Modelos de Mezclas Gaussianas (GMM, del inglés Gaussian Mixture Model) y su representación universal, conocida como Modelo Universal de Fondo (UBM, del inglés Universal Background Model) [8, 75]. El UBM ha de contener información acústica de los más diversos locutores posibles, para luego obtener un modelo por cada locutor derivado del UBM, utilizando la adaptación Máximo a Posterior (MAP, del inglés Maximum A-Posteriori Adaptation) [8, 75]. Este modelo adaptado contiene entre otra información, una gran cantidad de

información discriminatoria correspondiente al locutor. Es de notar que en la mayoría de la literatura sólo se adaptan los centros de las clases que conforman la distribución acústica del modelo UBM, esto se basa en que del modelo resultante sólo son utilizados las medias como parámetros discriminatorios de los posibles clasificadores.

Hace algunos años el enfoque de los super-vectores (SV) [21] se impuso como representación del locutor en los sistemas ASR, dicho enfoque tiene como base la adaptación MAP del UBM a los modelos GMM de los locutores, posibilitando en este marco que cada expresión de voz se represente por un super-vector que se obtiene de la concatenación de todos los vectores de medias de los componentes Gaussianos del GMM. Estos super-vectores forman un espacio de representación de altas dimensiones [21, 20], donde se llevarán a cabo todos los procesos restantes. Este enfoque significó el mayor avance en la evolución de los sistemas de reconocimiento del locutor al representarse cada expresión de voz como un vector, lo que permitió hacer comparables los resultados del reconocimiento del locutor con otros sistemas biométricos como el reconocimiento de la huella, el rostro, el iris, etc. En particular el espacio de los super-vectores posibilitó el uso de clasificadores discriminatorios como las Máquina de Vectores Soportes (SVM, del inglés Support Vector Machines) [21] y el modelado directo de la variabilidad de sesión[2] para su compensación. Más recientemente, dos nuevas soluciones han sido propuestas en el marco del enfoque de los super-vectores: el Análisis de Factores Conjunto (JFA, del inglés Joint Factor Analysis) [51] y el vector identidad (i-vector, del inglés identity vector en el sentido del reconocimiento del locutor o intermediate vector por su tamaño intermedio entre el de un super-vector y el de un vector de rasgos acústicos) [25], [26].

Los algoritmos basados en JFA e i-vector han mostrado los mejores niveles de rendimiento en las competencias del Instituto Nacional de Estándares y Tecnología (NIST, del inglés National Institute of Standards and Technology) en las Evaluaciones de Reconocimiento del Locutor (SRE, del inglés Speaker Recognition Evaluation) [18, 79, 53], pero presentan dos inconvenientes principales. En primer lugar, resulta difícil o imposible trabajar con la información secuencial/temporal del habla, ya que cada conjunto de vectores acústicos está representado por sólo un punto en el espacio de los super-vectores o en el espacio de los i-vectores. En segundo lugar, estos enfoques se basan en evaluaciones de modelos estadísticos, donde la influencia de una información específica se determina principalmente por la frecuencia de esta información. Es decir: si se produce un

[2]Dicha variabilidad se observa al contar con varias expresiones de locutores en diferentes momentos.

evento a menudo para un locutor dado, pero muy rara vez para los otros, apenas será tenido en cuenta por estos enfoques, lo cual podría parecer como una paradoja cuando el objetivo es discriminar los locutores.

En base a lo antes mencionado, se tiene como **problema de investigación** que las representaciones actuales de las expresiones de voz para el reconocimiento automático del locutor no reflejan de forma significativa los eventos acústicos frecuentes para un locutor dado pero poco frecuentes en la población, ni la información secuencial/temporal presente en la señal de voz de cada locutor, por tanto la información discriminatoria de dichas representaciones es insuficiente para desempeñar con mayor eficacia la verificación de personas por la voz.

Para resolver el problema, se propone como **objetivo general** de esta investigación, desarrollar nuevos métodos de representación de las expresiones de voz del locutor que tengan en cuenta los eventos discriminatorios independientemente de su frecuencia, contengan la información secuencial/temporal del habla y permitan modelar la variabilidad de sesión.

Para ello se proponen los siguientes **objetivos específicos**:

1. Aplicar métodos de reducción de dimensión sobre el espacio de los super-vectores para mostrar la gran cantidad de información redundante existente y la necesidad de utilizar un nuevo método de modelado para representar las expresiones de voz del locutor.

2. Proponer y evaluar un método de modelado para representar las expresiones de voz que incluya los eventos discriminatorios del locutor independientemente de su frecuencia de ocurrencia (información global).

3. Elaborar y evaluar métodos para enfrentar la variabilidad de sesión existente en las nuevas representaciones de las expresiones de voz.

4. Incorporar y evaluar la información secuencial/temporal de la voz al proceso de verificación del locutor.

En busca de mitigar las deficiencias propias de los enfoques actuales arribamos a la **hipótesis** principal:

Si se obtienen nuevas representaciones de las expresiones de voz que contengan la información global y temporal, y que permitan el modelado directo de la variabilidad de sesión, entonces la eficacia se elevaría en los sistemas de reconocimiento automático del locutor, en aplicaciones reales.

Para darle cumplimiento a los objetivos y demostrar la hipótesis planteada se propusieron las siguientes *tareas*:

1. Estudiar el estado actual de los algoritmos de reconocimiento automático del locutor independiente del texto adecuados para aplicaciones reales, específicamente los basados en un marco estadístico, para identificar las ventajas y desventajas de cada uno.

2. Analizar las representaciones actuales de la voz del locutor con el objetivo de mostrar sus limitaciones y la necesidad de un nuevo método de modelado para representar la voz.

3. Desarrollar e implementar un nuevo método de modelado para representar las expresiones de voz que tenga en cuenta los eventos discriminatorios del locutor indistintamente de su frecuencia.

4. Desarrollar e implementar una medida de similitud para evaluar la información intrínseca en la nueva representación del locutor.

5. Diseñar y realizar los experimentos que permitan comparar los resultados que se obtengan con la representación propuesta, con respecto a los resultados alcanzados por los métodos existentes en la literatura, utilizando para ello bases de datos internacionales de señales de voz.

6. Analizar los métodos actuales que enfrentan la variabilidad de sesión en el reconocimiento del locutor.

7. Elaborar e implementar nuevos métodos capaces de enfrentar la variabilidad de sesión existente en las nueva representación del locutor.

8. Diseñar y realizar los experimentos que permitan comparar los resultados que se obtengan con los métodos para la compensación de la variabilidad de sesiones propuestos, con respecto a los resultados alcanzados por los métodos existentes en la literatura, utilizando para ello bases de datos internacionales de señales de voz.

9. Desarrollar e implementar métodos capaces de capturar la información secuencial/temporal existente en la nueva representación del locutor.

10. Diseñar y realizar los experimentos que permitan evaluar la incorporación de la información temporal en el reconocimiento del locutor.

11. Proponer un algoritmo de reconocimiento del locutor que combine los métodos propuestos.

12. Implementar el algoritmo que integre los métodos propuestos y verificar su desempeño con señales de voz de bases de datos de pruebas internacionales y señales de voz obtenidas en condiciones reales.

13. Fusionar los resultados obtenidos con los resultados de los algoritmos actuales, para satisfacer el objetivo principal del trabajo.

Entre los **métodos de investigación** utilizados para llevar a cabo la investigación, se destaca el método general *hipotético-deductivo*, guiado por la observación de las problemáticas detectadas y el planteamiento de hipótesis que den respuestas a las preguntas ¿qué hacer? y ¿cómo lograrlo? que luego son corroboradas o validadas.

Como primer paso de este proceso científico se utilizó el método lógico *inductivo-deductivo*, realizándose un estudio del estado actual de los algoritmos de reconocimiento del locutor independiente del texto utilizados en condiciones reales, permitiendo obtener un conocimiento general de los problemas que afectan la eficacia de los mismos y proponer una hipótesis de partida para darle solución a las deficiencias detectadas.

Teniendo en cuenta que el problema detectado consta de varios elementos a resolver, se utilizó también el método *analítico-sintético* para descomponerlo en partes y profundizar en el estudio de cada una de ellas por separado, buscando soluciones parciales que luego son nuevamente integradas en el algoritmo general que se propone.

Se utilizó el método empírico de la *medición*, apoyado en procedimientos estadísticos y métodos matemáticos *algebraicos y aritméticos*, con el objetivo de representar las voces de locutores de manera numérica, poder determinar sus propiedades y relaciones de manera que pueda evaluarse la capacidad del algoritmo propuesto de discriminar entre un individuo y otro.

Finalmente, mediante el método empírico *experimental* se llevan a cabo las pruebas necesarias para validar cada uno de los métodos y el algoritmo que se propone, que a su vez se analizan y comparan con los resultados de los algoritmos de reconocimiento del locutor que aparecen en la literatura, apoyado en el método lógico de *comparación-clasificación*. De manera auxiliar se utiliza el método empírico *coloquial* para la presentación y discusión en sesiones científicas, de los resultados obtenidos.

La **novedad científica** de este trabajo radica en los aportes que se hacen al Reconocimiento del Locutor Independiente del Texto, específicamente permitiendo el uso de los eventos discriminatorios sin importar su frecuencia y la incorporación de

la información secuencial/temporal del habla para hacerle frente a la variabilidad de sesiones en las aplicaciones reales, lo cual es un aspecto no resuelto y de investigación activa en esta área. Los principales aportes son:

1. Un nuevo método para la creación de un Modelo Generador que permite obtener una representación binaria de la expresión de voz del locutor, que contiene eventos discriminatorios no contemplados en otras representaciones. El método incluye una nueva medida de semejanza entre dos representaciones binarias de locutores, que evalúa la información discriminatoria y temporal.

2. Un nuevo método de compensación de variabilidad de sesión que incluye la información común entre los locutores, dicha propuesta modificó tres métodos de compensación de variabilidad del estado del arte. El método incluye un nuevo criterio de selección (máscara) basado en la varianza de las especificidades de la representación binaria.

3. Dos nuevos métodos para obtener la información temporal contenida en la representación binaria. Un primer método obtiene información suprasegmental y un segundo método obtiene información de la dinámica entre las tramas.

La combinación de los resultados obtenidos con dichos métodos, soportados en la representación binaria de los locutores, con los resultados obtenidos con la representación i-vector, han elevado la eficacia del reconocimiento del locutor independiente del texto, con datos de competencias NIST. Adicionalmente, se pudo comprobar la gran redundancia existente en los super-vectores GMM, a partir de transformaciones topológicas, considerando que la voz yace sobre una variedad[3].

La significación práctica de este trabajo viene dada en primer lugar, por la incorporación del algoritmo propuesto en un sistema autóctono de reconocimiento del locutor independiente del texto. Los sistemas de reconocimiento del locutor existentes en el mercado internacional (Anexo A) presentan complejas restricciones para su adquisición y despliegue, además los precios suelen ser altos. Los métodos que se proponen en esta tesis, parten de la premisa de su eficiencia, de manera que puedan ser aplicados en un sistema real.

Es de señalar que esta investigación parte de un enfoque muy reciente y completamente diferente al actual en el reconocimiento del locutor, como resultados se

[3]Una variedad es el objeto geométrico estándar en matemática que generaliza la noción intuitiva de curva (1-variedad) y de superficie (2-variedad) a cualquier dimensión y sobre cuerpos diversos (no necesariamente el de los reales).

presentan en el trabajo algunos de los nuevos y primeros pasos en la evolución de este nuevo paradigma.

Este nuevo enfoque parte de una sencilla representación del habla, que se desplaza del área de trabajo probabilística continua a un espacio discreto binario y fue propuesta por J.F. Bonastre y otros investigadores en [1, 10, 11, 36]. Esta representación se basa en decisiones binarias locales, tomadas por cada una de las observaciones (vectores acústicos). A diferencia de los enfoques anteriores, esta área de trabajo binaria es capaz de modelar los eventos discriminatorios frecuentes y poco frecuentes, dado que representa un extracto de la voz mediante una matriz binaria, donde cada vector acústico está representado por un vector binario. Además como se demuestra en [11] y [36], la matriz binaria está ordenada temporalmente permitiendo la extracción de la información secuencial/temporal discriminatoria del locutor.

Por otra parte, el uso de un proceso de transformación muy simple de las matrices binarias permite la construcción de un nuevo espacio de super-vectores. El proceso de transformación consiste en una acumulación por columnas de la representación binaria, resultando un vector acumulativo que tiene la misma dimensión que los vectores binarios y representa de forma compacta la información discriminatoria de una expresión de voz del locutor. Esta especificidad de la representación binaria facilita compensar los efectos de la variabilidad de sesión, como se lleva a cabo en los enfoques JFA o i-vector, sin perder las cualidades intrínsecas del enfoque binario.

Los principales resultados y aportes teóricos que se presentan en esta tesis han sido presentados en eventos y publicaciones tanto nacionales como internacionales, las cuales aparecen listadas al final de este documento.

La estructura de la tesis consiste en introducción, tres capítulos, conclusiones, recomendaciones, referencias bibliográficas, glosarios y anexos. En el Capítulo 1 se presenta un estudio de los métodos actuales de Reconocimiento del Locutor, detallando los que se utilizan actualmente, de manera que se pueda comprender la magnitud de la problemática existente así como de las soluciones que se proponen. En el Capítulo 2 se propone un nuevo método de modelado, capaz de transformar la representación acústica de los locutores en una representación binaria. Muestra la necesidad de una nueva representación de la expresión de voz del locutor y describe la transformación propuesta. Además se presenta una nueva medida de similitud en el marco binario para evaluar la representación propuesta, utilizando bases de datos internacionales. En el Capítulo 3 se propone la incorporación de una nueva información en los métodos de compensación de la variabilidad de sesión en el marco binario, aplicada sobre tres

técnicas diferentes. Además se proponen dos vías para obtener la información temporal existente en las expresiones de voz. Se evalúan los métodos propuestos en bases internacionales.

Capítulo 1

Verificación del locutor independiente del texto

En este capítulo se presenta una panorámica sobre los métodos actuales del reconocimiento del locutor, describiendo sus principales etapas. Además se plasman algunas desventajas a enfrentar en el trabajo.

1.1. Métodos de Extracción de Rasgos Acústicos

Existen varios estudios en la literatura sobre la comparación de los resultados de reconocimiento del locutor utilizando rasgos de bajo nivel como Coeficientes Cepstrales de Frecuencia Mel (MFCC, del inglés Mel-Frequency Cepstral Coefficients [23] y [72], Coeficientes Cepstrales de Frecuencia lineal (LFCC, del inglés Linear Frequency Cepstral Coefficients [23], Coeficientes Cepstrales de Predicción lineal (LPCC, del inglés Linear Predictive Cepstral Coefficients [3] y otros.

La señal de voz incluye muchas características de las cuales no todas son importantes que estén contenidas en los rasgos para la discriminación del locutor, pero existe un grupo de características deseadas que deben estar presentes [55].

1. Sean robustas ante el ruido y la distorsión.

2. Se producen con frecuencia y naturalidad en el habla.

3. Sean fáciles de medir desde la señal de voz.

4. No sean afectadas por la salud del locutor o variaciones a largo plazo en la voz.

5. Sean difíciles de suplantar/imitar.

Los MFCC y los LFCC son las rasgos acústicos más ampliamente utilizados. Ellos son el cepstrum de la voz y se definen como la transformada inversa de Fourier del logaritmo del espectro de potencia a corto término (fig. 1.1). Son empleados para caracterizar tanto sonidos sonoros como sordos, con buenos resultados en la práctica en el reconocimiento del habla y del locutor [41].

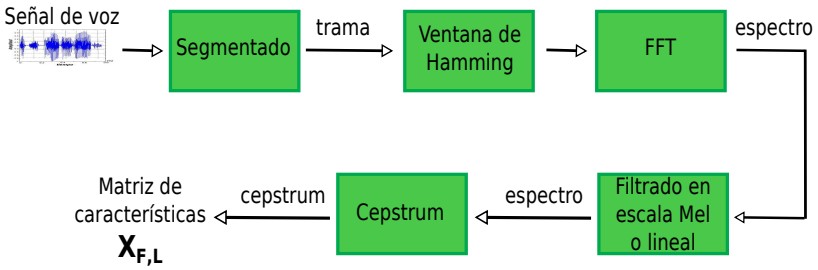


Figura 1.1: Pasos para la extracción de rasgos acústicos

El espectro de potencia a corto término se obtiene tomando porciones solapadas sucesivas de la señal (*Segmentado*), típicamente entre 20 y 30 ms, a las cuales se le aplica una ventana que suaviza los extremos (*Ventana de Hamming*), y se aplica la transformada de Fourier:

$$X[k] = \sum_{n=0}^{N-1} y[n]\ e^{-j(\frac{2\pi k}{N})n}, \tag{1.1}$$

donde $y[n] = x[n]w[n]$, $x[n]$ es la señal de voz, $w[n]$ la ventana de Hamming con $w[n] \neq 0$ para $0 \leq n \leq N-1$ y N es la cantidad de muestras.

Los MFCC fueron propuestos por primera vez para el reconocimiento del habla y su escala de frecuencia Mel imita el procesamiento de los sonidos por el oído humano. La cóclea del oído humano realiza un análisis espectral en una escala no lineal (escala Bark o Mel), que es lineal hasta 1000 Hz y aproximadamente logarítmica después [23]. Por tal motivo, al extraer los rasgos acústicos a corto término, es común efectuar una distorsión en frecuencia después del cálculo espectral, el espectro obtenido se procesa con un banco de filtros de M bandas de frecuencia, de acuerdo a la escala Mel.

Esta distorsión ocasiona que su resolución espectral se haga menor a medida que aumenta la frecuencia, provocando que la información existente en las altas frecuencias se afecte por el muestreo reducido de la escala Mel. Sin embargo, basado en la teoría de la producción de la voz [82], las características de los locutores asociadas con la estructura del tracto vocal, en particular la longitud del tracto vocal, se reflejan

más en la región de alta frecuencia del espectro de voz [88]. Motivado por esto en los últimos años se ha notado un aumento en el uso de los rasgos LFCC en los algoritmos de reconocimiento del locutor, lo cual conlleva a un cambio del banco de filtro, escala Mel por Lineal, al extraer los rasgos acústicos. En este trabajo se utilizarán principalmente los rasgos acústicos LFCC.

Por otra parte los rasgos acústicos son una "fotografía" del espectro en un cierto instante de tiempo, asumiendo que representan una señal estacionaria a corto término, sin información sobre su comportamiento en el transcurso del tiempo. Al hablar, los órganos articulatorios están cambiando su forma continuamente, este movimiento se refleja en el espectro en los cambios en las frecuencias y anchos de banda de los formantes, constituyendo un elemento identificativo del locutor [54].

La información dinámica de los rasgos acústicos o rasgos $delta(\Delta)$ [41, 70] se estima, calculando las derivadas de los rasgos en el tiempo y anexando dichas derivadas al vector de rasgos, elevando su dimensionalidad, lo que requiere un mayor volumen de datos para el entrenamiento de los clasificadores. Dada la matriz $X = \{x_1, ..., x_L\}$ que contiene los L vectores de rasgos se define:

$$\Delta x_{i,k} = \frac{\sum_{m=-F}^{F} m \; x_{i,k+m}}{\sum_{m=-F}^{F} m^2}, \tag{1.2}$$

aquí F es una función –lineal y decreciente– del índice del rasgo en cuestión, ya que los rasgos cepstrales de mayor orden varían más rápidamente [42].

Comúnmente se estiman también las derivadas en el tiempo de los rasgos Δ, conocidas como rasgos $delta - delta(\Delta\Delta)$, y se anexan al vector de rasgos, creciendo aun más la dimensionalidad del espacio de rasgos.

1.2. Métodos de clasificación

A partir de los rasgos acústicos del habla de cada locutor, previamente extraídos y seleccionados, se requiere encontrar un modelo que clasifique efectivamente al mismo y sea lo suficientemente robusto ante la variabilidad del habla.

Los métodos de clasificación han evolucionado en el tiempo. En la décadas del 70 predominó la clasificación comparando plantillas de palabras, en los 80 se clasificó aplicando la Distorsión Dinámica en el Tiempo (DTW, del inglés Dynamic Time Warping y la Cuantificación Vectorial (VQ, del inglés Vector Quantization). A partir de mediados de los 90 se desarrollan enfoques estadísticos de clasificación como los Modelos

Ocultos de Markov (HMM, del inglés Hidden Markov Model) y los Modelos de Mezclas Gaussianas (GMM), consolidándose el enfoque estadístico, en general, como la base de los algoritmos más eficaces para la clasificación del locutor hasta el momento.

1.2.1. Modelo de Mezclas Gaussianas (GMM)

El GMM es un modelo estocástico no supervisado que se ha convertido en el método de referencia dentro del área del reconocimiento del locutor y puede ser considerado una extensión del modelo de VQ, en el cual las clases son solapadas. Es decir un vector de rasgo no está asignado a la clase más cercana, sino que tiene una probabilidad distinta de cero desde el origen de cada clase.

Un GMM está compuesto por una mezcla finita de componentes Gaussianos, denotado por (λ) y se caracteriza por su función de densidad probabilística:

$$p(x|\lambda) = \sum_{k=1}^{K} P_k \, \mathcal{N}(x|\mu_k, \Sigma_k), \tag{1.3}$$

donde K es el número de componentes Gaussianas, P_k es la probabilidad a priori(peso de la mezcla) de la k-ésima componente Gaussiana y

$$\mathcal{N}(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{F/2}|\Sigma_k|^{1/2}} \; exp\left\{-\frac{1}{2}(x-\mu_k)^{'}\Sigma_k^{-1}(x-\mu_k)\right\}, \tag{1.4}$$

es la función de densidad Gaussiana, donde μ_k y Σ_k son el vector de medias y la matriz de covarianza correspondientes a la k-ésima mezcla. Las probabilidades a priori $P_k \geq 0$ están restringidas a $\sum_{k=1}^{K} P_k = 1$.

Por razones numéricas y computacionales, la matriz de covarianza del GMM es generalmente diagonal, es decir, contiene sólo la varianza por cada componente.

Entrenar un GMM consiste en la estimación de los parámetros $\lambda = \{p_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ a partir de la muestra de entrenamiento $X = \{x_1, ..., x_L\}$ obtenida de un extracto de voz del locutor. El enfoque clásico radica en estimar la Máxima Verosimilitud (ML, del inglés Maximum Likelihood) [75], partiendo del promedio del logaritmo de la verosimilitud de una secuencia de L vectores la muestra X con respecto al modelo λ y se define como:

$$LL_{arg}(X, \lambda) = \frac{1}{L}\sum_{l=1}^{L} log \sum_{k=1}^{K} P_k \, \mathcal{N}(x_l|\mu_k, \Sigma_k). \tag{1.5}$$

Cuanto mayor sea el valor, mayor es la indicación de que los vectores desconocidos

se originan a partir del modelo λ. Luego para maximizar la verosimilitud respecto a los datos dados se utiliza el algoritmo de Maximización de la Esperanza (EM, del inglés Expectation-Maximization) [9].

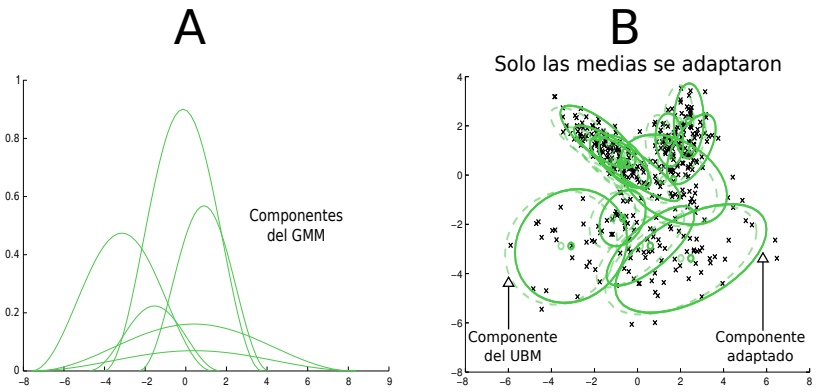


Figura 1.2: A partir de datos artificiales, la figura A muestra un ejemplo de un GMM con 6 componentes en una dimensión y la figura B muestra la adaptación de 12 componentes Gaussianos de un UBM a los datos de entrenamiento, utilizando el algoritmo MAP en dos dimensiones.

En el reconocimiento del locutor un UBM es un GMM entrenado vía EM, independiente del locutor, a partir de decenas o cientos de horas de datos de voz obtenidos de un gran número de locutores [75]. El propósito del UBM es contar con una representación de la distribución de las clases acústicas de una población, con el fin de obtener el modelo del locutor mediante la adaptación de los parámetros del UBM a la distribución de los rasgos acústicos del locutor. La *adaptación* de los modelos acústicos a las nuevas condiciones intrínsecas en los datos, es de suma importancia en el reconocimiento del locutor, debido a las diferencias entre locutores, el entorno, estilos del habla, etc.

En el año 2000 comienzan a aplicarse diferentes formas de adaptar o combinar los modelos GMM de los locutores con los UBM [75], predominando la adaptación Máximo A-Posterior (MAP, del inglés Maximum A-Posteriori) que ha sido clave en las mejoras de los clasificadores en esos años, alcanzando el estado del arte de dichos modelos durante la evaluación NIST SRE del 2004. La adaptación MAP consiste en obtener un GMM específico del locutor a partir del UBM. El modelo adaptado MAP se utiliza entonces como el modelo del locutor, de esta manera los parámetros del modelo no se estiman desde cero, ver imagen B en la fig. 1.2, además esto proporciona una conexión más estrecha entre el modelo del locutor y el UBM, permitiendo utilizar una técnica rápida para calcular la verosimilitud [75].

La adaptación MAP es un proceso de estimación de dos etapas, como el algoritmo de EM. La primera etapa es la misma que la del algoritmo EM, donde se calculan las estimaciones de las estadísticas suficientes de los datos de entrenamiento del locutor para cada componente en el UBM. La diferencia se encuentra en la segunda etapa del algoritmo de EM, para la adaptación, las "nuevas" estimaciones de los parámetros del modelo del locutor se combinan con las estimaciones de los parámetros del modelo de fondo UBM "viejo", usando un coeficiente para la combinación que es dependiente de los datos. El coeficiente que se utiliza para la combinación de los parámetros tiene que garantizar que: las componentes con altas verosimilitudes con los datos del locutor influyan más que las componentes del UBM en la actualización de los parámetros, y que en las componentes con bajas verosimilitudes con los datos del locutor influyan más las componentes del modelo UBM.

Dada la muestra de entrenamiento $X = \{x_1, ..., x_L\}$, y el UBM, $\lambda_{UBM} = \{p_k, \mu_k, \Sigma_k\}_{k=1}^{K}$, los vectores de medias adaptados $\mu_k^{'}$ a través del método MAP se obtienen por la suma pesada de los datos de entrenamiento del locutor y los vectores de medias del UBM:

$$\mu_k^{'} = \alpha_k \tilde{\mu}_k + (1 - \alpha_k)\mu_k, \tag{1.6}$$

donde $\mu_k^{'}$ es el vector de medias adaptado,

$$\alpha_k = \frac{n_k}{n_k + r} \tag{1.7}$$

es el coeficiente dependiente de los rasgos, con r un factor fijo de relevancia,

$$\tilde{\mu}_k = \frac{1}{n_k} \sum_{l=1}^{L} P(k|x_l)x_l \tag{1.8}$$

el vector de medias obtenido del vector de rasgo l,

$$n_k = \sum_{l=1}^{L} P(k|x_l) \tag{1.9}$$

el peso de la mezcla k y

$$P(k|x_l) = \frac{P_k \mathcal{N}(x_l|\mu_k, \Sigma_k)}{\sum_{i=1}^{K} P_i \mathcal{N}(x_l|\mu_i, \Sigma_i)} \tag{1.10}$$

es la probabilidad a posteriori para la i-ésima clase acústica (mezcla).

La relevancia del coeficiente r y por tanto α_k, controla el efecto de las muestras

de entrenamiento sobre el modelo resultante, con respecto al UBM.

En el reconocimiento del locutor, el modelo adaptado vía MAP y el UBM se conocen comúnmente como Modelo de Mezclas Gaussianas — Modelo Universal de Fondo (GMM-UBM, del inglés Gaussian Mixture Model - Universal Background Model).

Finalmente los índices de verosimilitud dependen del modelo conocido $\lambda_{cliente}$ y el modelo de fondo λ_{UBM} a través del promedio del logaritmo de la razón de probabilidad:

$$LLR_{arg}(X, \lambda_{cliente}, \lambda_{UBM}) = \frac{1}{L} \sum_{l=1}^{L} \{log\ p(x_l|\lambda_{cliente}) - log\ p(x_l|\lambda_{UBM})\}, \qquad (1.11)$$

que esencialmente mide la *diferencia* entre el modelo del cliente y el de fondo respecto a la muestra $X = \{x_1, ..., x_L\}$. Notar que el uso común del UBM, para todos los locutores, hace que el rango de los valores de puntuación entre diferentes locutores, sea comparable.

Existen métodos de adaptación alternativos al MAP y la decisión de cuál utilizar depende de la cantidad de datos de entrenamiento disponibles [60]. La Regresión Lineal de Máxima Verosimilitud (MLLR, del inglés Maximum Likelihood Linear Regression) [57] ha mostrado ser un método eficaz para el entrenamiento del modelo con expresiones de voz de corta duración, que aunque originalmente fue desarrollada para el reconocimiento del habla, ha sido aplicada con éxito en el reconocimiento del locutor [46, 60]. Tanto la adaptación MAP como la MLLR forman las bases para los clasificadores basados en *super-vectores*, que se explicarán en la Sec. 1.2.3.

Es de señalar que los GMM han sido utilizados por 20 años [76] como la base del reconocimiento del locutor, pero presentan algunas problemáticas como:

- Por su naturaleza las componentes Gaussianas se solapan en el espacio de trabajo, lo cual es beneficioso pero origina la posibilidad de que algunos componentes se encuentren totalmente solapados, sobre todo cuando la "*cantidad de clases acústicas*" presentes en el extracto de voz es mucho menor que la "*cantidad de Gaussianas del modelo*". Por lo general en la literatura suele utilizarse 512 componentes en un espacio de 50 dimensiones, que representa un equilibrio entre el costo computacional del modelo y su eficacia ante el reconocimiento del locutor. Esta cantidad de Gaussianas implica que una gran parte de ellas se encuentren solapadas dentro de otras y con bajo peso, aportando una probabilidad cercana

a cero para la mayoría de las muestras representadas por su distribución de densidades. Basado en esto, se considera que las GMM-UBM contienen gran cantidad de información redundante dentro del modelo [37].

- Alto costo computacional, sobre todo para el entrenamiento del UBM.

1.2.2. Máquina de Vectores Soportes (SVM)

La Máquina de Vectores Soportes (SVM) es un potente clasificador discriminatorio adoptado en el año 2006 en el área de reconocimiento del locutor. Este clasificador ha sido aplicado a nivel espectral [20, 21], a nivel prosódico [28] y sobre los rasgos de alto nivel [19]. Hasta el año 2010 fue el clasificador más utilizado en el reconocimiento del locutor, sobre todo a partir del éxito de su combinación con el GMM, aumentado considerablemente la eficacia del reconocimiento [20, 21]. En la tarea de verificación del locutor (ASV), como se ilustra en la fig. 1.3, la SVM es un clasificador discriminatorio binario que modela la frontera de decisión entre dos clases utilizando un hiperplano de separación, a diferencia del GMM que es un clasificador generativo que modela las distribuciones de probabilidad de las clases.

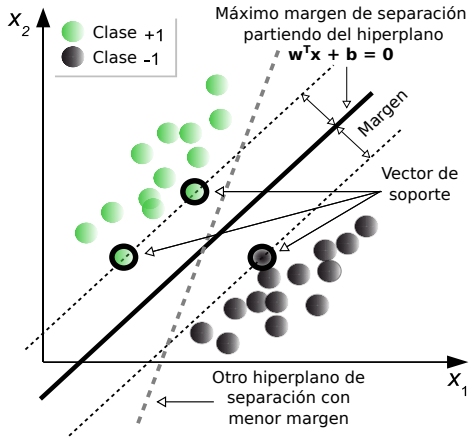


Figura 1.3: Estructura principal de la SVM, el hiperplano separa las muestras de entrenamiento positivas (+1) y las negativas (-1).

En la ASV, las clases se dividen en 2: etiquetados como +1, los vectores de entrenamiento del locutor cliente y como −1 los vectores de entrenamiento de una población de impostores (pueden emplearse los locutores utilizados para crear el UBM). Empleando los vectores de entrenamiento etiquetados la SVM encuentra un hiperplano que separa y maximiza el margen de separación entre las dos clases. Formalmente, la función discriminatoria de la SVM está dada por [20],

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K(x, x_i) + d. \tag{1.12}$$

Aquí los $t_i \in \{+1, -1\}$ son los valores de las salidas ideales, $\sum_{i=1}^{N} \alpha_i t_i = 0$ y $\alpha_i > 0$. Los vectores de soporte x_i, sus correspondientes pesos α_i y término de sesgo d, se determinan a partir de un conjunto de entrenamiento usando un proceso de optimización. La función núcleo $K(\cdot, \cdot)$ está diseñada de modo que se pueda expresar como $K(x, y) = \phi(x)'\phi(y)$, donde $\phi(x)$ es un mapeo desde el espacio de los rasgos de entrada hacia algún espacio de características de altas dimensiones donde se comporten de forma lineal. En un espacio de altas dimensiones dos clases se separan más fácilmente con un hiperplano. Intuitivamente, se corresponde a una frontera de decisión no lineal en el espacio de entrada original (por ejemplo, el espacio de los rasgos LFCC).

1.2.3. Los super-vectores: un paso en la evolución

¿Qué es un super-vector?

En el año 2006 uno de los problemas que enfrentaba el reconocimiento del locutor fue cómo representar con sólo un punto en el espacio los extractos de una señal voz que, en general, tienen un número variable de vectores de rasgos acústicos. Es entonces que se presenta por primera vez un camino robusto para representar un extracto de voz utilizando un único vector, llamado *super-vector*.

El super-vector nace de la combinación de muchos vectores de bajas dimensiones en un vector de altas dimensiones, en el caso de los GMM, se concatenan o apilan los vectores de medias con dimensión F de K componentes Gaussianas adaptadas creando un super-vector de dimensión FK [21]. Formalmente, dado una matriz de rasgos acústicos $A_{F,L}$ y su modelo $\lambda = \{p_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ adaptado GMM-UBM, o simplemente un GMM, del cual sólo se tendrá en cuenta la matriz de medias μ, el proceso de obtención del super-vector consiste en:

$$\mu_{\mathbf{s}} = \left\{ \begin{array}{cccc} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_K \\ \downarrow & \downarrow & \downarrow & \downarrow \\ \left\{ \begin{array}{cccc} x_{1,1} & x_{1,2} & \dots & x_{1,K} \\ \vdots & \vdots & \vdots & \vdots \\ x_{F,1} & x_{F,2} & \dots & x_{F,K} \end{array} \right\} \end{array} \right\} \quad super - vector_s = SV_s = \left\{ \begin{array}{c} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_{FK} \end{array} \right\},$$

donde SV_s representa el super-vector del locutor s.

Es importante resaltar que los super-vectores de diferentes extractos de voz surgen desde un "sistema de coordenadas común", o sea la concatenación sigue el mismo orden de las componentes del UBM para todos los extractos de voz. De esta manera los elementos de los super-vectores están alineados y posibilitan el cálculo de la similitud.

Esta evolución permitió utilizar los super-vectores como parámetros directos de los clasificadores SVM, como se muestra en la fig. 1.4.

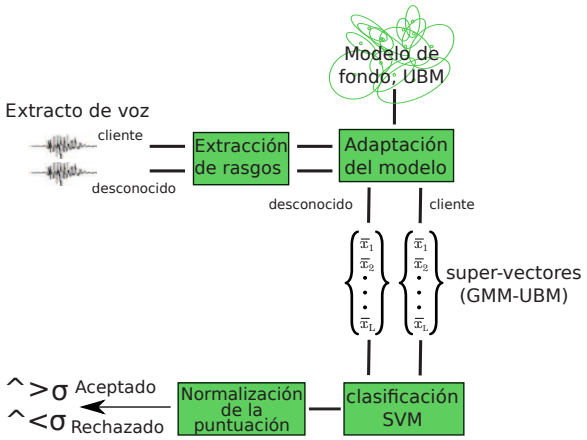


Figura 1.4: Estructura de los sistemas de reconocimiento del locutor sobre el espacio de los super-vectores.

Un ejemplo de un sistema de reconocimiento del locutor con SVM fue el que utilizó super-vectores GMM-UBM con Núcleo Lineal(GSL, del inglés GMM-UBM Supervector Linear Kernel) [21, 24, 56], que es un clasificador híbrido, donde se utilizó el modelo generativo GMM-UBM para crear los parámetros de entrada del clasificador discriminatorio SVM.

El desarrollo de métricas en el espacio de las GMM [7] llevó a la idea de emplear la divergencia Kullback-Liebler (KL) para definir nuevos núcleos secuenciales basados en los super-vectores. Supongamos que tenemos un UBM, $\lambda_{UBM} = \{p_k, \mu_k, \Sigma_k\}_{k=1}^K$, y dos extractos de voz a y b los cuales están descritos por sus modelos adaptados GMM-UBM (Subsección 1.2.1), $\lambda_a = \{p_k, \mu_k^a, \Sigma_k\}_{k=1}^K$ y $\lambda_b = \{p_k, \mu_k^b, \Sigma_k\}_{k=1}^K$ correspondientes (observe que sus modelos sólo difieren en la matriz de media). Entonces el núcleo de divergencia KL se define como,

$$K(\lambda_a, \lambda_b) = \sum_{k=1}^{K} (\sqrt{p_k}\Sigma_k^{-1/2}\mu_k^a)' \ (\sqrt{p_k}\Sigma_k^{-1/2}\mu_k^b). \tag{1.13}$$

Notar que su formulación matemática beneficia su implementación, o sea que todos

los vectores de medias Gaussianas μ_k pueden ser normalizados con $\sqrt{P_k}\Sigma_k^{-1/2}$ antes del proceso de entrenamiento de la SVM. En este caso aunque sólo los vectores de medias del GMM son incluidos en el super-vector, la información de la varianzas y los pesos se encuentran presentes implícitamente por el rol de la normalización.

El trabajo en el espacio de los super-vectores significó el mayor paso, hasta el momento, en la evolución de los algoritmos para el reconocimiento del locutor, a continuación se presentan algunas de las ventajas y desventajas de este espacio.

- Ventajas

  - La representación del locutor mediante un único vector o punto en el espacio de altas dimensiones permitió utilizar fácilmente los clasificadores discriminatorios para el reconocimiento del locutor.
  - Por primera vez el espacio de trabajo permitió el modelado directo de la variabilidad de sesión, como se verá en la Sección 1.3.
  - La eficacia obtenida por los algoritmos en el espacio de los super-vectores aventaja cualquier otra representación hasta ese momento [27, 55].

- Desventajas

  - Gran cantidad de información redundante presente en los super-vectores, acarreada de los GMM-UBM.
  - Este enfoque se basa en la estadística, donde la influencia de una información específica se determina principalmente por la frecuencia de esta información. Es decir, si se produce un evento a menudo para un locutor dado, pero muy rara vez para los otros, apenas será tenido en cuenta por este enfoque, lo cual podría parecer como una paradoja cuando el objetivo es discriminar los locutores.
  - Resulta imposible trabajar con la información secuencial/temporal del habla, ya que cada conjunto de vectores acústicos está representado por sólo un punto en el espacio de los super-vectores.
  - Posee altas dimensiones y un complejo proceso de entrenamiento que presenta un alto costo computacional.

1.3. Compensación de la sesión en el marco de las GMM

Un nuevo avance en el reconocimiento del locutor se desarrolló entre los años 2007 y 2010, basado en el enfoque de los super-vectores [52, 86], explícitamente la *compensación de la variabilidad de sesión*. Esta nueva forma de enfrentar cualquier variación en las expresiones de voz (respecto al canal, el entorno, el contenido fonético o ruido aditivo) se debe a que cada expresión o extracto de voz se presenta ahora como un único punto en el espacio de los super-vectores, permitiendo cuantificar y eliminar directamente la variabilidad no deseada en ellos.

¿Significa esto que se necesitan varias expresiones de entrenamiento grabadas a través de distintos canales telefónicos o micrófonos o entornos de un mismo locutor para su entrenamiento? No necesariamente, por el contrario, el modelo de variabilidad entre las sesiones se entrena con datos independientes y luego es utilizado al entrenar un super-vector, de un nuevo locutor, para mitigar su variabilidad.

Las técnicas establecidas para manejar estos problemas operan a nivel de rasgos [74], a nivel del modelo generativo [35, 49], en el marco de la SVM [81, 80] y a nivel de los valores de puntuación como por ejemplo: la Normalización del Auricular (H-norm, del inglés Handset Normalization) [73] y la Normalización de la prueba (T-norm, del inglés Test Normalization) [4].

Diversos autores han desarrollado de manera independiente diferentes técnicas de compensación de la variabilidad de sesión, tanto en el marco de los modelos generativos GMM-UBM de los locutores como en el marco del clasificador discriminatorio SVM. Los métodos que siguen el Análisis de Factores (FA, del inglés Factor Analysis) propuesto por Kenny en [49] están diseñados para el reconocedor basado en GMM y utilizan explícitamente las propiedades estocásticas del mismo, mientras que los métodos desarrollados en el espacio de los super-vectores, como la Proyección de Atributos No Deseados (NAP, del inglés Nuisance Attribute Projection) propuesto por Solomonoff en [80], se basan en el álgebra lineal.

La hipótesis subyacente de estos dos enfoques asume que existe un subespacio de bajas dimensiones donde se encuentra presente la variabilidad de sesión con un limitado solapamiento de la información específica del locutor.

1.3.1. Análisis de factor (FA)

El FA fue introducido por Kenny [49, 50] y Vogt [85], esta técnica opera sobre los modelos generativos con los enfoques estadísticos tradicionales (como el algoritmo de EM) en busca de modelar la variabilidad entre sesiones.

El enfoque más reciente es el Análisis de Factores Conjunto (JFA, del inglés Joint Factor Analysis) [47, 51].

El modelo JFA considera la variabilidad de los super-vectores como una combinación lineal de los locutores y los componentes del canal. Dado una muestra de entrenamiento, el super-vector obtenido $\boldsymbol{M}$, dependiente de la información específica del locutor y de la sesión, se descompone en dos componentes estadísticamente independientes, como sigue

$$M = s + c, \tag{1.14}$$

donde $\boldsymbol{s}$ y $\boldsymbol{c}$ son referidos como el super-vector del locutor y el super-vector del canal respectivamente. Dado F, la dimensión de los vectores de rasgos acústicos y K el número de mezclas en el UBM, entonces los super-vectores $\boldsymbol{M}$, $\boldsymbol{s}$ y $\boldsymbol{c}$ yacen en un espacio paramétrico de dimensión FK. Luego la variabilidad del canal es modelada de la forma,

$$c = Ux, \tag{1.15}$$

Durante el entrenamiento el factor $\boldsymbol{x}$ debe ser estimado de forma conjunta con el factor del locutor $\boldsymbol{y}$, de la siguiente forma:

$$s = m + Vy + Dz, \tag{1.16}$$

sustituyendo las ec. 1.15 y 1.16 en la ec. 1.14 tenemos,

$$M = m + Vy + Ux + Dz, \tag{1.17}$$

donde $\boldsymbol{m}$ es un super-vector independiente del locutor y del canal obtenido del UBM, $\boldsymbol{V}$ es una matriz rectangular de bajo rango que define el subespacio de los locutores (vectores propios del subespacio, nombrada matriz de voces-propias (del inglés eigenvoices matrix)), $\boldsymbol{D}$ es una matriz diagonal con dimensión $FK \times FK$ que contiene la información residual del subespacio de los locutores y $\boldsymbol{U}$ es una matriz rectangular de bajo rango que define el subespacio de las sesiones, nombrada matriz

de canales-propios (del inglés eigenchanel matrix). Los vectores $\boldsymbol{y}$, $\boldsymbol{z}$ y $\boldsymbol{x}$ son factores dependientes del locutor y de la sesión en sus respectivos espacios y se asume que cada uno es una variable aleatoria con distribución normal $\mathcal{N}(0, I)$.

El reconocimiento del locutor aplicando JFA consiste en estimar primero los subespacios (es decir $\boldsymbol{V}$, $\boldsymbol{D}$ y $\boldsymbol{U}$) [47, 52] desde bases de datos de desarrollo apropiadamente etiquetadas y luego estimar los factores del locutor y la sesión (es decir $\boldsymbol{y}$, $\boldsymbol{z}$ y $\boldsymbol{x}$) para un nuevo extracto de voz del cliente. Finalmente el super-vector dependiente de la sesión $\boldsymbol{c}$ se descarta y es utilizado el super-vector dependiente del locutor que está dado por la ec. 1.16, donde en el caso especial en que $y = 0$, entonces $s = m + Dz$ describe exactamente el mismo proceso de adaptación MAP. Por lo tanto, el super-vector del locutor en el JFA se puede ver como una extensión del MAP con la matriz de voces-propias $\boldsymbol{Vy}$ incluida.

La similitud o puntuación viene dada por el cálculo del logaritmo de la verosimilitud (LLR) entre el modelo del locutor cliente s y el UBM con el extracto de voz de prueba X.

La técnica JFA dominó las evaluaciones NIST SRE 2008 [18] y 2010 [79] y fue indicado claramente su potencial por diferentes grupos de investigadores en el área.

1.3.2. Espacio de Variabilidad Total (T)

Recientemente ha surgido un nuevo enfoque en el reconocimiento del locutor, donde el método de FA fue utilizado como extractor de rasgos para una nueva variante, llamada Espacio de Variabilidad Total (T, del inglés Total variability), planteado por Dehak y otros en [26].

Como se describió, el modelado clásico de variabilidad de sesión JFA, basado en los factores del locutor y de la sesión, consiste en definir dos subespacios distintos: el subespacio de los locutores definido por la matriz de *voces-propias* V y el subespacio de sesión representado por la matriz *canales-propios* U. El nuevo enfoque propone definir un sólo subespacio, el espacio de variabilidad total T, que contenga simultáneamente las variabilidades del locutor y de sesión. Esta propuesta fue motivada porque se mostró que el factor de la sesión en el JFA, encargado de modelar solamente los efectos de variabilidad de sesión, también contenía información específica del locutor.

Este nuevo espacio está definido por una matriz de variabilidad total que contiene los vectores-propios (del inglés eigenvectors) correspondientes a los mayores valores-propios (del inglés eigenvalues) de la matriz de covarianza que encierra la variabilidad del espacio. En T no se hace distinción entre los efectos de los locutores y los efectos

de las sesiones contenidas en el espacio de los super-vectores.

La expresión definida en la ec. 1.17 se vuelve a escribir de la siguiente manera:

$$M = m + Tw, \tag{1.18}$$

donde m es el super-vector con información independiente del locutor y de la sesión (el super-vector creado a partir las medias del UBM), T es una matriz rectangular de bajo rango y w nombrado vector intermedio o i-vector con dimensión W, es un vector aleatorio que sigue una distribución normal estándar $\mathcal{N}(0, I)$, donde sus componentes son los factores. En este modelo se supone que el vector M siga una distribución normal con vector de media m y como covarianza la matriz TT'.

Considerando a F como la dimensión de los vectores de rasgos acústicos y K el número de mezclas en el UBM, la matriz de variabilidad total T presenta dimensiones $FK \times W$ y se obtiene empleando el mismo proceso de aprendizaje utilizado para la matriz de voces-propias V, excepto por una diferencia importante: en el entrenamiento de las voces-propias, todas las grabaciones de un locutor se consideran que pertenecen a la misma persona; sin embargo en el caso del entrenamiento de la matriz de variabilidad total, todas las expresiones de voz de un locutor se consideran que han sido producidas por diferentes locutores. Concluyendo, esta técnica puede verse como análisis de factor que permite proyectar un extracto de voz a un espacio de variabilidad total de bajas dimensiones.

El factor total w es una variable oculta, la cual se puede definir por su distribución Gaussiana utilizando las estadísticas de Baum-Welch de un extracto de voz dado, resultando que la media de la distribución se corresponde exactamente con el i-vector.

Las estadísticas de Baum-Welch utilizadas para obtener el i-vector son extraídas utilizando el UBM. Dado un extracto de voz $X = \{x_1, ..., x_L\}$ y un UBM λ_{UBM} compuesto por K componentes definidos en un espacio de rasgos de dimensión F, las estadísticas de Baum-Welch son obtenidas a través de:

$$N_k = \sum_{l=1}^{L} P(k|x_l, \lambda_{UBM}), \tag{1.19}$$

$$F_k = \sum_{l=1}^{L} P(k|x_l, \lambda_{UBM})\ x_l, \tag{1.20}$$

donde N_k es la estadística de cero orden y F_k la de primer orden, con $k = \{1, ..., K\}$ índice de las Gaussianas y $P(k|x_l, \lambda_{UBM})$ corresponde con la probabilidad a posterior

de la componente Gaussiana k modelando el vector de rasgos x_l. Además, con el fin de estimar los i-vectores, también se necesita calcular las estadísticas de primer orden centralizadas de Baum-Welch, basadas en los vectores medios de los componentes del UBM.

$$\tilde{F}_k = \sum_{l=1}^{L} P(k|x_l, \lambda_{UBM}) \ (x_l - m_k), \tag{1.21}$$

donde m_k es el vector medio de la componente Gaussiana k del UBM. Tanto la estadística de cero orden como la de primer orden presentan una complejidad computacional de $O(KL)$ [32].

Luego para obtener el i-vector de la expresión de voz X se utiliza la siguiente ecuación:

$$w = H^{-1} \ T'\Sigma^{-1}\tilde{F}, \tag{1.22}$$

con L definido como:

$$H = I + T'\Sigma^{-1}NT. \tag{1.23}$$

Se define $N(X)$ como una matriz diagonal de dimensión $FK \times FK$ cuyos bloques diagonales son $N_k I(k = \{1, ..., K\})$, $\tilde{F}(X)$ es un super-vector de dimensión FK obtenido por la concatenación de todas las estadísticas de primer orden de Baum-Welch $\tilde{F}_k$ de X. La covarianza diagonal Σ es una matriz de dimensión $FK \times FK$ estimada durante el entrenamiento del factor y modela la variabilidad residual no capturada por la matriz de variabilidad total T.

La complejidad computacional al calcular un i-vector w es $O(W^3 + W^2K + WFK)$ [32, 2, 59], donde el término W^3 proviene del cálculo de la inversa de la matriz H mientras que el término W^2K es provocado por el cálculo de $I + T'\Sigma^{-1}NT$. Obsérvese que cuando K es grande $(K > W)$ el término W^2K causa un costo computacional enorme.

El trabajo en el espacio de los i-vectores ha marcado una pauta en la evolución de los algoritmos para el reconocimiento del locutor y constituye la base de la mayoría de los sistemas actuales, con las siguientes ventajas y desventajas.

- Ventajas

  - Por primera vez se tiene una representación de la expresión de voz del locutor, de bajas dimensiones y alta eficacia, además, encaminó el problema

del reconocimiento del locutor al problema clásico de reconocimiento de patrones biométricos.

- Permite el modelado directo de la variabilidad de sesión.
- Presenta una representación compacta de la voz, que provoca un menor costo computacional al calcular la puntuación entre dos i-vectores.

▪ Desventajas

- Los paradigmas anteriores y este enfoque se basan en la estadística, donde la influencia de una información específica se recopila principalmente por la frecuencia de esta información.
- Resulta difícil trabajar con la información secuencial/temporal del habla, ya que cada conjunto de vectores acústicos está representado por sólo un punto en el espacio de los i-vectores.
- Aunque la estimación del i-vector es menos costosa que el cálculo del supervector, es un complejo proceso de entrenamiento con un alto costo computacional.

1.4. Compensación de la sesión en el marco de los i-vectores

La variabilidad de sesión es conocida por ser un factor importante en la degradación del funcionamiento de los sistemas, compensar esta variabilidad es obligatorio en los modernos sistemas de reconocimiento del locutor.

Es de notar que al fusionar en un solo espacio (espacio de variabilidad total) el espacio del locutor y el de las sesiones, se acarrea al espacio de los i-vectores el problema de la variabilidad de sesiones, por lo tanto se convierte en indispensable la aplicación de técnicas de compensación de la variabilidad de sesión sobre los i-vectores en lugar de aplicarlas en el espacio definido por las GMM, como es el caso del JFA. Las técnicas de compensación de la variabilidad necesitan grandes cantidades de datos para lograr encontrar el comportamiento dentro de las clase (locutores) y entre clases, provocando que altas dimensiones en los datos atenten con los recursos de las computadoras, lo cual queda resuelto en el dominio de los i-vectores, permitiendo estos aplicar nuevas técnicas de compensación de variabilidad de sesión dada la baja dimensión de su

espacio comparado con los super-vectores. En esta sección se presentarán cuatro de estas técnicas de compensación.

1.4.1. Proyección de Atributos No Deseados (NAP)

La Proyección de Atributos No Deseados (NAP) fue introducida por Solomonoff en [80] y consiste en encontrar una matriz de proyección apropiada para intentar eliminar los atributos no deseados del espacio de trabajo, o sea eliminar o mitigar la variabilidad intra-clases proyectando los vectores de rasgos x (i-vectores) sobre un subespacio ortogonal complementario. Para utilizar el NAP, se requiere bases de datos de señales de voz etiquetadas con la información del canal y del locutor, con el fin de entrenar los parámetros de transformación. La matriz de proyección es formulada como:

$$\tilde{x} = (I - U'U)x, \tag{1.24}$$

donde I es la matriz de identidad y U es la matriz de proyección ortogonal de $r \times d$ dimensiones, donde d es la dimensión del espacio y r es la cantidad de ejes principales (los vectores-propios asociados con los mayores valores-propios) obtenidos al resolver el problema del valor-propio

$$S_w u = \lambda u. \tag{1.25}$$

donde $u = \{u_0, u_1, ..., u_{d-1}\}$ son los vectores de solución con sus correspondientes valores-propios $0 = \lambda_0 \leq \lambda_1 \leq ... \leq \lambda_{d-1}$. Los vectores-propios asociados con los valores-propios con valor cero quedan fuera y son utilizados los restantes vectores-propios. Se define S_w como la matriz de covarianza que contiene la dispersión intra-clases de los vectores de rasgos,

$$S_w = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (x_i^s - x_s)(x_i^s - x_s)', \tag{1.26}$$

donde S es el número de locutores, n_s el número de extractos de voz correspondientes a cada locutor s, x_s es el vector medio de las expresiones de voz de cada locutor ($x_s = \frac{1}{n_s} \sum_{i=1}^{n_s} x_i^s$), y x_i^s es la i-ésima expresión de voz del locutor s.

Dada la naturaleza del método NAP puede ser utilizado en cualquier clasificador que evalúe vectores de observación de altas dimensiones. Por ejemplo, una medida de similitud entre dos vectores *x1* y *x2* puede ser:

$$S(x_1, x_2) = (x_1 - U'Ux_1)(x_2 - U'Ux_2) = x_1x_2 - Ux_1Ux_2. \tag{1.27}$$

Otra medida de similitud puede ser la distancia del coseno, formulada como sigue

$$S_{cos}(x_1, x_2) = \frac{x_1x_2 - Ux_1Ux_2}{||x_1 - U'Ux_1||||x_2 - U'Ux_2||}. \tag{1.28}$$

1.4.2. Normalización de la Covarianza Intra-clase (WCCN)

La Normalización de la Covarianza Intra-clase (WCCN, del inglés Within-class Covariance Normalization) fue propuesta inicialmente por Hatch en [34] como una técnica de compensación de variabilidad para entrenar el clasificador SVM, su objetivo fue minimizar el error esperado debido a las falsas aceptaciones y falsos rechazos durante el entrenamiento. Una adaptación de WCCN, propuesta por Dehak y otros, para trabajar sobre el espacio de los i-vectores fue presentada en [26].

En busca de normalizar los vectores se utiliza la siguiente función de proyección:

$$\varphi(x) = B'x \tag{1.29}$$

donde x es el vector de rasgo y B' es una matriz superior obtenida de la descomposición de Cholesky de $W^{-1} = BB'$. Obsérvese que W es semidefinida positiva y simétrica por construcción y se alcanza utilizando la ecuación 1.26 ($W = S_w$), similar a la técnica NAP.

Pueden presentarse varias medidas de similitud, entre ellas las más utilizadas son:

$$S(x_1, x_2) = (B'x_1)'(B'x_2), \tag{1.30}$$

o la propuesta por Dehak y otros en [26]:

$$S_{cos}(x_1, x_2) = \frac{(B'x_1)'(B'x_2)}{||B'x_1||||B'x_2||}. \tag{1.31}$$

Esta normalización está dirigida a compensar la variabilidad de sesión, garantizando la conservación de las direcciones (dimensiones) en el espacio, en contraste con LDA y NAP.

1.4.3. Análisis de Discriminante Lineal (LDA)

El Análisis de Discriminante Lineal (LDA, del inglés Linear Discriminant Analysis) [71], es una técnica de reducción de dimensiones ampliamente utilizada en el reconocimiento de patrones y utilizada en [26] para compensar la variabilidad de sesiones. La idea de este método es obtener nuevos ejes ortogonales para mejorar el poder discriminatorio entre clases. Los ejes encontrados tienen que cumplir los siguientes requerimientos: maximizar la varianza entre las clases y minimizar la varianza intra-clase.

En el caso del LDA se considera como una clase a todos los extractos de voz de un mismo locutor y el problema de optimización del LDA se puede definir con la siguiente razón:

$$J(v) = \frac{v' S_b v}{v' S_w v}. \tag{1.32}$$

La expresión anterior, comúnmente denominada coeficiente de Rayleigh para la dirección del espacio v, representa la información de la razón entre la matriz de covarianza inter-clase S_b y la matriz de covarianza intra-clase S_w en la dirección del espacio definida por el vector-propio v.

Las matrices de covarianza son calculadas como sigue:

$$S_b = \sum_{s=1}^{S} (x_s - \bar{x})(x_s - \bar{x})', \tag{1.33}$$

$$S_w = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (x_i^s - x_s)(x_i^s - x_s)', \tag{1.34}$$

donde $\bar{x}$ es el vector medio global de la población de locutores, o sea el vector medio de todos los i-vectores que representan a la población. Obsérvese que como los i-vectores tienen teóricamente una distribución normal estándar $\mathcal{N}(0, I)$, como se asume en la descripción del espacio de variabilidad total T, entonces el vector medio global es igual al vector nulo. Esta afirmación, plasmada en [26], no suele ser cierta en la práctica como se demuestra en [15] y [14], implicando el necesario cálculo del vector medio global. S es el número de locutores, n_s el número de expresiones de voz correspondientes a cada locutor s, x_s es el vector medio de las expresiones de voz de cada locutor ($x_s = \frac{1}{n_s} \sum_{i=1}^{n_s} x_i^s$), y x_i^s es la i-ésima expresión de voz del locutor s. Obsérvese la matriz de covarianza S_w es equivalente a la definida en la ecuación 1.26.

El propósito del LDA es maximizar el coeficiente de Rayleigh en busca de definir

la matriz de proyección A, la cual está compuesta por los mejores vectores-propios (los correspondientes a los más altos valores-propios). Para lograr la maximización se resuelve el problema del valor propio generalizado,

$$S_w v = \lambda S_w v. \tag{1.35}$$

donde λ es la matriz diagonal de los valores-propios y utilizando los vectores-propios seleccionados se crea la matriz de proyección A. Finalmente la similitud entre dos vectores *x1* y *x2* se formaliza:

$$S_{cos}(x_1, x_2) = \frac{(A'x_1)'(A'x_2)}{||A'x_1||||A'x_2||}. \tag{1.36}$$

1.4.4. Combinación del LDA con WCCN

Una de las desventajas del enfoque WCCN es que se centra en la atenuación de las dimensiones que presentan gran variabilidad intra-clases, pero también puede eliminar parte de la información sobre la variabilidad entre-clases, que se encuentra contenida dentro de las dimensiones atenuadas. En busca de aliviar este problema se aplica un LDA para maximizar la separación entre clases y minimizar la dispersión intra-clases, para luego normalizar los vectores utilizando el WCCN sobre el nuevo espacio de trabajo [26], [45].

Una vez que se ha calculado con LDA la transformación A, puede entonces calcularse la matriz de proyección del WCCN sobre el espacio $A'x$. La similitud entre dos vectores *x1* y *x2* es calculada utilizando la distancia del coseno:

$$S_{cos}(x_1, x_2) = \frac{(A'x_1)'W^{-1}(A'x_2)}{\sqrt{(A'x_1)'W^{-1}(A'x_1)}.\sqrt{(A'x_2)'W^{-1}(A'x_2)}}. \tag{1.37}$$

donde $W = S_w$ definida en la ecuación 1.26.

1.4.5. Análisis de Discriminante Lineal Probabilístico (PLDA): modelos generativos

Una evolución sobre el enfoque i-vector fue introducida en [48] utilizando Análisis de Discriminante Lineal Probabilístico (PLDA, del inglés Probabilistic Linear Discriminant Analysis) [68]. Este nuevo modelo generativo está estrechamente relacionado con la técnica JFA para el reconocimiento del locutor y consiste en una extensión probabilística del método LDA. Sobre este enfoque,

dos principales métodos han sido los más desarrollados hasta el momento en el área, Análisis de Discriminante Lineal Probabilístico Gaussiano (G-PLDA, del inglés Gaussian Probabilistic Linear Discriminant Analysis) presentado en [68] y Análisis de Discriminante Lineal Probabilístico de Cola Pesada (HT-PLDA, del inglés Heavy-Tailed Probabilistic Linear Discriminant Analysis) propuesto en [48], de los cuales describiremos el más utilizado para el reconocimiento del locutor, G-PLDA.

El método generativo G-PLDA, sobre el espacio de los vectores de rasgos, asume que cada vector x de dimensión d perteneciente al locutor s, puede ser descompuesto:

$$x = \mu + \phi y_s + \Gamma z + \epsilon \tag{1.38}$$

En el área del reconocimiento del locutor este modelo consta de dos partes: (i) la parte específica del locutor $s = \mu + \phi y_s$ la cual describe la variabilidad entre clases (locutores) y (ii) el componente del canal o el componente ruidoso $\Gamma z + \epsilon$ el cual es diferente para cada sesión y representa la variabilidad intra-clases. La matriz rectangular ϕ, con r_{voces} columnas ($r_{voces} < d$), proporciona una base para el subespacio de los locutores, generalmente llamado "voces-propias" y la matriz Γ igualmente rectangular, con $r_{canales}$ columnas, proporciona una base para el subespacio de las sesiones, llamado "canales-propios". Las variables y_s y z asumen una distribución normal estándar y el término ϵ sigue una distribución Gaussiana con media cero y una matriz de covarianza diagonal Σ; todas las variables se suponen estadísticamente independientes.

El caso particular de que $r_{canales} = d$ (matriz completa Γ) es equivalente a la versión propuesta en [48], donde los canales-propios son removidos de la ecuación 1.38 y el ruido residual Σ es asumido como una matriz de covarianza completa. El modelo G-PLDA se convierte en:

$$x = \mu + \phi y_s + \epsilon \tag{1.39}$$

donde ϕ es la matriz de voces-propias, y_s el factor del locutor y ϵ es el residuo.

El cálculo de la similitud en los sistema de verificación del locutor que utilizan G-PLDA se lleva a cabo empleando la razón de verosimilitud entre dos vectores. Dado dos vectores $x_{cliente}$ y x_{prueba} la razón de verosimilitud puede ser calculada como sigue:

$$S(x_{cliente}, x_{prueba}) = log \frac{P(x_{cliente}, x_{prueba}|H_1)}{P(x_{cliente}|H_0)P(x_{prueba}|H_0)}, \tag{1.40}$$

donde H_1 denota la hipótesis de que los dos vectores representen al mismo locutor

y H_0 denota la hipótesis de que sean vectores de diferentes locutores [48, 17, 67, 22].

Finalmente, los modelos G-PLDA asumen que los vectores de entrada siguen una distribución Gaussiana, sin embargo, como se demuestra en [48, 14] la estimación ML de los parámetros del PLDA, bajo el supuesto Gaussiano, no logra producir modelos precisos para los i-vectores. Para resolver este problema se introdujo, antes de aplicar G-PLDA, un enfoque sencillo que preserva el supuesto de la distribución Gaussiana, pero introduce una etapa de pre-procesamiento donde se aplica un LDA para reducir la dimensión, y aun más importante, se aplica una normalización de la covarianza intra-clases y de la longitud del vector [31, 14, 13].

1.5. Esquema del sistema de verificación del locutor sobre el marco de los i-vectores

En este epígrafe se describe, a través de esquemas, el sistema de verificación de locutores independiente del texto que integra los métodos de actualidad presentados en este capítulo, el cual puede ser utilizado en aplicaciones reales.

Los métodos existentes en la literatura actual necesitan, antes de realizar el proceso de verificación, entrenar a priori un conjunto de parámetros con sus datos característicos correspondientes, fig. 1.5. En primera instancia se necesita entrenar un UBM, capaz de contener la información de la distribución de las clases acústicas de una población de locutores dada. Luego para el enfoque i-vector se necesita obtener el espacio de variabilidad total T para la representación i-vector de expresiones de voz. Es de señalar que para crear este espacio es necesario grandes volúmenes de datos de voz, que contengan la mayor diversidad de sesión posible. Además para compensar la variabilidad de sesión sobre el espacio de los i-vectores se necesita entrenar la matriz de proyección del método LDA y el modelo PLDA.

A partir de contar con los parámetros necesarios, fig. 1.5, se presenta entonces el esquema del sistema de verificación del locutor independiente del texto.

Obsérvese que si se desea utilizar la medida del coseno en la verificación, se sustituye el PLDA por la similitud por coseno. Además se puede sustituir el método LDA por el método WCCN, el NAP o combinaciones de ellos, siempre con su respectiva proyección.

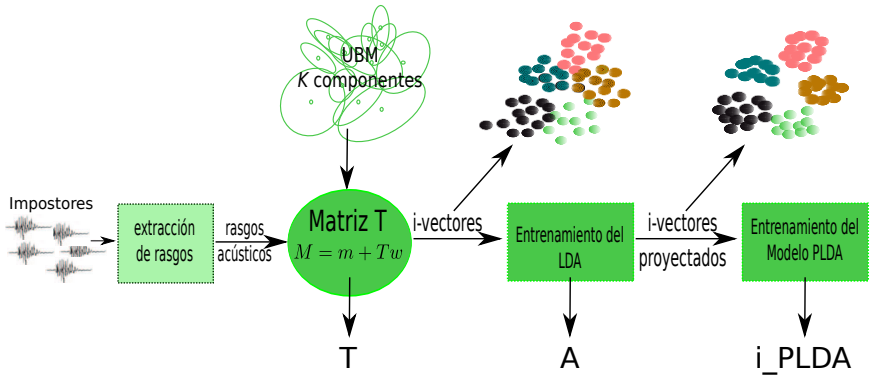


Figura 1.5: Parámetros utilizados en el enfoque de los ivectores

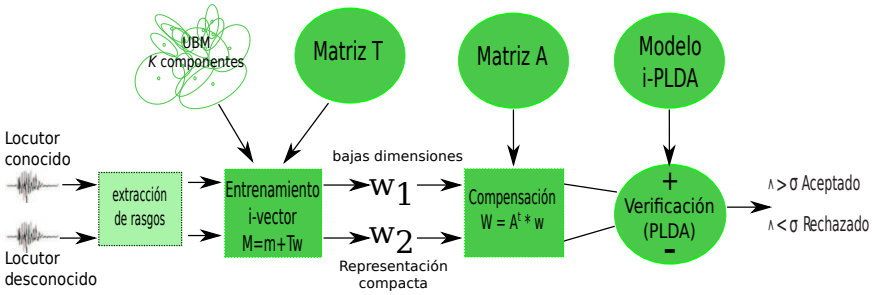


Figura 1.6: Sistema de verificación del locutor

1.6. Conclusiones parciales

Como se ha descrito en el capítulo las técnicas en el área del reconocimiento del locutor han evolucionado notablemente en la última década; siempre basadas en un marco estadístico. Comenzando por los GMM-UBM, luego el enfoque de los supervectores y finalmente el espacio de los i-vectores, dichas técnicas se han ido acercando al problema tradicional de reconocimiento de patrones biométricos.

Cada paso en la evolución ha mejorado la eficacia de los algoritmos, como se ha mostrado en las últimas evaluaciones NIST SRE 2008 [18], 2010 [79] y 2012 [53]. Es de señalar que los sistemas que se basan en la extracción de los i-vectores y que utilizan, como medida de similitud las técnicas PLDA o LDA, son el estado-del-arte en el área del reconocimiento del locutor independiente del texto. No obstante, a pesar de los buenos resultados que estos métodos aún poseen, se han detectado un grupo de desventajas o debilidades en los algoritmos:

- Todos se basan en un marco estadístico, donde la influencia de una información específica se recopila principalmente por la frecuencia de esta información. Todos los enfoques parten del UBM, que contiene la distribución de las clases acústicas

de una población, o sea las características particulares de un locutor que no sean frecuentes dentro de la población tendrán una pobre representación dentro del UBM.

- Los super-vectores presentan gran cantidad de información redundante, la cual provoca sus altas dimensiones.
- Las representaciones de los extractos de voz de cada locutor han sido enfocadas a obtener un vector único, cada matriz de rasgos acústico deriva en un punto en algún espacio. Por lo tanto resulta imposible trabajar con la información secuencial/temporal del habla.
- El cálculo del i-vector presenta alta complejidad computacional $O(W^3+W^2K+WFK)$.
- Se asume que los i-vectores obtenidos del espacio de variabilidad total T sigan una distribución Normal, lo cual no se logra, si observamos que la matriz de covarianza total está lejos de ser la matriz identidad.
- Los supuestos detrás de los modelos generativos GPLDA son:
  - En general asumen que todos los factores son estadísticamente independientes.
  - Se asume que los factores del locutor y de la sesión siguen una distribución Normal y el término residual sigue una distribución Gaussiana con media cero y matriz de covarianza diagonal.

El presente trabajo se encamino a enfrentar las *tres primeras debilidades*.

Capítulo 2

Nuevo enfoque para el reconocimiento del locutor: Marco Binario

En este capítulo se presenta un reciente y nuevo enfoque para el reconocimiento del locutor independiente del texto, partiendo de las desventajas presentadas anteriormente y de la demostración, de forma empírica de la información redundante presente en el espacio de los super-vectores. La idea consiste en desarrollar un método capaz de llevar la información obtenida, del área de trabajo probabilístico continuo a un espacio discreto y binario.

2.1. La información redundante en los super-vectores

La primera tarea consistió en confirmar la necesidad de encontrar una nueva representación. Partiendo de las desventajas ya analizadas del espacio de los super-vectores, se decidió demostrar un nuevo inconveniente, que consiste en la gran cantidad de información redundante presente en este espacio, evaluando en términos de reducción de la dimensión el espacio de los super-vectores. Para ello se propuso reducir las dimensiones del espacio y demostrar que el sistema mantiene similar o mejor eficacia [37].

A partir del estudio realizado de las investigaciones, durante la última década, en el campo del reconocimiento del locutor independiente del texto, se llegó a la conclusión de que se han desarrollado clasificadores estadísticos, discriminatorios y combinaciones de ellos, todos utilizando la misma representación del espacio acústico

y bajo el supuesto de que este espacio es lineal.

Jansen y Niyogi en [44] demostraron que las características acústicas de la voz se encuentran en una *variedad* de baja dimensión que está incrustada en un espacio acústico de alta dimensión, por lo que una variedad de baja dimensión puede tener una estructura altamente no lineal que los métodos lineales no serían capaces de descubrir. Partiendo del supuesto de que los sonidos acústicos, en general, se encuentran en una variedad de altas dimensiones [44] y que el rango de los sonidos acústicos de la voz humana es un subconjunto de todos los sonidos, entonces los sonidos obtenidos de la voz se encuentran en una subvariedad de pocas dimensiones del espacio de altas dimensiones de todos los sonidos posibles. Basado en este razonamiento se propuso utilizar un conjunto de técnicas no lineales de reducción de la dimensión y comparar sus desempeños con los métodos lineales, en busca de demostrar la cantidad de información redundante presente en los super-vectores.

Como se describió en la Subsección 1.2.3, el espacio de los super-vectores se corresponde con el espacio de los rasgos acústicos, cuyos vectores de medias son los centros de las clases acústicas, por lo que se propone descomponer los super-vectores en sus vectores de medias y trabajar sobre el espacio acústico. La reducción de la dimensión se llevó a cabo, principalmente, utilizando los algoritmos Laplaciano [6] y Isomap [84].

2.1.1. Algoritmos clásicos de aprendizaje de variedad

La literatura muestra una serie de algoritmos de aprendizajes de variedad, también conocidos como algoritmos de reducción de dimensiones no lineales, los cuales enfrentan las limitaciones de los métodos lineales. Entre estos algoritmos se destacan el Isomap y el Laplaciano [84] que comienzan creando un grafo para luego utilizar su información geométrica en busca de reducir la dimensión.

El problema general de la reducción de la dimensión consiste en: dado un conjunto de observaciones $X = \{x_1, ..., x_N\}$ de N puntos en $\mathcal{R}^D$ y asumiendo que estas observaciones tienen una dimensión intrínseca d, encontrar un conjunto $Y = \{y_1, ..., y_N\}$ en $\mathcal{R}^d$ (donde $d < D$ y a menudo $d << D$) tal que y_i "represente" a x_i para $i = 1, ..., N$. Es de aclarar, que en términos matemáticos la dimensión intrínseca significa que los puntos del conjunto de datos X yacen en o cerca de una variedad con dimensiones d, que se encuentra incrustada en el espacio D-dimensional.

Método Isomap

La principal asunción que hace el algoritmo Isomap es que la distancia, a lo largo de la curva entre dos puntos, no es la línea recta que los une, sino el camino más corto a través de los puntos de la curva que los conectan. La idea básica consiste en construir un grafo cuyos nodos son los puntos, definidos por los vectores columnas de la matriz X, donde un par de nodos son adyacentes si y solo si los dos puntos están cerca en $\mathcal{R}^D$, para entonces tomar la distancia geodésica entre los dos puntos a lo largo la variedad, como el camino más corto en el grafo. Finalmente a partir de la matriz de adyacencia y conociendo los caminos más cortos entre todos los puntos se utiliza un escalado multidimensional (MDS, del inglés Multidimensional Scaling) – que es un método clásico para obtener información de disimilitud embebida en un espacio métrico [84] – para extraer las representaciones en baja dimensión (como vectores en $\mathcal{R}^d$, $d << D$).

Método Laplaciano

El algoritmo de mapas-propios Laplaciano (del inglés eigenmaps) propuesto por Belkin y Niyogi en [6], se basa en las ideas de la teoría de grafos espectrales.

Dado X una matriz en el espacio $\mathcal{R}^D$, se construye un grafo pesado con N nodos, uno por cada punto, y un conjunto de arcos que conectan los puntos vecinos. Luego se obtiene el mapa embebido que es proporcionado mediante el cálculo de los vectores-propios del grafo Laplaciano, como se explica a continuación:

1. Construcción del grafo de adyacencia representado por la matriz $W_{N,N}$. Se coloca un arco entre el nodo i y el nodo j si x_i y x_j están cercanos. Existen dos variantes:

   - ϵ-vecindad (parámetro $\epsilon \in \mathcal{R}$), el nodo i y j están conectados por un arco si $||x_i - x_j||^2 < \epsilon$, donde la norma es la norma Euclidiana usual en $\mathcal{R}^D$. El peso del arco está dado por $W_{i,j} = exp^{-||x_i - x_j||^2/2\sigma^2}$ y en otro caso $W_{i,j} = 0$.
     - Ventajas: motivación geométrica.
     - desventajas: dificultad para escoger el parámetro ϵ y σ.
   - k vecinos más cercanos (parámetro $k \in \mathcal{N}$), el nodo i y j están conectados por un arco si i está entre los k-vecinos más cercanos de j o j está entre los k-vecinos más cercanos de i (obsérvese que esta relación es simétrica); el peso del arco está dado por $W_{i,j} = 1$ y 0 en otro caso.
     - Ventajas: es un esquema ponderado simple.

- desventajas: es menos intuitivo geométricamente.

2. Obtención de los mapas-propios. Cálculo de los vectores-propios y valores-propios resultantes de resolver el problema del valor propio generalizado $Lv = \lambda Pv$. Donde P es una matriz diagonal de pesos tal que $P_{i,i} = \sum_j W_{i,j}$, obsérvese que W es simétrica por construcción. Finalmente, dado un grafo y su matriz de adyacencia, donde los valores diferentes de ceros son los pesos de los arcos, W, obtenemos la matriz Laplaciana o operador Laplaciano L a partir de $L = P - W$.

Los valores-propios y los vectores-propios del operador Laplaciano revelan una gran cantidad de información acerca de geometría de los datos, en este trabajo se utilizó esta técnica para capturar información local sobre la variedad.

2.1.2. Naturaleza de los super-vectores en el reconocimiento del locutor

Los super-vectores (SV) se construyen por la concatenación de los centros de las componentes Gaussianas de cada modelo del locutor, como se explicó en el epígrafe 1.2.3 y s representa el índice de cada locutor,

$$SV_s = \{x_{1,1}, x_{2,1}, ..., x_{F,1}, x_{1,2}, ..., x_{F,2}, x_{1,K}, ..., x_{F,K}\}.$$

Es necesario notar que como resultado de esta construcción, cada vector de medias define un conjunto específico de dimensiones, o sea cada posición dentro de los vectores de medias define una dimensión del SV y la unión de todos los vectores de medias, define el lugar del punto en el espacio de alta dimensión del SV. Además el orden en que son elegidos los centros de las Gaussianas en la construcción del SV no es importante, siempre que se utilice el mismo orden para crear todos los SV de los locutores.

Este proceso es sólo una simple transformación de la matriz de medias del modelo a un SV con dimensión FK, convirtiendo la matriz μ_s en un punto sobre un espacio de altas dimensiones, el cual es utilizado para clasificar al locutor a través de un clasificador SVM. Para una mejor comprensión, la fig. 2.1 ilustra, desde otro punto de vista la construcción de un SV, en un espacio de dos dimensiones.

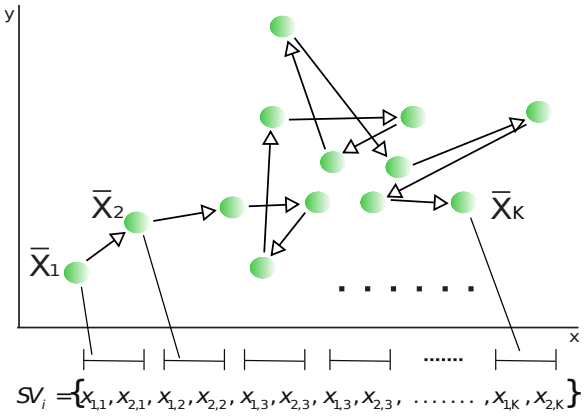


Figura 2.1: Cada punto representa los dos primeros coeficientes de las componentes Gaussianas de un locutor, el cual define sus clases acústicas [37].

La fig. 2.2 ilustra, a manera de ejemplo en las mismas dos dimensiones, la distribución de las componentes Gaussianas de S locutores, donde cada punto es el centro de la componente para un locutor y los colores representan las clases acústicas de la población.

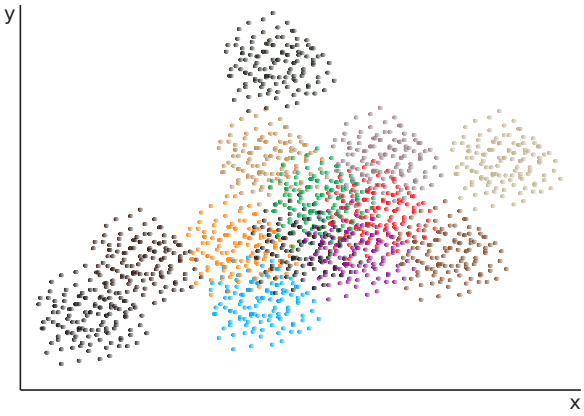


Figura 2.2: Representación artificial en dos dimensiones del espacio acústico [37].

Los investigadores en el área, conocen que existe mucha información redundante en los modelos acústicos del locutor obtenidos a partir de la adaptación GMM-UBM, la cual puede ser observada fácilmente incrementando la cantidad de componentes del UBM. Como resultado se obtendrá una mejora en el EER, que no se justifica con el gran tamaño que alcanzan los SV de cada locutor, lo cual a menudo genera más problemas que los pequeños beneficios en porcentaje de mejora del EER logrado [75]. Si observamos la fig. 2.2, donde los puntos tienen una naturaleza similar al espacio donde yacen las clases acústicas, se puede ver, que en el centro de la nube de puntos

es donde existe el mayor solapamiento entre las clases acústicas y cuanto más grande sea el número de muestras acústicas más pequeña será la distancia entre las clases, aumentando la no linealidad del espacio. Además de este análisis también nos basamos en trabajos anteriores, ver [44], [40]. Por lo tanto se decidió utilizar la información topológica presente en las clases acústicas para reducir la dimensión del espacio y por ende el tamaño del SV, sin afectar en gran medida el resultado del clasificador.

2.1.3. Algoritmo propuesto para reducir la dimensión de los super-vectores

El algoritmo propuesto realiza un análisis global del espacio inicial donde se encuentran las clases acústicas para finalmente realizar la reducción de la dimensión. Se asume que todas las clases acústicas yacen en una variedad y se desea encontrar la información geométrica de la estructura topológica de estas clases acústicas que mejor caractericen al locutor y se conoce que cada punto en el espacio de altas dimensiones se define por el número de componentes Gaussianas.

El algoritmo cuenta de los siguientes pasos:

1. Se obtienen los modelos de los locutores clientes, de los impostores y de los locutores de pruebas utilizando la adaptación GMM-UBM, de los cuales sólo se utilizan las matrices de medias,

$$\{\mu_s\}_{s=1}^{S}, \quad \textit{donde } \mu_s = \left\{ \begin{array}{cccc} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_K \\ \downarrow & \downarrow & \downarrow & \downarrow \\ \left\{ \begin{array}{cccc} x_{1,1} & x_{1,2} & \dots & x_{1,K} \\ \vdots & \vdots & \vdots & \vdots \\ x_{F,1} & x_{F,2} & \dots & x_{F,K} \end{array} \right\} \end{array} \right\},$$

   S es el número de locutores, F la dimensión de las componente Gaussianas y K es el número de Gaussianas del GMM.

2. Entonces, se construye el espacio inicial para las K componentes de todos los locutores en sus F dimensiones.

$$\{A^f_{S,K}\}^F_{f=1}, \quad donde\ A^f_{S,K} = \left\{ \begin{array}{c} \begin{array}{cccc} 1 & 2 & \dots & K \\ \downarrow & \downarrow & \downarrow & \downarrow \end{array} \\ \left\{ \begin{array}{cccc} \mu_1(f,1) & \mu_1(f,2) & \dots & \mu_1(f,K) \\ \vdots & \vdots & \vdots & \vdots \\ \mu_S(f,1) & \mu_S(f,2) & \dots & \mu_S(f,K) \end{array} \right\} \end{array} \right\}$$

es una matriz de $S \times K$ dimensiones y F es el número de espacios construidos. Por cada f se realiza una reducción de la dimensión teniendo en cuenta la información topológica de cada espacio.

3. Por cada subvariedad $A^f_{S,K}$ se obtiene una nueva representación en un espacio lineal $H : \mathcal{R}^K \rightarrow \mathcal{R}^G$, donde intervino la información topológica para la descripción de las nuevas medias. Para obtener la nueva proyección se utiliza cualquiera de las técnicas propuestas en el trabajo, Isomap, Laplaciano u otra, aplicada a cada matriz $H(A^f_{S,K}) = A^f_{S,G}$, donde H es la técnica utilizada y G es la nueva dimensión del espacio, resultando en una reducción del número de las componentes Gaussianas por cada modelo ($G << K$).

4. Luego la matriz de medias de cada locutor es re-ensamblada a partir del nuevo espacio, para finalmente concatenar sus vectores obteniendo un SV de dimensión FG, estos nuevos super-vectores participan tanto en el entrenamiento de la SVM como en la prueba, para cada locutor correspondiente.

2.1.4. Evaluación experimental

Los experimentos desarrollados para evaluar el algoritmo utilizaron las siguientes bases de datos: NIST 2004 para el entrenamiento del UBM, 1348 locutores masculinos de NIST 2005 para los datos de desarrollo con 5 minutos de conversación tanto para el entrenamiento del cliente como para la prueba y 380 locutores de Fisher 2004 como impostores, utilizados para entrenar el clasificador SVM.

La realización del los experimentos de reconocimiento del locutor tuvo lugar en el contexto de la plataforma de código abierto ALIZE [12] y el algoritmo propuesto junto con las técnicas de reducción de la dimensión fue implementado sobre Matlab. El desempeño de los métodos se evaluó utilizando las curvas DET y medidos en términos de EER. La función de costo se calculó siguiendo los criterios de NIST 2006 [69].

Para todos los experimentos se utilizaron los modelos de los locutores clientes, impostores y los de prueba, obtenidos vía la adaptación GMM-UBM. Para el sistema de referencia se obtuvieron los super-vectores de dimensión FK y participaron en el entrenamiento del clasificador SVM los SV de los clientes y los impostores. Para la puntuación final se utilizaron los SV de los locutores de prueba.

Se desarrollaron dos experimentos sobre las competencias NIST (fig. 2.3 y 2.4) que consistieron en la evaluación del desempeño, en términos del reconocimiento del locutor independiente del texto, de la nueva representación del locutor en un espacio de pocas dimensiones. Estas nuevas representaciones fueron obtenidas utilizando el algoritmo propuesto y las técnicas Isomap y Laplaciano, sobre un conjunto cerrado de locutores.

El primero (fig. 2.3) consistió en:

Sistema de referencia, $K = 512$ componentes Gaussianas con $F = 50$ dimensiones para un super-vector por locutor de $FK = 25600$ dimensiones.

Reducción con Isomap, partiendo de tomar los 12 vecinos más cercanos en un espacio de dimensión $F = 50$ y número de componentes $K = 512$, se obtuvo como resultado un nuevo espacio reducido con $G = 128$ componentes Gaussianas con la misma dimensión, para un super-vector por cada locutor de $FG = 6400$ dimensiones.

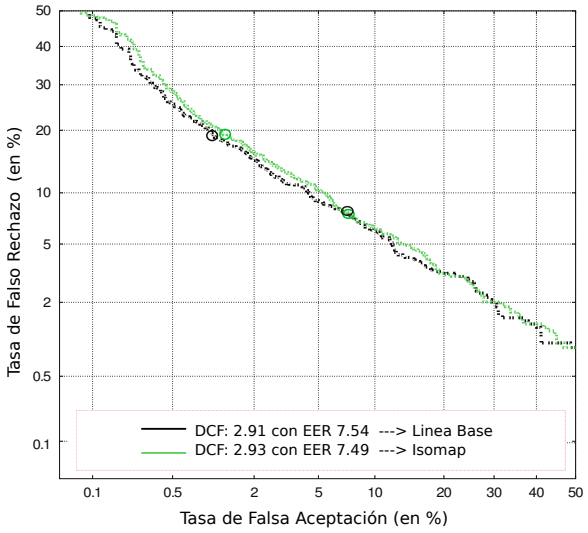


Figura 2.3: Desempeño del algoritmo propuesto utilizando Isomap, frente al sistema de referencia.

En el segundo experimento se incluyó una técnica de reducción lineal de la dimensión, Análisis de Componentes Principales (PCA, del inglés Principal Component Analysis), en busca de evaluar la propuesta frente a una técnica

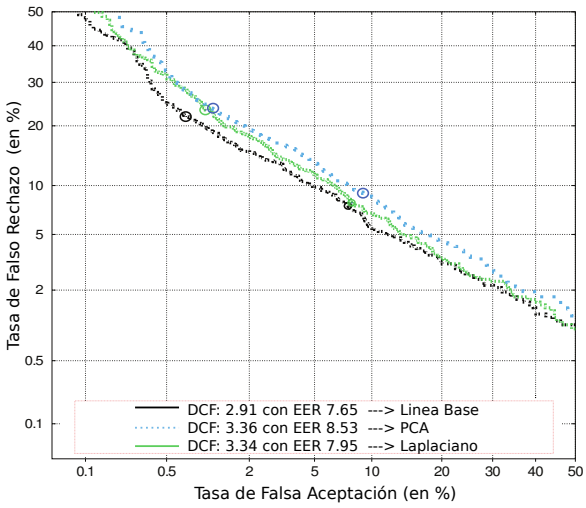


Figura 2.4: Desempeño del algoritmo propuesto utilizando el Laplaciano frente al sistema de referencia y el PCA.

de reducción lineal.

Los parámetros del segundo experimento (fig. 2.4) son:

Sistema de referencia, $K = 128$ componentes Gaussianas con $F = 50$ dimensiones para un super-vector por cada locutor de $FK = 6400$ dimensiones.

Reducción con Laplaciano, partiendo de tomar los 12 vecinos más cercanos en un espacio de dimensión $F = 50$ y número de componentes $K = 128$, se obtuvo como resultado un nuevo espacio reducido con $G = 64$ componentes Gaussianas con la misma dimensión, para un super-vector por cada locutor de $FG = 3200$ dimensiones.

Reducción con PCA, partiendo de un espacio de dimensión $F = 50$ y número de componentes $K = 128$, se obtuvo como resultado un nuevo espacio reducido con $G = 64$ componentes Gaussianas con la misma dimensión, para un super-vector por cada locutor de $FG = 3200$ dimensiones.

El primer experimento utilizando el Isomap refleja resultados muy similares al sistema de referencia, 7.54 % de EER (base) contra 7.49 % de EER y 2.91 % de DCF (base) por un 2.93 %.

Aunque no se obtuvieron mejoras, este experimento está enfocado a la reducción de las dimensiones del espacio de los super-vectores, lográndose una significativa reducción que muestra la información redundante que existe en el espacio original de los super-vectores. Si se compara la dimensión de los SV en el primer experimento, el sistema de referencia cuenta con 25600 dimensiones y la propuesta de reducción con Isomap tiene 6400, lo que representa 1/4 del tamaño original, resultando una reducción apreciable manteniendo casi la misma eficacia.

En el segundo experimento, el sistema de referencia tiene 6400 dimensiones y la propuesta de reducción con Laplaciano 3200, la mitad del tamaño, resultando una valiosa reducción también, aunque en este caso se pierde algo de eficacia.

Obsérvese como la reducción lograda con PCA, aunque tienen el mismo nivel, introduce una mayor pérdida en la eficacia.

2.1.5. Conclusiones parciales

Este resultado constituye uno de los primeros pasos en la utilización de la información topológica en el campo de reconocimiento del locutor. Con el fin de evaluar la presencia de información redundante en el espacio de los super-vectores se trató de reducir la dimensión del espacio, para ello se utilizó el algoritmo propuesto basado en las técnicas Isomap y el Laplaciano. Los resultados mostraron claramente la información redundante presente en los super-vectores, con una reducción de la dimensión en un factor de cuatro (de 25600 a 6400) prácticamente sin pérdida, en términos de rendimiento. Además, el enfoque no lineal (Laplaciano) superó el resultado obtenido a partir de una técnica lineal clásica (PCA), lo que muestra la importancia de tener presente la naturaleza interna de los datos como lo hace el enfoque topológico.

Este resultado [37] mostró la gran cantidad de información redundante dentro del espacio de los super-vectores GMM, lo cual representa una *nueva desventaja* de los sistemas actuales y conllevó a la necesidad de buscar nuevas alternativas para el reconocimiento del locutor.

2.2. Nueva Representación: Matriz binaria

Con la intención de reducir las desventajas presentadas en el espacio probabilístico ha sido necesario desarrollar un método capaz de llevar la información contenida en la voz del área de trabajo probabilística continua a un espacio discreto y binario. Recientemente fue propuesto en [1] un nuevo enfoque binario capaz de representar mejor las características discriminatorias del locutor a partir de una expresión de voz. Este nuevo enfoque ha continuado su desarrollo en el campo de la diarización del locutor [63] y en la verificación del locutor [10, 11].

Basándonos en la propuesta inicial de una representación binaria, se ha desarrollado una nueva transformación del espacio probabilístico a un espacio discreto y binario que refleja nuevas características discriminatorias de la voz [36].

2.2.1. Modelo Generador y Matriz binaria

El objetivo principal de la técnica de modelado de una expresión de voz del locutor es obtener, a través de una transformación –acústica a binaria– (Modelo Generador (GM, del inglés Generator Model)), una representación compacta de un locutor en un espacio binario, donde se expresan nuevas características discriminatorias entre los locutores, con respecto al espacio acústico. Una vez en el espacio binario, varias aplicaciones han sido desarrolladas para beneficiarse de la facilidad de comparación, reducción de almacenamiento y otras propiedades de los vectores binarios. Es de señalar que el GM es entrenado, una sola vez, *a priori* durante la fase de desarrollo.

La lógica detrás del modelado se basa en el aumento de la capacidad discriminatoria de un UBM; el nuevo modelo se compone de las componentes Gaussianas de un UBM clásico que representan las clases acústicas de la población y de un conjunto de mono-Gaussianas que yacen dentro de cada clase acústica. Implicando que el método de modelado binario de los locutores esté diseñado en torno a dos niveles principales.

En el primer nivel se encuentra el UBM, que juega un rol estructural, definiendo una agrupación solapada del espacio acústico en regiones acústicas particulares, cada una de las regiones o clases es asociada a un componente del UBM. Para el entrenamiento del UBM se utiliza el algoritmo EM y un amplio conjunto de rasgos acústicos obtenidos de impostores; luego se realiza una selección de las componentes Gaussianas, que consiste en elegir aquellas componentes del UBM que con mayor probabilidad modelen cada trama al menos una vez, según la ec. 1.4, frente a los mismos datos utilizados para entrenar el UBM. Obsérvese que podría parecer lógico que fueran seleccionadas todas las Gaussianas, lo cual no suele suceder producto de la naturaleza de los GMM, pueden existir componentes con muy bajo peso que estén solapados completamente por otro componente con mayor peso, siendo estos últimos los que sobresalen al calcular las probabilidades de los rasgos que se encuentren en la región cubierta por dichas componentes, ver fig. 1.2. Finalmente, asumiendo que fueron seleccionadas mucho menos componentes que los originales, se ajustan los pesos del nuevo modelo y se crea una nueva matriz de covarianza donde estarán solamente los vectores de varianzas correspondientes a las Gaussianas elegidas, resultando en un nuevo UBM con menos componentes.

Esta selección, en busca de eliminar las Gaussianas con poca o ninguna información discriminatoria del locutor, es una de las diferencias importantes de método propuesto frente a otras variantes existentes para crear el modelo generador [1, 10, 11].

En el segundo nivel se obtienen las mono-Gaussianas, denotadas por "especifici-

dades" o puntos característicos las que están dedicadas a reflejar la nueva información discriminatoria de los locutores. Las especificidades son obtenidas a partir de los modelos adaptados de los mismos impostores utilizados en el primer nivel, para ello se utiliza el método de adaptación MAP, sobre el UBM resultante del primer nivel y los rasgos acústicos de los impostores. Luego se crea un conjunto con los vectores de medias de los nuevos modelos de los locutores y utilizando este conjunto como nuevos rasgos acústicos (obsérvese que estos vectores son los centros de las clases acústicas por locutores), se procede a vincular, utilizando la función de densidad Gaussiana descrita en la ec. 1.4, cada vector con aquella Gaussiana del nuevo UBM que represente la región del espacio donde se encuentra el vector.

Este proceso obtiene gran cantidad de vectores vinculados a cada componente del nuevo UBM, lo que provoca tan altas dimensiones que imposibilita su uso. Para resolver este problema se entrena un GMM a partir de cada subconjunto de vectores vinculado a cada componente del UBM, obteniéndose un GMM por cada componente del UBM. Finalmente, se realiza el mismo proceso de selección utilizado para obtener el nuevo UBM, por cada subconjunto de especificidades, en busca de eliminar las especificidades que no aportan información discriminatoria dentro de la clase acústica. Este proceso genera por cada componente del UBM un subconjunto de mono-Gaussianas que describe, en términos de la información característica del locutor, la región correspondiente a la clase acústica. Notar que cada subconjunto puede diferir en la cantidad de mono-Gaussianas y la cantidad de ellas está estrechamente relacionada con la naturaleza de la región acústica que se describe, lo que representa otra importante diferencia frente a los otros métodos existentes [1, 10, 11].

Concluimos, con un GM que consta de un UBM con K componentes Gaussianas y E especificidades (la suma de la cantidad de especificidades en cada región). Aunque resulte un poco complejo el entrenamiento del GM, es de destacar que es un proceso que se realiza completamente a priori, y utilizarlo resulta muy simple, como se muestra en la fig. 2.5.

El objetivo de este nuevo Modelo Generador consiste en la obtención de una representación binaria para cada expresión de voz, que represente las mejores relaciones entre cada vector de rasgos acústicos y su grupo de especificidades, organizado según las componentes del UBM. La transformación, como se ilustra en la fig. 2.5, se define como $\varphi : \mathcal{R}^F \rightarrow \mathcal{N}^E$, entre un vector acústico de dimensión F y un vector binario de dimensión E, donde E se corresponde con el número de especificidades en el GM ($E >> F$) y cada posición del vector binario está vinculada a una especificidad. Esta

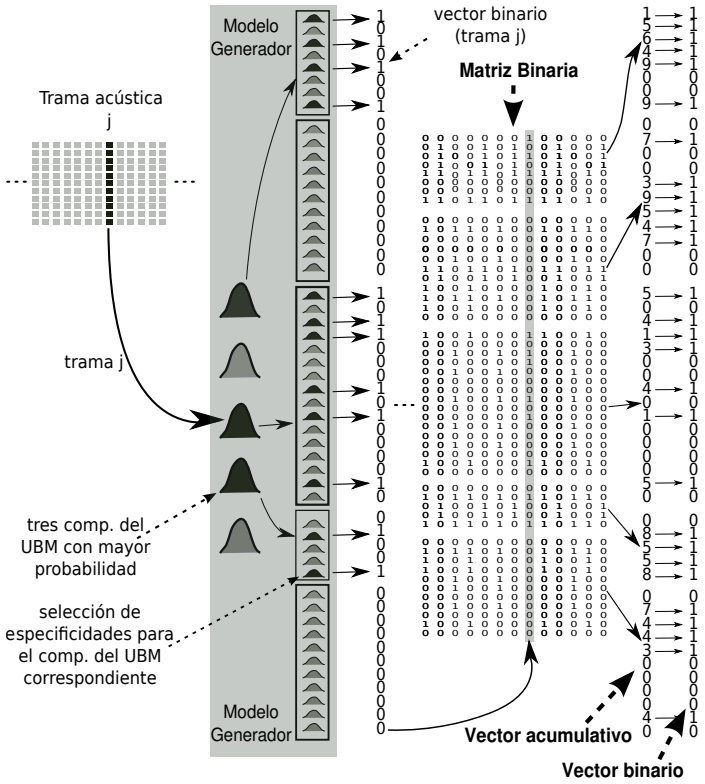


Figura 2.5: Pasos para obtener la matriz binaria, el vector acumulativo y el vector binario

transformación, proyecta individualmente cada trama acústica al espacio binario, en general las posiciones seleccionadas son etiquetadas a 1 en el vector binario indicando que la mono-Gaussiana correspondiente fue "activada" por la trama acústica. La selección de las posiciones se realiza mediante un cálculo de probabilidades, primero se seleccionan las 3 componentes del UBM (cantidad establecida de forma empírica) con mayor probabilidad dado el vector acústico, este proceso implica el cálculo de $p_k = P(k|x_l, \lambda_{UBM})$, la probabilidad de todas las componentes del UBM (K) por cada trama (x_l), donde L es la cantidad total de tramas. Como resultado el número de operaciones del proceso es $K \times L$ y su complejidad computacional $\mathbf{O(KL)}$. Luego dentro de cada una de las 3 componentes seleccionadas y utilizando las probabilidades de las mono-Gaussianas correspondientes dado el vector, son seleccionadas aquellas especificidades con mayor probabilidad que un umbral dado (0.001 calculado de forma empírica), el resto son etiquetadas a 0. Las especificidades seleccionadas se consideran activadas y el número de posiciones establecidos a 1 por cada vector binario guardan relación dinámica con las características acústicas del vector de rasgos, lo que constituye la principal diferencia antes los enfoques propuestos en [1, 10, 11]. Considerando

a C la cantidad máxima de mono-Gaussianas dentro de las componentes del UBM, el segundo proceso necesita realizar $3 \times C \times L$ operaciones para una complejidad computacional de $\mathbf{O(CL)}$.

Como resultado, al aplicar el GM, se obtiene por cada expresión de voz una matriz binaria dispersa (BK, del inglés Binary Key) [1] de dimensión $E \times L$, donde E es el número de especificidades y L la cantidad de tramas acústicas de un extracto de voz, esta matriz binaria es una representación "exhaustiva" en el tiempo de la señal en el nuevo espacio binario [36].

Con esta transformación, es posible obtener información espacial sobre las características acústicas, la cual no es posible con los métodos clásicos como el GMM, JFA o i-vector. Además es fácilmente apreciable que la nueva representación binaria contiene los eventos acústicos (sin enfatizar en la frecuencia) existentes en la señal de un locutor, tanto frecuentes como pocos frecuentes.

Como conclusión, la matriz binaria de una expresión de voz presenta varias características importantes.

1. Es capaz de capturar los eventos, independientemente de su frecuencia, presentes en la matriz de rasgos acústicos del locutor.

2. Posibilita obtener un solo vector por locutor permitiendo un modelado directo de la compensación de la variabilidad de sesión.

3. Mantiene la información temporal presente en la matriz de rasgos acústicos.

4. Es una matriz dispersa permitiendo un fácil y rápido trabajo sobre ella.

5. El tamaño para almacenar la información es sumamente pequeño teniendo en cuenta que el valor medio de la cantidad de especificidades activadas por vector binario es 8.

6. Este enfoque puede ser utilizado en cualquier otra área de procesamiento de la voz.

7. Los dos procesos que implica la transformación para obtener la matriz binaria, presentan una complejidad computacional similar a la extracción de las estadísticas de Baum-Welch en el enfoque i-vector.

2.2.2. Vector Acumulativo y Vector Binario

El proceso para obtener la información global de la matriz binaria es muy simple y consiste en contar las etiquetas 1 por cada fila de BK, lo que implica como resultado un vector de dimensión E que contienen el nivel de activaciones de cada especificidad, este nuevo vector es denotado por "modelo general", "vector global" [36] o como lo llamaremos en este trabajo, Vector Acumulativo (CV, del inglés Cumulative Vector). Formalmente el vector acumulativo de un extracto de voz se define por,

$$CV[i] = \sum_{l=1}^{L} BK(i,l), \tag{2.1}$$

donde $i = 1, ..., E$ y $CV \in \mathcal{N}$. Con esta representación compacta de BK se logra tener presente la información de los eventos, independientemente de su frecuencia, que contiene la expresión de voz de un locutor dado. Ver figura 2.5. El proceso de extracción del CV realiza $E \times L$ operaciones para una complejidad computacional de **O(EL)**.

Más sencilla es aún la representación de la expresión de voz por un Vector Binario (BV, del inglés Binary Vector), obtenido al establecer en 1 los valores del CV diferentes de 0, resultando en un BV de las mismas dimensiones que CV. Ver figura 2.5.

Notar que existe una alta relación entre BV y CV, el primero refleja las especificidades que mejor representan a la expresión de voz, mientras el segundo pesa la influencia de cada especificidad (el nivel de activación) [11].

Por tanto, una expresión de voz está representada por dos vectores simples y dispersos que componen una nueva representación del locutor, causando una reducción drástica del volumen de información necesaria para el reconocimiento del locutor en comparación con las GMM.

Finalmente, como los valores de los CV de cada expresión de voz, son el nivel de activación de cada especificidad en la expresión completa, poseen una estrecha relación con la duración de la señal, se requiere realizar una normalización de cada CV para eliminar el efecto de la duración, como se define a continuación,

$$CV = \frac{CV}{||CV||}, \tag{2.2}$$

Esta normalización asegura que todos los vectores tengan longitud uno, lo cual es deseado para los clasificadores en general, ajustando así los valores de los CV

sin importar la duración de la señal. Esta normalización es aplicada para todos los vectores acumulativos en la investigación realizada.

2.3. Medida de similitud: Intersección y Diferencia Simétrica

La nueva representación binaria de la voz involucró la necesidad de encontrar un nuevo método para medir la similitud entre dos locutores. Para resolver este problema se utilizaron dos operaciones de conjuntos, la intersección y la diferencia simétrica, para comparar dos locutores A y B, donde cada locutor a evaluar cuenta con el BV y CV. Para describir la utilización de las operaciones de conjunto partiremos de la fig. 2.6.

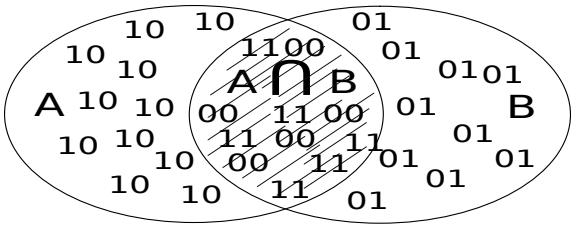


Figura 2.6: Operaciones de conjuntos entre dos BV.

$A \cap B = \{x/x \in A \ y \ x \in B\}$	A intersección B, tendrá los elementos comunes a los dos conjuntos, {1,1} y {0,0}.
$A \triangle B = (A - B) \cup (B - A)$ $A \triangle B = \{x/x \in A \cup B \wedge x \notin A \cap B\}$	La diferencia simétrica es la unión sin la intersección, los elementos no comunes entre los dos conjuntos {1,0} and {0,1}.

Obsérvese, los elementos {0,0} en la intersección no fueron utilizados en este trabajo, aunque pueden ser muy interesantes por identificar las especificidades no activadas pero si *comunes* entre dos locutores.

2.3.1. La intersección como medida de similitud

En [1] y [11] los autores proponen una medida de similitud que refleja el tamaño de la intersección, definida por la cantidad de pares {1,1}.

$$IS(A,B) = \frac{|A \cap B|}{|A|}, \qquad |A| = |B| \tag{2.3}$$

Esta medida funciona sólo en los vectores binarios y se corresponde con el número de componentes activos comunes en ambas representaciones y se denota como Similitud de la Intersección (IS, del inglés Intersection Similarity). La medida IS presenta algunas ventajas con respecto a la clásica ML como son: es fácil de programar, su costo computacional es muy bajo y permite caracterizar cada expresión de voz con un simple vector binario, reduciendo drásticamente el tamaño en memoria.

Pero también presenta desventajas: utiliza sólo el vector binario como una simplificación del vector acumulativo, lo que provoca una pérdida de información de la expresión de voz del locutor, como es, por ejemplo, el nivel de activación de las especificidades.

2.3.2. Nueva similitud empleando la intersección y la diferencia simétrica

Se denota como Similitud de la Intersección y la Diferencia Simétrica (ISDS, del inglés Intersection and Symmetric Difference Similarity), una nueva medida que es manejada por el BV de A y B [36], pero aplicada en los correspondientes CV. El uso de los CV para obtener una similitud entre dos expresiones de voz, incorpora nueva información en la comparación, porque los vectores acumulativos no sólo contienen la selección de las mejores especificidades (las cuales modelan la expresión de voz), sino que también contienen la información de cuánto influyen en el modelo dichas especificidades. Esta característica es la diferencia principal de esta nueva propuesta con respecto a [1] y [11].

La medida ISDS se basa en dos términos independientes que utilizan la intersección y la diferencia simétrica entre dos conjuntos. Estos dos términos pueden ser vistos como medidas de similitud por separado; el primer término refleja la similitud que existe en la intersección de los dos conjuntos, con respecto a sus valores acumulativos y se define como sigue:

Dados dos representaciones de locutores A y B, obtenidos vía el GM, se utilizan sus BV para manejar la operación de conjunto, como sigue,

$$ID(A,B) = \frac{1}{\sum_{i=1}^{|A\cap B|} |a_i - b_i|}, \tag{2.4}$$
$$\{\forall a \in A, \forall b \in B \mid \exists (a,b) \in A \cap B \ \ and \ \ a \neq b\},$$

donde a_i y b_i son los valores acumulados de la i-ésima especificidad. Esta medida se encarga de comprobar la similitud que existe en la intersección de los CV.

El segundo término consiste en la razón entre la importancia de la intersección de los vectores y su diferencia simétrica, la cual se basa en las especificidades activadas que difieren entre ellos. Derivando en la suma de los valores acumulativos de los elementos que se encuentran en la intersección divida entre la suma de los valores acumulativos que se encuentran en la diferencia simétrica de las representaciones A y B.

$$ISD(A,B) = \frac{\sum_{i=1}^{|A\cap B|} a_i + b_i}{\sum_{j=1}^{A-B} a_j + \sum_{j=1}^{B-A} b_j} \tag{2.5}$$
$$\{\forall a \in A, \forall b \in B \mid A - B \neq \phi\}$$

Note que, dada la naturaleza de los conjuntos, donde se asegura que todos tengan el mismo número de elementos, los casos $A \supset B$ o $B \supset A$ no existen, por lo tanto $A - B \neq \phi \Leftrightarrow B - A \neq \phi$.

Por último, la multiplicación de los términos (ec. 2.4 y ec. 2.5) conforma la nueva medida de similitud ISDS.

$$ISDS(A,B) = \frac{\sum_{i=1}^{|A\cap B|} a_i + b_i}{(\sum_{j=1}^{A-B} a_j + \sum_{j=1}^{B-A} b_j) * \sum_{i=1}^{|A\cap B|} |a_i - b_i|} \tag{2.6}$$
$$\{\forall a \in A, \forall b \in B \mid A - B \neq \phi \; and \; \exists a \neq b \mid (a,b) \in A \cap B\}$$

Obsérvese que la nueva medida es fácil de programar y tiene un bajo costo computacional. A continuación se proponen los pasos a seguir en un algoritmo que conlleva a la verificación del locutor.

Sea A el locutor cliente y B el locutor de prueba desconocido. Cada locutor contiene su matriz de rasgos X^A y X^B.

Algoritmo 1

1. Obtener la representación binaria de cada uno, a través del modelo generador. $BK^A = GM(X^A)$ y $BK^B = GM(X^B)$.

2. Obtener su vector acumulativo y binario, $\{CV^A, BV^A\}$ y $\{CV^B, BV^B\}$. Denotaremos por GT a los vectores $\{CV, BV\}$ de un locutor.

3. Obtener la puntuación de la verificación entre el locutor cliente y el desconocido:

$$S(GT^A, GT^B) = ISDS(GT^A, GT^B) \tag{2.7}$$

4. Si $S \geqslant \theta$ se acepta, sino se rechaza.

2.3.3. Evaluación experimental

Con el fin de probar la viabilidad del modelo generador, la nueva representación binaria, y la nueva medida de similitud propuesta para la tarea de verificación del locutor se realizaron dos experimentos. El primero consistió en una verificación del locutor dependiente del texto y utilizando señales microfónicas leídas a diferentes velocidades. El segundo se realizó independiente del texto, con voz espontánea y con señales telefónicas.

Experimento de reconocimiento del locutor dependiente del texto

En el primer experimento se utilizó la base de datos SALA [64] para crear un UBM inicialmente de 512 mezclas, para ello se emplearon 1990 secuencias de dígitos de 500 locutores, que en su conjunto se ensamblaron para un total de 2.5 horas de voz. Finalmente, el UBM utilizado en el modelo generador se compone de 459 componentes, seleccionadas del modelo inicial y 22000 mono-Gaussianas. En estos experimentos se utilizaron, como rasgos acústicos, 12 MFCC + 12 deltas con normalización cepstral de la media y la varianza.

La verificación del locutor se desarrolló utilizando las sesiones microfónicas de la base de datos Ahumada (Nist 2001) [65] nombradas M1, M2, M3, M4, M5 y M6, cada sesión cuenta con 100 locutores pronunciando las mismas expresiones de voz, pero leídas a tres velocidades: normal, suave y rápido. Las expresiones de la sesión M1, con una velocidad de lectura normal, son utilizadas como el conjunto de locutores conocidos (clientes) y las restantes sesiones como conjunto de prueba, con tres velocidades de lectura diferentes, para cada velocidad de lectura, en general, se realizan 10000 comparaciones, 100 como verdaderas y 9900 como impostores.

Este experimento se centró en probar la robustez de la representación binaria obtenida vía el modelo generador y evaluar las medidas IS e ISDS. La fig. 2.7 muestra la evaluación siguiendo los pasos descritos en la Subsección 2.3.2, para ambas medidas IS e ISDS, los resultados se presentan en términos de EER para cada sesión de pruebas y se encuentran agrupados por velocidad de lectura de la expresión (normal, suave y rápido).

Este resultado muestra en primera opción, que la nueva representación binaria del locutor obtenida por la transformación con el GM propuesto en este trabajo es capaz de discriminar locutores, mostrando un EER promedio de 1.42 % cuando se utiliza

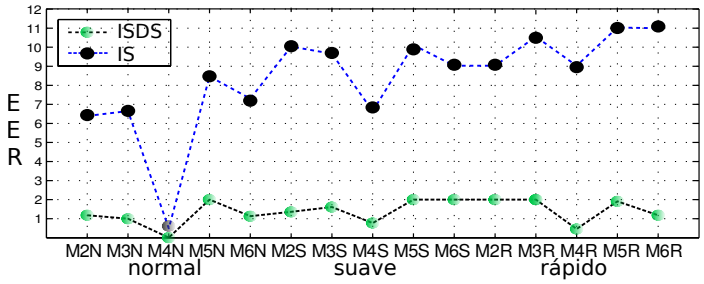


Figura 2.7: Verificación del locutor con diferentes velocidades de lecturas.

la medida ISDS. Es importante observar la estabilidad que se obtiene con la nueva representación ante la variabilidad de la velocidad de lectura, el cual provoca por lo general pérdidas de eficacia en el reconocimiento del locutor.

Respecto a la evaluación del funcionamiento de ambas medidas, se puede apreciar la importancia de la inclusión del nivel de activaciones de las especificidades, a través del CV, en la representación y su uso en la medida ISDS, aportando una mejora en la eficacia, como promedio para todas las velocidades, del 82.9 % de EER. Finalmente, este experimento refleja como la incorporación de los CV incrementa el poder discriminatorio de la representación de la voz del locutor, en relación con el vector binario.

Experimento de reconocimiento del locutor independiente del texto

En el segundo experimento, se realizó la verificación del locutor independiente del texto entre las sesiones telefónicas T1 y T2 de la base de datos Nist 2001 Ahumada. Cada una de las sesiones cuenta con 100 locutores, con alrededor de 1.5 minutos de voz espontánea; las expresiones de T1 se utilizan como conjunto de locutores conocidos (clientes) y las expresiones de T2 como conjunto de locutores desconocidos, para un total de 10000 comparaciones, 100 pruebas como verdaderas y 9900 como impostores. Este experimento se concentró en probar la robustez del GM y comprobar de nuevo la medida IS contra la medida ISDS pero para voz espontánea, además se utilizó como sistema de referencia el enfoque GMM-MAP [75] sobre la plataforma ALIZE [12], (ver Anexo B).

La fig. 2.8 presenta los resultados de este experimento en términos de EER y DCF, donde además se presentan las ventajas de utilizar los CV ante los BV.

Estos resultados fueron publicados en [36] donde se observa que el GM es capaz de obtener una representación binaria discriminatoria entre locutores, con una eficacia comparable (0.01 % de EER de diferencia) a la obtenida con una representación GMM-

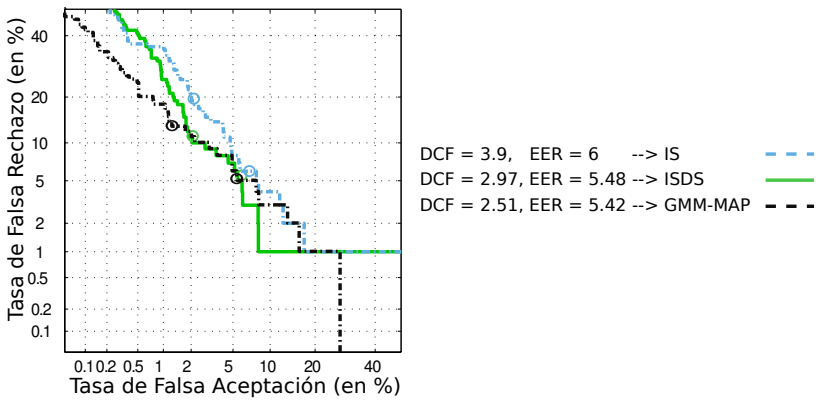


Figura 2.8: Verificación del locutor independiente del texto.

MAP, pero con una eficiencia computacional mucho mayor, al evaluarse las distancias entre dos vectores. Por otra parte se pudo ratificar que la utilización del nivel de activación de las especificidades con el CV, continúa presentando una eficacia mayor frente al uso del BV y de la correspondiente medida IS.

2.3.4. Conclusiones parciales

El contenido de los epígrafes 2.2 y 2.3 presenta los resultados obtenidos a partir de un nuevo enfoque basado en una representación binaria de la expresión de voz del locutor, donde se propuso una nueva medida de similitud asociada con la representación global (vector acumulativo) de la información existente en la matriz binaria, que tiene en cuenta los eventos (desde los más frecuentes hasta los pocos frecuentes en una señal de voz). Por otra parte, en comparación con los métodos reportados en la literatura actual, esta nueva forma de representación y la medida de similitud propuesta, requieren considerablemente menos recursos de cómputo y memoria, mostrando un nivel de rendimiento comparables con el enfoque GMM-MAP.

Podemos resaltar como conclusiones:

1. La nueva representación binaria del locutor, obtenida del nuevo Modelo Generador propuesto, contiene información altamente discriminatoria del locutor comparable con el sistema de referencia GMM-MAP, teniendo en cuenta los eventos discriminatorios existentes en la voz, independientemente de su frecuencia.

2. La medida de similitud propuesta ISDS presenta mejor eficacia que la medida IS.

3. La nueva representación aunque obtiene un EER comparable, sólo un 0.01 % mayor, al sistema de referencia GMM-MAP, requiere de mucho menos recursos computacionales O(EL), siendo más eficiente.

Además de estas ventajas, este enfoque abrió nuevos caminos para continuar las investigaciones, en el enfrentamiento a problemas de variabilidad de sesión aprovechando la representación binaria dispersa y en la explotación de la información temporal existente en ella. Estos nuevos puntos de vistas fueron enfrentados, los nuevos métodos obtenidos y sus resultados serán expuestos en el próximo capítulo.

Capítulo 3

Compensación de la Variabilidad de Sesión e Información Temporal sobre el enfoque binario

Este capítulo trata el tema de la compensación de la variabilidad sesión sobre la representación propuesta y sobre la obtención de la información temporal en el reconocimiento del locutor. Finalmente se presenta el esquema del algoritmo de verificación del locutor que contiene los métodos propuestos.

3.1. Compensación de la variabilidad: Información común

Los algoritmos de compensación de la variabilidad de sesión necesitan grandes cantidades de datos para lograr encontrar el comportamiento de la dispersión intra-clases y entre clases (locutores), y si además, los datos presentan altas dimensiones entonces se necesita tener en cuenta los recursos computacionales.

El espacio de trabajo binario presenta altas dimensiones, para obtener el vector acumulativo (CV) representativo de un locutor se realiza un sencillo proceso de conteo de las etiquetas en "1" por cada especificidad. Esto implica que la dimensión de los CV es igual a la cantidad de especificidades en el modelo generador (GM). Para mitigar este problema se propone una máscara capaz de seleccionar las especificidades que más información discriminatoria aportan para el reconocimiento.

Por otra parte, vale la pena señalar que el GM puede contener información común

o redundante, por ejemplo especificidades redundantes. Esta característica no deseada podría estar vinculada a dos causas diferentes.

En primer lugar, la estructuración básica del espacio acústico se realiza gracias al UBM, un GMM entrenado utilizando el algoritmo de Maximización de la Esperanza y el criterio de máxima verosimilitud. Dicho método es bien conocido por los investigadores en el área [37] que incorpora cierta redundancia, o –en términos de representación espacial– acepta un solapamiento entre sus clases (componentes Gaussianos). Además el algoritmo de selección de las mono-Gaussianas, dentro de una región dada, podría aceptar también cierta redundancia o solapamiento.

En segundo lugar, todas las representaciones propuestas contienen en mayor o menor medida, información común a todos los locutores, como puede ser la relacionada con la influencia de un mismo genero en el timbre y tono de la voz, el uso de un mismo idioma y su fonética común, etc. Dicha información común constituye información redundante. Para enfrentar este problema se propone incorporar dicha información común en los métodos de compensación de la variabilidad de sesión.

3.1.1. Selección de Especificidades: Máscara

La máscara para los vectores acumulativos, propuesta a continuación, está enfocada a reducir la dimensión al descartar las especificidades con poca o ninguna información (varianza) dentro del espacio de los CV [39].

El proceso para obtener la máscara consiste en un algoritmo de selección, basado en las especificidades con poca o ninguna variación dentro de la población que no aportan información discriminatoria dentro del CV, por lo que su aporte en la comparación de dos vectores se considera nula o negativa.

Dado $X = [x_1, x_2, ..., x_S]$ una matriz, donde sus columnas son los CV correspondientes a los locutores de una población de S impostores y las filas sus E especificidades. Entonces la máscara es obtenida por:

$$mask_j = \begin{cases} 1 & \text{if } \frac{1}{S}\sum_{i=1}^{S}\left(X_{j,i} - \overline{x}_j\right)^2 \geqslant \theta \\ \\ 0 & \text{otro caso} \end{cases}$$

donde $\bar{x_j}$ es el valor medio de la especificidad correspondiente para $j = 1, 2, ..., E$, θ es el umbral de selección.

Como resultado se obtiene una máscara definida por un vector booleano donde los coeficientes con varianza mayor que el umbral dado son etiquetados en “1” y “0”

en el resto.

El umbral de la varianza θ fue calculado de la siguiente forma:

1. Se obtiene el valor de la varianza por cada especificidad en la población, almacenándose en un vector de la misma dimensión que el CV.

2. Se ordenan los índices de las especificidades de forma descendiente respecto a sus valores de varianza.

3. Se escoge como umbral inicial el primer valor de varianza mayor que 0.

4. Las especificidades con valor de varianza mayor que el umbral son seleccionadas.

5. Utilizando solamente las especificidades seleccionadas de los CV, se realiza un experimento de reconocimiento del locutor empleando la medida ISDS propuesta. El valor del EER y el umbral de varianza son almacenados.

6. Aumentando el valor del umbral se repite el paso 4 y 5 hasta que el EER sufra un aumento significativo, finalmente se selecciona como umbral el valor de varianza que refleje el valor mínimo de EER.

La figura 3.1 muestra el proceso de obtención del umbral de varianza.

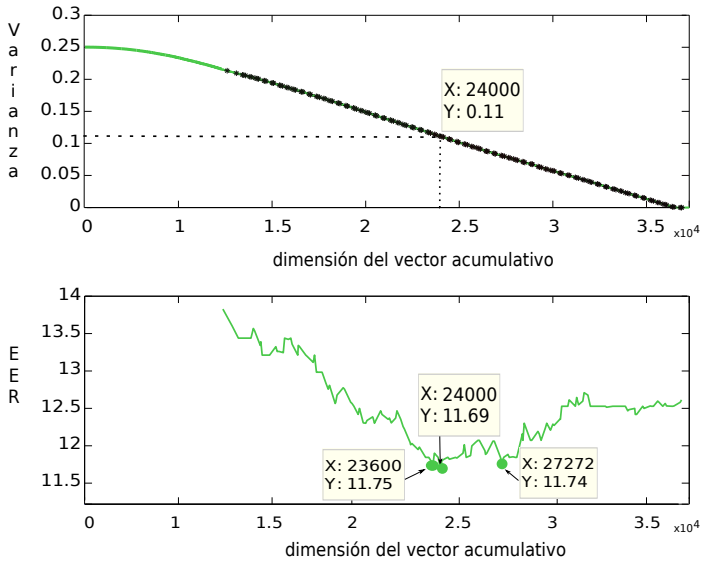


Figura 3.1: En la imagen superior se muestra la curva con los valores de varianza así como el umbral seleccionado, en la imagen inferior la curva con los valores de EER para los tres probables umbrales.

Como resultado del proceso se obtuvo una máscara binaria de E elementos donde se etiquetaron con "1" las k especificidades con varianza mayor que el umbral θ, $k < E$.

Algo interesante sucedió en este proceso, al contabilizar y promediar el comportamiento (número de activaciones) de cada especificidad en la población general, se observó que no sólo las especificidades con muy bajo nivel de activación y muy poca varianza fueron eliminadas en la máscara, sino que existieron muchas especificidades con alto nivel de activación y poca varianza que también se eliminaron. Se considera que este hecho está dado por la naturaleza de los GMM, porque existen componentes Gaussianos con gran varianza que sobresalen dentro de los demás, cubriendo varias clases acústicas, lo que implica que en la mayoría de los locutores provoquen altos niveles de activaciones. Este fenómeno causa, que estas especificidades sean eliminadas de la máscara, pero lejos de perjudicar el proceso reconocimiento del locutor, lo beneficia porque son componentes genéricas o muy comunes entre los locutores, lo cual no es deseado si se trata de discriminar entre ellos.

El empleo de la máscara permite obtener el CV con una menor dimensión, donde se encuentren presentes los valores de los CV con los mismos índices establecidos en "1" en la máscara.

3.1.2. Proyección de Atributos no Deseados y la Información Común (C&NAP)

La representación binaria propuesta en el capítulo anterior, no está aislada de la degradación del rendimiento por la variabilidad de la sesión. Por lo tanto, compensar estas variabilidades se convierte en una etapa obligatoria del reconocimiento del locutor. Existen varios algoritmos enfocados en mitigar este problema, como son: Proyección de Atributos No Deseados (NAP) propuesto en [81], la Normalización de la Covarianza dentro de la Clase (WCCN) propuesto en [34] y el Análisis de Discriminante Lineal (LDA) propuesto en [26], los cuales serán modificados para mejorar sus desempeños sobre las representaciones propuestas.

Para incorporar la información común en los métodos de compensación de variabilidad se propuso inicialmente una modificación al método NAP, la cual se denominó Proyección de Atributos no Deseados y Comunes (C&NAP, del inglés Common and Nuisance Attribute Projection) [39].

El método NAP consiste en obtener una matriz de covarianza que contenga la máxima cantidad de información sobre la dispersión intra-clase y la menor cantidad posible de información de la dispersión entre-clases, sobre una base de datos de una

población etiquetada por locutores. Por lo tanto, el rendimiento del método NAP depende en gran medida de la información contenida en la matriz de covarianza S_w (ec. 1.26), que contiene la dispersión intra-clase de los atributos no deseados.

A continuación, los principales supuestos del NAP:

1. La información discriminatoria deseada en una clase (locutor) está contenida en su vector de medias.

2. Al eliminar el vector medio de las diferentes observaciones (sesiones) de la misma clase, se obtienen todos los atributos no deseados de esta clase.

3. Seleccionado los vectores propios asociados a los mayores valores propios (obtenidos mediante la solución de la ec. 1.25), se debe dominar la mayor cantidad de información sobre la variabilidad de la sesión (atributos no deseados).

Existen algunos puntos débiles en estos supuestos, como se demuestra por Baker y otros en [5]. Por ejemplo, en los primeros vectores propios asociados a los mayores valores propios, se cuenta con información sobre la variabilidad de sesión, pero también existe información sobre la variabilidad de los locutores. Para resolver este problema, en [5] proponen incorporar la información de la dispersión entre-clases en la matriz de proyección, siguiendo la lógica del análisis discriminatorio lineal (LDA).

La propuesta para fortalecer el desempeño del NAP se centra en otro punto de vista: el método NAP no tiene en cuenta la información común compartida entre los locutores, lo que afecta en gran medida los sistemas de reconocimiento. En el supuesto 1 del método NAP, el objeto medio de la clase no sólo contiene información característica de la misma, también contiene información común a las otras clases. Estos atributos comunes pueden estar presentes en la señal de voz por el genero, por el lenguaje utilizado, por el contenido fonético de expresiones similares, y por la redundancia o superposición de las mono-Gaussianas.

La idea del enfoque que se propone C&NAP, consiste en eliminar no sólo los atributos no deseados, sino también los atributos comunes entre los locutores, al incorporar en la matriz de proyección la información común a todas las clases. De este modo se obtiene una matriz S_w, que dominará no sólo la información de la dispersión intra-clase, sino que también la información común entre-clases, la cual es un atributo no deseado. Una forma intuitiva de entender el proceso, consiste en restar la información común del objeto medio de cada clase, antes de sustraer el objeto medio de las observaciones correspondientes a su clase.

Una forma sencilla consiste en emplear $\bar{x}$ como la información común, donde $\bar{x}$ es el objeto medio global de la base de datos de la población de locutores, modificando la ec. 1.26.

$$S_w = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(x_i^s - (x_s - \overline{x}))\ (x_i^s - (x_s - \overline{x}))', \tag{3.1}$$

donde S es el número de locutores (clases), n_s el número de expresiones de voz correspondientes a cada locutor s, x_s es el vector medio de las expresiones de voz de cada locutor ($x_s = \frac{1}{n_s}\sum_{i=1}^{n_s} x_i^s$), y x_i^s es la i-ésima expresión de voz del locutor s. El resto del algoritmo es ejecutado como el NAP.

El método C&NAP tiene como objetivo reducir los efectos de la variabilidad de sesión para cada locutor, manteniendo la separación entre-clases al tratar de remover los atributos comunes y no deseados de la representación de los locutores.

La implicación matemática de la inclusión del objeto medio global $\bar{x}$ en la matriz de covarianza S_w, consiste en un desplazamiento del origen de coordenadas hacia el objeto medio global, donde será calculada la matriz de covarianza.

Implicación matemática:

Si a la ec. 1.26: $S_w^{NAP} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(x_i^s - x_s)(x_i^s - x_s)'$, se le incluye el objeto medio global $\bar{x}$, obtenemos entonces

la ec. 3.1: $S_w^{C\&NAP} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(x_i^s - (x_s - \overline{x}))\ (x_i^s - (x_s - \overline{x}))'$, esta incorporación implica que:

$$S_w^{C\&NAP} = S_w^{NAP} + \overline{x}\overline{x}'$$

Demostración:

Las dos matrices de dispersión difieren solamente por el término $\overline{x}\overline{x}'$, que es una matriz de rango 1. La aplicación de esta matriz aporta datos de un subespacio $\overline{x}$ de dimensión 1 (una línea).

Partimos de:

$S_w^{C\&NAP} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}((x_i^s - x_s) + \overline{x})\ ((x_i^s - x_s) + \overline{x}))'$

$S_w^{C\&NAP} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}\{(x_i^s - x_s)(x_i^s - x_s)' + 2(x_i^s - x_s)\overline{x}' + \overline{x}\overline{x}'\}$

y trabajando sobre el segundo término tenemos que:

$$\frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(x_i^s - x_s)\overline{x}' \;=\; \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\{\sum_{i=1}^{n_s}(x_i^s - x_s)\}\overline{x}' \;=\; \frac{1}{S}\sum_{s=1}^{S}\frac{2}{n_s}A\overline{x}'$$

donde $A = \sum_{i=1}^{n_s}(x_i^s - x_s) \;=\; \sum_{i=1}^{n_s} x_i^s - \sum_{i=1}^{n_s} x_s \;=\; n_s x_s - n_s x_s \;=\; 0.$

De modo que:

$$S_w^{C\&NAP} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}\{(x_i^s - x_s)(x_i^s - x_s)'\} + \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}\overline{x}\overline{x}'$$

$$S_w^{C\&NAP} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}\{(x_i^s - x_s)(x_i^s - x_s)'\} + \overline{x}\overline{x}'$$

$$S_w^{C\&NAP} = S_w^{NAP} + \overline{x}\overline{x}'.$$

Finalmente, dado un vector propio v de $S_w^{C\&NAP}$, tenemos que $S_w^{C\&NAP}v \;=\; \lambda v \;=\; S_w^{NAP}v + \overline{x}\overline{x}'v \;=\; S_w^{NAP}v + \alpha\overline{x}$, donde α es un escalar igual a $\overline{x}'v$, lo cual completa la demostración.

Análisis de las distribuciones de la varianza espectral

En esta sección se presenta una herramienta visual con la intención de analizar gráficamente el comportamiento de las distribuciones de varianza espectral de los datos [39], esta herramienta fue utilizada por Bousquet y otros en [14]. Dado un conjunto de CV expresados en una base dada, se denomina "gráfico espectral" a un gráfico que relaciona las dimensiones del espacio con la variabilidad de los locutores y la sesión, reflejadas en el espectro de los valores propios correspondientes a sus matrices de covarianza.

Inicialmente se calcula la matriz de variabilidad total S_t, obtenida a partir del cálculo de la covarianza de todos los CV: $S_t = \frac{1}{S}\sum_{s=1}^{S}(x_i - \overline{x})(x_i - \overline{x})')$, donde S es el número de locutores de la población y $\overline{x}$ la media global. A continuación se obtienen las matrices S_w que contienen la dispersión intra-clase (ec. 1.26) y la dispersión intra-clase + la información común (ec. 3.1). Luego se obtiene la matriz $S_b = S_t - S_w$ que contiene la información de la dispersión entre-clases.

El gráfico espectral (fig. 3.2) muestra las curvas definidas por los valores propios de la matrices S_w y S_b, los cuales contienen la proporción de la varianza de los locutores

y de las sesiones por dimensiones del espacio.

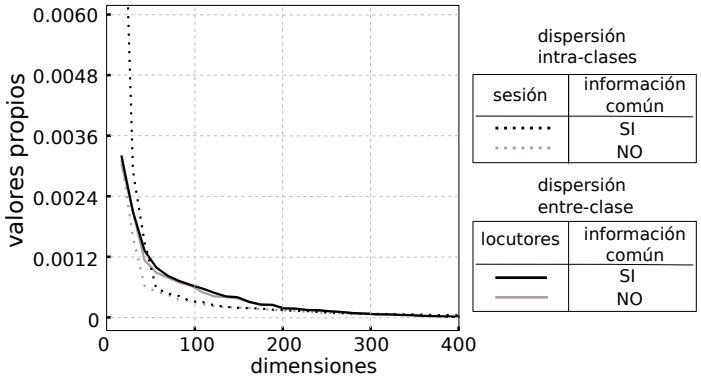


Figura 3.2: Gráfico espectral de los valores propios con/sin información común. El eje y muestra los valores de varianza de los locutores y la sesión hasta 400 dimensiones que muestra el eje x.

A continuación dos importantes conclusiones de la visualización.

- Un resultado significativo es que sobre las primeras 100 dimensiones la información de la varianza de sesión, utilizando los atributos comunes, está por encima de la varianza sin usar la información común; incluso en la primera dimensión, la energía es 15 veces mayor. Esto demuestra que la matriz de covarianza utilizando la información común contiene mucha más información de la dispersión intra-clase.

- La energía del espectro de la varianza entre clases S_b se mantiene; menos de 250 dimensiones contienen una proporción importante de la información de variabilidad entre locutores. Además, en el espectro de la varianza intra-clase S_w, la mayor proporción de información de la variabilidad de sesiones se observa en las mismas 250 dimensiones.

A modo de conclusión, la incorporación del objeto medio global fortalece el método NAP, como se demostrará en los experimentos, su inclusión es muy sencilla y no causa costos computacionales adicionales.

3.1.3. Información Común para la Normalización de la Covarianza Intra-clase (C&WCCN)

El enfoque utilizado en la matriz de covarianza del C&NAP puede ser fácilmente extrapolable a otras técnicas, ya que la información de la dispersión intra-clase se enriquece mediante la incorporación de los atributos comunes [39].

En el caso del WCCN, la modificación se denominó Atributos Comunes en la Normalización de la Covarianza Intra-clase (C&WCCN, del inglés Common Attribute in Within-class Covariance Normalisation).

Dado un conjunto de muestras representativas de señales etiquetadas por locutor en una población, la idea consiste en normalizar la dispersión existente dentro de las clases que definen los locutores. Como se describe en la Sección 1.4.2, el núcleo del método es la matriz de covarianza W, obtenida por la ec. 1.26. La propuesta, al igual que en el método C&NAP consiste en incluir la información común de la población en el proceso, por lo que para el método C&WCCN equivale a obtener W utilizando la ec. 3.1.

Esta pequeña modificación fortalece el método en busca de mejorar su desempeño sobre el marco binario propuesto, como se verá en los experimentos. Además mantiene los conceptos básicos del WCCN, como son la normalización de la dispersión intra-clase garantizando la conservación de la direcciones de las varianzas en el espacio.

3.1.4. Análisis de Discriminante Lineal Desplazado (S-LDA)

En el caso del método LDA, se denominó la variante propuesta como Análisis de Discriminante Lineal Desplazado (S-LDA, del inglés Shifted Linear Discriminant Analysis).

Como se describe en la sección 1.4.3, el método LDA se basa en dos matrices capaces de contener la mayor cantidad de información sobre la dispersión intra-clase S_w y entre-clases S_b. El método LDA presenta casi los mismos supuestos que el NAP para obtener la matriz S_w. Donde no se tiene en cuenta la información común del espacio de los locutores, asumiendo que el vector medio de un locutor x_s sólo contiene la información deseada (discriminatoria) de la representación de un locutor, lo cual como se ha demostrado no es correcto, porque también contiene información común de la población entera.

Para mitigar este problema, se propone solamente aumentar la información a tener en cuenta de la dispersión intra-clase, para lograrlo se introduce, como en las propuestas anteriores, la información común de la población, desplazando el centro del área de trabajo hacia el centro real de la población, utilizando la ec. 3.1 para obtener la matriz S_w.

Como se comprobará en los siguientes experimentos, la incorporación del objeto global medio de la población en los diversos métodos fortalece su desempeño en la compensación de la variabilidad de sesión.

3.1.5. Evaluación experimental

Los experimentos presentados en esta sección fueron desarrollados utilizando como rasgos: 19 coeficientes LFCC, junto con la energía, los deltas y los deltas-deltas, formando un vector de rasgos de dimensión 60.

En busca de una mejor relación con las evaluaciones NIST SRE se implementó un nuevo Modelo Generador (GM), empleando un UBM inicial de 512 componentes Gaussianas entrenado con la base de datos NIST SRE 2005. Dicho UBM fue reducido a 459 componentes. Las especificidades fueron obtenidas de una selección de los vectores medios en los modelos adaptados GMM-UBM de las expresiones de voz de la misma base. Esto hace una diferencia en las especificidades respecto al GM utilizado en las evaluaciones del capítulo anterior, obteniéndose ahora $E = 36872$ especificidades, siendo E la dimensión de los CV y BV.

Para la transformación al espacio binario de los rasgos acústicos, se utilizaron las 3 componentes del UBM con mayor probabilidad por trama y aquellas especificidades correspondientes a una probabilidad a posterior mayor que 0.001 se etiquetaron en "1".

Con el fin de fijar el umbral de la varianza en la máscara en $\theta = 0.11$, se utilizaron los CV obtenidos a partir de 2450 señales multilingües de 124 locutores tomadas de la base NIST SRE 2004. Los experimentos de reconocimiento del locutor utilizados para fijar el umbral (fig. 3.1) fueron realizados con la base NIST SRE 2008, en su condición det 7. Como resultado las especificidades con mayor varianza que θ fueron etiquetadas en "1", obteniéndose una selección de $k = 24000$. El uso de la máscara implicó que todos los CV y BV tendrán dimensión k.

Los experimentos de verificación del locutor independiente del texto se realizaron sobre la base NIST SRE 2008, con locutores masculinos, en la condición det 7 (conversaciones por teléfono en inglés). Esta condición utiliza 439 locutores clientes y 671 pruebas, realizando 6.615 verificaciones.

Para el análisis de las distribuciones de la varianza espectral y para entrenar las matrices de proyección de los métodos de compensación de la variabilidad de sesión, se utilizaron las bases de datos NIST SRE 2004 y 2005, con 3285 sesiones telefónicas en idioma inglés de 262 locutores.

Todos los resultados de las evaluaciones se expresan en términos de EER y MinDCF.

Primero se evaluó el impacto del enfoque de la selección basada en la varianza de las especificidades. La Tabla 3.1 presenta los resultados del experimento en el marco

de la representación binaria mediante la compensación de la variabilidad de sesión con el método NAP, sin y con la máscara, respectivamente. La medida de similitud ISDS, presentada en [36] se utiliza como referencia en la medición sin compensación y la medida de similitud $S(x, y)$ definida en la ec. 1.27 en la medición con la compensación NAP.

Tabla 3.1: Técnicas de compensación sobre los vectores acumulativos, sin la máscara (36872-dimensiones) y con la máscara (24000-dimensiones).

Máscara	compensación	similitud	DCF*100	EER %
no	no	$ISDS(x, y)$	5.5	12.5
si	no	$ISDS(x, y)$	4.8	11.6
no	NAP	$S(x, y)$	6.3	18
si	NAP	$S(x, y)$	4.3	10.24

Todos los resultados del experimento muestran una mejora en la eficacia cuando se utiliza la máscara, lo que implica una ganancia al lograrlo con una reducción del 35 % de las especificidades. Si se observan las dos primeras filas de la Tabla 3.1, se puede observar que la máscara mejora la eficacia del algoritmo, debido a las características genéricas de las especificidades que se han eliminado, pues no contienen información discriminatoria pero si contienen información perjudicial para el rendimiento del clasificador. El mismo problema ocurre en la captura de la dispersión de los atributos no deseados con el método NAP, en las últimas dos filas de la Tabla, tal dispersión intra-clase está estrechamente relacionada con la información discriminatoria contenida en las especificidades (cuanto mayor sea el conocimiento de la dispersión intra-clase, mejor será el rendimiento con respecto a la captura de los atributos). Luego, la aplicación de la máscara sobre los CV conlleva a una mejor representación de la dispersión de los atributos no deseados en el NAP. El resto de los experimentos en el trabajo se realizaron después de aplicar esta máscara.

Una primera aplicación de la propuesta de compensación de la variabilidad de sesión C&NAP, se llevó a cabo aplicando la ec. 3.1 para obtener la matriz de compensación y utilizando SVM como clasificador [62]; se compararon los resultados con los métodos del estado del arte [26]. En este trabajo se utilizó, la base NIST 2004 para obtener las matrices de proyección y la matriz de variabilidad total T (con 400 dimensiones) así como para el entrenamiento de la SVM.

La Tabla 3.2 muestra los resultados del experimento.

Este experimento obtuvo resultados comparables (5.97 % de EER) entre el marco binario propuesto y la mejor variante del sistema de referencia (5.82 % de EER) del

Tabla 3.2: Evaluación del enfoque C&NAP en el marco binario y su comparación con i-vector.

	DCF*100	EER %
CV + C&NAP + SVM	3.37	**5.97**
CV + ISDS	4.86	11.69
i-vector	3.09	7.09
i-vector + LDA	2.90	**5.82**
i-vector + WCCN	2.86	6.65
i-vector + WCCN + LDA	2.83	5.92

estado del arte aplicando el método i-vector. Además se observa la importancia de la inclusión de la información común en el marco binario, provocando un 48.9 % de mejora de la eficacia medida por el EER.

Una evaluación más completa de la compensación de la variabilidad de sesión en la representación binaria fue llevado a cabo en otros experimentos, donde se utilizaron para las matrices de proyección y para la matriz de variabilidad total T (con 400 dimensiones) las bases NIST 2004 y 2005, lo que significó un aumento de la variabilidad de sesión.

Los principales resultados del experimento se presentan en la Tabla 3.3, donde se comparan los métodos de compensación de la variabilidad de sesión NAP, WCCN con la propuesta de incluir los atributos comunes en los métodos C&NAP y C&WCCN [39]. Se obtiene la puntuación con dos métodos de similitud: el producto y el coseno.

Para la aplicación de los métodos WCCN y C&WCCN sobre los CV fue necesario utilizar una técnica de reducción de la dimensión (PCA), obteniéndose una significativa reducción de 24000 dimensiones a $W = 400$. Esto implica que la complejidad computacional del proceso de extracción del CV sea $\mathbf{O(EL + EW)}$, donde el término EL está dado por la obtención de CV y el término EW por la proyección del CV al nuevo espacio de dimensión W.

Tabla 3.3: Métodos de compensación de la variabilidad de sesión C&NAP y C&WCCN sobre los vectores acumulativos CV.

compen-	similitud	det 7		similitud	det 7	
sación	por producto	DCF*100	EER %	por coseno	DCF*100	EER %
NAP	$S_{prod-NAP}(x,y)$	7.2	20.2	$S_{cos-NAP}(x,y)$	4.3	10.24
C&NAP	$S_{prod-NAP}(x,y)$	**3.9**	**6.37**	$S_{cos-NAP}(x,y)$	**2.7**	**5.45**
WCCN	$S_{prod-WCCN}(x,y)$	5.2	11.61	$S_{cos-WCCN}(x,y)$	4.1	10.47
C&WCCN	$S_{prod-WCCN}(x,y)$	**4.5**	**7.28**	$S_{cos-WCCN}(x,y)$	***2.7***	***5.37***

Para calcular la similitud por producto en el NAP se utilizó la ec. 1.27 y en el

WCCN la ec. 1.30; para el cálculo de la similitud por coseno en el NAP se utilizó la ec. 1.28 y en el WCCN la ec. 1.31.

Una mejora significativa en términos de eficacia se obtiene utilizando el enfoque propuesto de inclusión en los atributos no deseados de la información común, para ambos métodos de puntuación, alcanzándose, con C&NAP un promedio de mejora del 57 % en el EER y alcanzándose con C&WCCN un promedio de mejora del 43 % en el EER. Además, hay que señalar que el uso de la similitud por coseno trae mejores resultados que el empleo de la similitud por producto.

Con el objetivo de comprobar la generalización de los métodos C&NAP y el C&WCCN, se llevó ha cabo un nuevo experimento en el marco de los i-vectores, cuyos resultados se muestran en la Tabla 3.4.

El experimento con los i-vectores se realizó sin aplicar la normalización de la longitud y usando como método de similitud el coseno.

Tabla 3.4: Generalización de los métodos de compensación de la variabilidad de sesión sobre los i-vectores.

compensación	DCF(*100)	EER %
NAP	2.44	4.8
C&NAP	2.42	4.46
WCCN	2.31	4.01
C&WCCN	2.01	3.87
LDA-NAP	2.25	4.04
LDA-*C&NAP*	**2.16**	**3.86**
LDA-WCCN	2.25	3.93
LDA-*C&WCCN*	**2.01**	***3.72***

La Tabla 3.4 muestra una comparación de los métodos de compensación de la variabilidad de sesión, comúnmente aplicados para el enfoque i-vector y los métodos propuestos en este trabajo: C&NAP y C&WCCN. En la primera parte de la tabla se comparan los métodos propuestos con las técnicas que funcionan sólo con la matriz de dispersión intra-clase (NAP y WCCN). En la segunda parte se comparan con la técnica que utiliza la matriz de dispersión intra-clase y la matriz de dispersión entre-clase (LDA) en combinación con los métodos NAP y WCCN.

Como refleja la tabla los métodos propuestos muestran una mejoría en la eficacia comparados con el NAP, el WCCN y las combinaciones de LDA-NAP y de LDA-WCCN; lo que implica que la incorporación de la información común en el enfoque i-vector también incrementa su rendimiento, permitiendo confirmar la generalización de los métodos propuestos.

El próximo experimento, en la Tabla 3.5, refleja los resultados de la fusión lineal de las puntuaciones del enfoque i-vector utilizando LDA, (fig. 1.6 utilizando la similitud por coseno), más los métodos propuestos de compensación C&NAP y C&WCCN y el enfoque binario, con los vectores acumulativos, empleando los mismos métodos propuestos.

Tabla 3.5: Fusión lineal de las puntuaciones entre el enfoque i-vector y la representación con los CV.

fusión	DCF(*100)	EER %
(i-vector : LDA-$C\&NAP$) + (CV : C&NAP)	**2.03**	**3.33**
(i-vector : LDA-$C\&WCCN$) + (CV : C&WCCN)	**2.06**	**3.52**

En la primera fila de la Tabla 3.5 se evaluó la utilización del método C&NAP, obteniéndose una mejora de 13.7 % de EER sobre el segundo mejor resultado mostrado en la Tabla 3.4 (3,86 % de EER). En la segunda fila de la Tabla se evaluó la utilización del método C&WCCN, obteniéndose una mejora de 5.37 % de EER sobre el mejor resultado que se muestra en la Tabla 3.4 (3,72 % EER).

3.1.6. Conclusiones parciales

Esta sección tuvo como objetivo asociar las posibilidades de la nueva representación binaria de las expresiones de voz, con los nuevos métodos de compensación de la variabilidad de sesión, publicados en [39] y [62].

Se propuso una nueva variante de la matriz de dispersión intra-clase en el marco de la representación binaria y en el enfoque i-vector, que es capaz de contener no sólo los atributos no deseados, sino también la información común, que tampoco es deseada. Empleando la medida de similitud coseno, este enfoque logró una disminución relativa del 47.7 % en el EER promedio; el mejor resultado obtenido utilizando el método propuesto (C&WCCN) fue de 5.37 % en el EER. Este resultado muestra la importancia de incluir en la matriz de dispersión intra-clase, de los CV, la información común, lo que confirma las conclusiones obtenidas del análisis de la varianza espectral.

También se presenta un método simple (la máscara) para seleccionar las especificidades más discriminatorias de los vectores acumulativos, capaz de eliminar hasta un tercio (12.872 de 36.872) de las especificidades del modelo generador e incrementando

la eficiencia y la eficacia en el reconocimiento del locutor.

En el marco de los i-vectores se logró también una mejora en el rendimiento (5 % en el EER promedio) utilizando los métodos propuestos de compensación de la variabilidad de sesión, debe observarse que dicha mejora no es comparable a la obtenida en el dominio de la representación binaria. Esto sucede porque en el dominio I-vector hay menos información común, a diferencia de los vectores acumulativos, donde su contenido se relaciona con las repeticiones de las especificidades. Observe, por último, que ambos dominios contienen información complementaria, responsable de la mejoría en la fusión de la puntuación, cumpliéndose así la hipótesis principal de la investigación.

Un inconveniente de estas técnicas de compensación, es que su rendimiento se ve afectado si el número de muestras para el entrenamiento de las matrices de proyección no es lo suficientemente grande. Los resultados pueden ser mejorados si se incrementan las bases de datos de entrenamiento.

3.2. Información Temporal sobre el enfoque binario

Esta sección trata sobre la información temporal obtenida de la matriz binaria de las expresiones de voz. No solo la información global (CV) de la voz puede ser extraída de la matriz binaria sino también la información temporal existente en ella, la cual refleja los continuos cambios del tracto vocal. Esta nueva información está dividida en dos tipos, la que representa la dinámica de los sonidos y la que se obtiene a nivel de tramas, ambas aportan poder discriminatorio a los rasgos del locutor.

3.2.1. Modelo de Trayectoria

A partir del comportamiento de la voz humana, que contiene una secuencia de sonidos que pueden representarse por fonemas, sílabas, palabras, etc, se propuso obtener información de la dinámica de la voz creando un modelo de trayectoria, aplicando ventanas (bloques) solapadas y consecutivas sobre la matriz binaria. Cada ventana tiene un tamaño k y entre dos ventanas consecutivas existe un desplazamiento z, obteniendo en una expresión de voz V ventanas, donde $V = (L - k)/z$ y L es la cantidad de tramas acústicas de la expresión. Cada bloque, definido por una ventana, es una sub-matriz binaria, facilitando la obtención de un CV y un BV para cada uno de ellos.

Se define el Modelo de Trayectoria (TM, del inglés Trajectory Model) de una expresión de voz, como un conjunto de vectores acumulativos y binarios, $TM = \{c_1, c_2, \ldots, c_V\}$, donde cada elemento contiene un CV y BV correspondiente al segmento temporal [11, 36] (ver fig. 3.3). Notar que V depende de la duración de la expresión de voz.

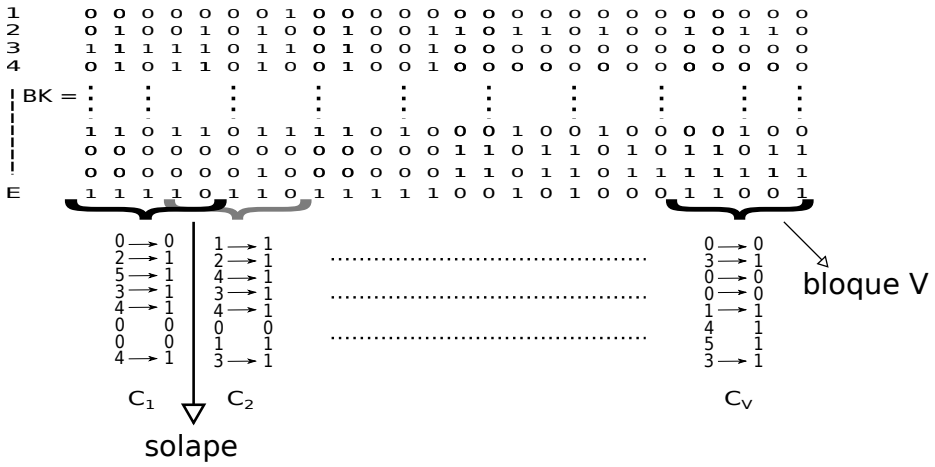


Figura 3.3: Proceso para la obtención del Modelo de Trayectoria.

Esta propuesta presenta como ventaja que, un segmento de habla contaminado por ruido debe ser considerado como una pérdida respecto al carácter discriminatorio para el reconocimiento del locutor, por lo tanto, una segmentación y superposición adecuada de los bloques en la expresión de voz podría reducir el efecto del ruido y aumentar la robustez de la representación.

A continuación se presentan dos algoritmos para utilizar el TM dado los rasgos acústicos de una expresión de voz del locutor cliente A y los de una expresión de voz desconocida B.

El algoritmo presentado es una extensión del Algoritmo 1, propuesto en la sesión 2.3.2 y realiza una comparación utilizando la información global (CV) de A con el modelo de trayectoria de B y viceversa. En busca de facilitar la notación se nombró GT a la unión de un CV y un TM en el mismo contexto, $GT = \{CV, TM\}$.

Algoritmo 2

1. Obtener la matriz binaria para ambas matrices acústicas, $\{BK^A, BK^B\}$.

2. Obtener el vector acumulativo para ambas matrices binarias, $\{BK^A, BK^B\} \Rightarrow \{CV^A, CV^B\}$.

3. Obtener el modelo de trayectoria para ambas matrices binarias, $\{BK^A, BK^B\} \Rightarrow \{TM^A, TM^B\}$.

4. Crear las duplas, $GT^A = \{CV^A, TM^A\}$, $GT^B = \{CV^B, TM^B\}$.

5. Obtener la puntuación de la verificación entre el locutor cliente y el desconocido:

$$S(GT^A, GT^B) = \frac{1}{2}(\frac{1}{V^B}\sum_{v=1}^{V^B} ISDS(TM^B[v], CV^A) + \frac{1}{V^A}\sum_{v=1}^{V^A} ISDS(TM^A[v], CV^B)) \tag{3.2}$$

6. Si $S \geqslant \theta$ se acepta, sino se rechaza.

El siguiente algoritmo combina, la información global (Algoritmo 1) con el modelo de trayectoria Algoritmo 2.

Algoritmo 3

Los pasos 1, 2, 3, 4 y 6 son los mismos que en el Algoritmo 2.

5 Obtener la puntuación de la verificación entre el locutor cliente y el desconocido:

$$S(GT^A, GT^B) = \frac{1}{3}(\frac{1}{V^B}\sum_{v=1}^{V^B} ISDS(TM^B[v], CV^A) \tag{3.3}$$

$$+\tfrac{1}{V^A}\textstyle\sum_{v=1}^{V^A} ISDS(TM^A[v], CV^B) + ISDS(CV^B, CV^A)).$$

3.2.2. Información temporal en las tramas

Como el modelo de trayectoria no es capaz de capturar la información temporal a nivel de trama se propone un nuevo enfoque para obtener dicha información temporal.

La idea de este nuevo enfoque surge de la siguiente hipótesis: si tenemos una representación binaria (BK), ordenada en el tiempo, que contiene la información global de cada trama (las columnas de BK) de la expresión de voz, entonces es posible extraer la información temporal a nivel de la trama, que provienen de los continuos cambios en corto tiempo, que ocurren en el tracto vocal y obtener así nuevas características discriminatorias de cada locutor [38].

Esta información temporal se basa en capturar los cambios que suceden trama por trama, que equivale a obtener la información de la naturaleza dinámica de las especificidades respecto a sus activaciones y desactivaciones en el tiempo. Este proceso sigue una idea similar a la extracción de los rasgos deltas [29].

Se trata de contar los cambios de una trama a otra en la matriz binaria, siempre secuencialmente a lo largo de la expresión y pesarlos por la importancia de la trama dentro de la expresión de voz. Dado una matriz binaria BK de una expresión de voz, con dimensión $E \times L$, donde E es el número de especificidades y L la cantidad de tramas de la expresión, se define como matriz de cambio a:

$$D_{i,j}^{cambio} = |bk_{i,j} - bk_{i,j+1}|, \tag{3.4}$$

donde $i = 1, 2, ..., E$, con $j = 1, 2, ..., L-1$ y D^{cambio} es una matriz binaria que contiene los cambios entre tramas de la especificidad correspondiente.

Como la matriz de cambio D presenta altas dimensiones se propone obtener un solo Vector Temporal (TV, del inglés Temporal Vector) capaz de compactar la información temporal de las tramas pesadas por su importancia relativa,

$$TV[i] = \sum_{j=1}^{L} \Big(D_{i,j} * \sum_{i=1}^{E} D_{i,j} \Big). \tag{3.5}$$

Obteniéndose un TV de dimensión E, capaz de contener la información temporal de la expresión de voz del locutor.

La complejidad computacional del proceso de extracción de la información temporal por tramas es $\mathbf{O(EL)}$ y realiza $(EL + EL + EL)$ operaciones. La cantidad de operaciones está dada por el cálculo de la matriz de cambio (EL), el cálculo de los pesos de los cambios en cada trama (EL) y la suma de los cambios por especificidades (EL). Note que cada uno de estos cálculos se puede realizar de forma secuencial.

Finalmente, se normaliza el TV por su norma euclidiana, en busca que todos los vectores tengan la misma longitud.

$$TV = \frac{TV}{||TV||}. \tag{3.6}$$

3.2.3. Evaluación experimental

Para evaluar el impacto de la información temporal sobre el desempeño de la verificación del locutor independiente del texto, se realizaron dos experimentos. El

primero consistió en incluir y evaluar la información temporal a *nivel de segmentos* (Modelo de Trayectoria) y el segundo, en incluir y evaluar la información temporal a *nivel de tramas* (Vector Dinámico).

Empleo de la información temporal a nivel de segmentos en el reconocimiento del locutor

En el primer experimento se utilizó el GM ya creado y empleado en la Sección 2.3.3, el cual consta de 459 componentes para el UBM y 22000 mono-Gaussianas. Además se utilizaron, como rasgos acústicos, 12 MFCC + 12 deltas con normalización cepstral de la media y la varianza.

La verificación del locutor independiente del texto se realizó con las sesiones telefónicas T1 y T2 de la base de datos Ahumada. Cada una de las sesiones cuenta con 100 locutores, con alrededor de 1.5 minutos de señal de voz espontánea; las expresiones de T1 se utilizan como conjunto de locutores conocidos (100 clientes) y las expresiones de T2 como conjunto de prueba, para un total de 10000 comparaciones.

Un primer experimento (Tabla 3.6), utilizando el *Algoritmo 2* y la medida ISDS, se realizó en busca de seleccionar el tamaño de los bloques y el desplazamiento que reportara la mejor eficacia en términos de EER.

Tabla 3.6: Evaluación del Algoritmo 2 con diferentes tamaños y desplazamientos de los bloques.

Tamaño por bloque	Desplazamientos						
	300	250	200	150	100	50	10
500	7.03	7.02	7.06	7.19	7.55	7.43	7.47
400	7.12	7.11	6.90	7.2	7.23	7.3	7.4
300	–	7.15	7.07	**6.8**	6.98	7.35	7.6
200	–	–	–	7.9	7.1	7.41	8
100	–	–	–	–	–	8.2	8.5

Como resultado se obtuvo que el modelo de trayectoria TM que mejor eficacia reporto utilizó k = 300 tramas por cada bloque con un desplazamiento de z = 150 tramas. Estos parámetros representan segmentos de una expresión de voz de 3 segundos con un desplazamiento de 1.5 segundos. Estos son los parámetros utilizados en los siguientes experimentos donde se utilizó el modelo de trayectoria.

Un segundo experimento se centró en probar la robustez del modelo de trayectoria utilizando como medida de puntuación la ISDS, además se utilizó como sistema de referencia el enfoque GMM-MAP [75] sobre la plataforma ALIZE [12] (Ver ANEXO

B). La Tabla 3.7 presenta los resultados del experimento en términos de EER y DCF, para los tres algoritmos de comparación propuestos en esta tesis.

Tabla 3.7: Resultados del empleo de la información temporal a nivel de segmentos.

	DCF*100	EER %
IS	3.9	6
Algoritmo 1	2.97	5.48
Algoritmo 2	4.33	6.8
Algoritmo 3	**2.19**	**5**
GMM-MAP	2.51	5.42

Estos resultados mostraron, con el *Algoritmo 3*, que es posible capturar información discriminatoria temporal a nivel de segmentos de una expresión de voz, obteniendo una mejora en la eficacia del 8 % de EER ante los resultados del Algoritmo 1, que sólo utiliza la información global. También queda plasmado la robustez de la propuesta, comparada con el sistema de referencia GMM-UBM obteniéndose una mejora del 7 % del EER, resultados publicados en [11] y [36].

Empleo de la información temporal a nivel de tramas en el reconocimiento del locutor

Para comprobar el desempeño de los métodos propuestos al incluir la información temporal a nivel de tramas se utilizó como rasgos: 19 coeficientes LFCC, junto con la energía, los deltas y los deltas-deltas, resultando vectores de rasgos con dimensión 60. Se empleó el mismo GM y su representación binaria, presentado en las evaluaciones de la Sección 3.1.5: UBM con 459 componentes Gaussianas, con $E = 36872$ especificidades de las cuales fueron seleccionadas 24000, utilizando la máscara.

El proceso de verificación de locutores independiente del texto se realizó sobre la base NIST SRE 2008, con locutores masculinos, la condición det 7 (conversaciones teléfono-teléfono en inglés). Esta condición utiliza 439 locutores clientes y 671 pruebas, realizando 6615 verificaciones. Para obtener la puntuación entre dos locutores se utilizaron dos variantes, la medida del coseno (ec. 1.36) y la verosimilitud (ec. 1.40) en el caso del PLDA.

Se utilizó la base NIST SRE 2004 y 2005 para obtener la matriz de variabilidad total T, las matrices de proyección de los métodos de compensación de la variabilidad de sesión y para el entrenamiento del clasificador PLDA.

La Tabla 3.8 muestra una comparación, en el marco de la representación binaria propuesta, del poder discriminatorio de la información global (CV), el modelo de

trayectoria (TM) utilizando el mejor algoritmo de comparación, el Algoritmo 3 y la información temporal por tramas (TV). También se evalúa el desempeño de la información común incluida en la matriz que contiene la dispersión intra-clase de los locutores, propuesta en el método S-LDA en la sección 3.1.4.

Para la aplicación del método S-LDA sobre los CV, TM, TV fue necesario utilizar una técnica de reducción de la dimensión (PCA), obteniéndose una significativa reducción de 24000 dimensiones a $W = 400$. Esto implica que la complejidad computacional del proceso de extracción de CV o TV sea $\mathbf{O(EL + EW)}$, donde el término EL está dado por la obtención de CV o TV y el término EW por la proyección al nuevo espacio de dimensión W. Obsérvese que el proceso de obtención de TM presenta la misma complejidad computacional que el proceso para obtener CV.

Tabla 3.8: Evaluación de las representaciones CV, TM, TV

representación	compensación y medida	DCF *100	EER %	compensación y medida	DCF *100	EER %
CV	LDA+cos	5.64	10.50	S-LDA+cos	2.86	5.46
CV	LDA+PLDA	2.76	4.92	S-LDA+PLDA	2.70	4.54
TM	LDA+cos	5.55	10.52	S-LDA+cos	3.01	5.70
TM	LDA+PLDA	2.81	4.90	S-LDA+PLDA	2.78	4.80
TV	LDA+cos	5.36	10.45	S-LDA+cos	2.80	5.27
TV	LDA+PLDA	**2.78**	**4.77**	S-LDA+PLDA	**2.6**	**4.09**

Se divide el análisis de los resultados de la Tabla 3.8 en dos aspectos para una mejor comprensión. Primero respecto a la compensación de la variabilidad de sesión, donde la inclusión de la información común muestra, una vez más, su eficacia, el EER promedio para todos los métodos de representación sin el empleo de la información común es 7.67 % y al incorporar la información común el EER promedio se reduce a 4.97 %, para una ganancia del 35 %.

Como segundo aspecto se comparan los resultados utilizando la información temporal, ante el uso de la información global, el desempeño de forma general para ambos clasificadores y con la inclusión o no de la información común. Los resultados en las tres representaciones son muy similares, sobre todo entre el TM y el CV. Pero cuando se observa el desempeño, se aprecia que la representación de la información temporal a *nivel de tramas* TV, consigue los mejores resultados, con una reducción relativa (6 % del DCF y 14 % de EER) con respecto al TM y también con respecto al CV (3 % del DCF y 9.9 % del EER). Este resultado mostró la importancia de obtener e incorporar la información temporal a nivel de tramas en la representación TV, la cual

demostró ser más discriminatoria para la verificación del locutor, que la información global y el modelo de trayectoria.

Por último, la Tabla 3.9 muestra los resultados de los métodos del estado del arte en la verificación del locutor independiente del texto (fig. 1.6), así como la fusión lineal de las puntuaciones con los resultados propuestos.

Tabla 3.9: Evaluación del enfoques i-vector y la fusión de la puntuación con la representación binaria.

representación	compensación y medida	DCF *100	EER %
i-vector	LDA+cos	2.24	3.80
i-vector	LDA+PLDA	1.89	2.96
Fusión de las puntuaciones entre i-vector y CV o TV.			
(i-vector:LDA+PLDA) + (CV:S-LDA+PLDA)		1.88	2.89
(i-vector:LDA+PLDA) + (TV:S-LDA+PLDA)		***1.83***	***2.57***

En el caso del enfoque i-vector se presentan los resultados actuales utilizando los dos clasificadores, se debe tener en cuenta que aunque el PLDA tenga mejor eficacia, si se desea desarrollar un sistema de verificación del locutor independiente del texto que sea eficiente en el procesamiento de grandes bases de datos, es recomendable utilizar la medida de similitud del coseno.

Es de señalar que en la fusión de la puntuación con ambas representaciones, se decidió no utilizar la representación TM debido a que la combinación de los costos computacionales de la verificación del locutor utilizando el Algoritmo 3 y el método PLDA, se incrementaría significativamente.

Como se muestra en la fusión de las puntuaciones, la combinación de los dos enfoques mejora la eficacia de la verificación, y en el caso del uso de la información temporal por tramas TV, se obtiene una ganancia del 13 % de EER. Este importante resultado (***2.57 EER***) al combinar los dos enfoques, evidencia la información complementaria que existe entre ellos y la incorporación de nueva información discriminatoria que aporta la representación temporal del enfoque propuesto, aprobado para publicación en [38].

Comparación de la complejidad computacional entre el enfoque i-vector y el enfoque binario propuesto

Para la evaluación de la eficiencia de los dos enfoques (la representación por i-vector y la representación binaria propuesta) se centro la comparación en la etapa

de la extracción del vector representativo en cada una, donde radica su principal diferencia. En el caso de la etapa de clasificación ambos enfoques utilizan los mismos algoritmos, por lo tanto presentan la misma eficiencia.

La extracción del vector representativo en el enfoque i-vector consta de dos pasos, primero la extracción de las estadísticas de Baum-Welch a partir de los rasgos acústicos de una expresión de voz dado el UBM y segundo, la obtención del factor (i-vector) utilizando el UBM, las estadísticas y la matriz de variabilidad total.

En el caso del enfoque binario propuesto la extracción del vector representativo también consta de dos partes, primero la extracción de la estadística de Baum-Welch de cero orden, a partir de los rasgos acústicos de una expresión de voz dado el UBM perteneciente al Modelo Generador. El segundo paso cuenta con la activación o no de las especificidades del Modelo Generador utilizando también los rasgos acústicos.

Para el experimento se utilizaron todas las expresiones de voz de locutores masculinos en la base NIST SRE 2008, 1110 locutores. La información de habla como promedio por expresión es de 8264 tramas (rasgos acústicos), lo que equivale a un tiempo promedio por expresión de 82 segundos.

Tabla 3.10: Evaluación de la eficiencia de los dos enfoques.

	Número de operaciones	Complejidad computacional	Tiempo promedio (seg) (1 CPU)	Tiempo promedio (seg) (4 CPU)
i-vector	$T(KL + KL + W^3 + W^2K + WFK)$	$O(KL + W^3 + W^2K + WFK)$	$10.16 + 15.3 = 25.5$	$4 + 5.1 = 9.1$
binario	$T(KL + 3 \times CL + 3 \times EL + EW)$	$O(KL + CL + EL + EW)$	$\mathbf{7.1 + 10.1 = 17.2}$	$\mathbf{2.5 + 3.6 = 6.1}$

En la Tabla 3.10 el parámetro K es la cantidad de componentes Gaussianos del UBM correspondientes a cada enfoque, L es la cantidad de tramas de los rasgos acústicos y F su dimensión. En el enfoque binario C es la máxima cantidad de mono-Gaussianas que presenta un sub-modelo, E es la cantidad total de mono-Gaussianas del Modelo Generador. Finalmente, para ambos enfoques W es la dimensión del vector discriminatorio que se utilizará en la clasificación.

Como se demuestra en la Tabla 3.10 el enfoque binario propuesto presenta mejor eficiencia en todos los casos. Obsérvese que la duración del contenido de voz en una señal (L), en el caso del enfoque i-vector, afecta solamente la extracción de las estadísticas y en el enfoque binario afecta todo el proceso. Este problema implica que si aumenta considerablemente L la propuesta binaria puede presentar un tiempo mayor que la extracción de los i-vectores, pero es conocido en el área del reconocimiento del

locutor que con 1 o con 1.5 minutos de voz es suficiente para obtener una representación discriminatoria del locutor. Por otra parte lo que suele suceder en la realidad es que las señales de voz para el reconocimiento del locutor tienden a ser cortas. Concluyendo que el enfoque binario propuesto presenta una mejor eficiencia que los i-vectores.

3.2.4. Conclusiones parciales

Esta sección tuvo como objetivo capturar e incluir, la información temporal existente en las expresiones de voz, en los métodos de reconocimiento del locutor independiente del texto [11, 36] y [38].

Se presentó un método para extraer la información temporal a nivel de segmentos, lográndose la siguiente ventaja: un modelo de trayectoria TM con una segmentación y superposición adecuada de los bloques en la expresión de voz puede reducir el efecto de eventos ruidosos y aumentar la robustez de la representación.

Se logró obtener e incluir en la verificación del locutor una representación de la expresión de voz que refleja la información temporal a nivel de tramas TV, observando una mejora de la eficacia frente a la información global CV, de un 9.9 % en el EER. La representación de la información temporal, en general, es un nuevo paso en el desarrollo de nuevas representaciones dentro del marco binario, la cual es imposible obtener en el enfoque de los i-vectores. Obsérvese además, que el proceso de extracción de CV o TV es mucho más eficiente que el proceso de extracción del i-vector, $O(KL+CL+EL+EW)$ respecto a $O(KL+W^3+W^2K+WFK)$.

Se evaluó además una nueva variante de la matriz de dispersión intra-clase para el marco binario, mediante el S-LDA, capaz de tener en cuenta la información común de la población y permitiendo una mejora en la eficacia promedio del 35 % del EER.

Además es importante notar que ambos enfoques contienen información discriminatoria complementaria, responsable de la mejora (13 % EER) en la fusión lineal de la puntuación, reafirmando el provecho de la utilización de la información temporal.

3.3. Esquema del sistema de verificación del locutor sobre el marco binario

En este epígrafe se describe, a través de esquemas, el sistema de verificación del locutor independiente del texto que integra los métodos propuestos en la tesis, el cual

puede ser utilizado en aplicaciones reales.

Los métodos propuestos también necesitan, antes de realizar el proceso de verificación, entrenar a priori un conjunto de parámetros con sus datos característicos correspondientes. En primer lugar se necesita obtener el Modelo Generador (fig. 3.4), partiendo del mismo UBM obtenido en el enfoque i-vector, se realiza una selección de las componentes Gaussianas más representativas de la población, para luego dividir cada región acústica en subregiones llamadas especificidades.

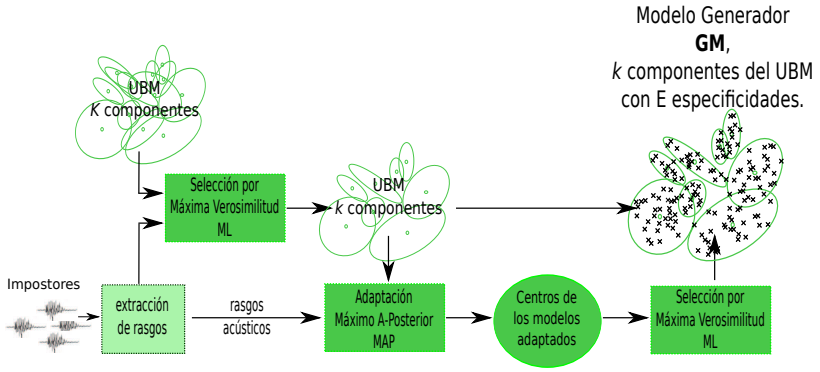


Figura 3.4: Modelo Generador

Luego para compensar la variabilidad de sesión es necesario obtener la matriz de proyección (B) del S-LDA y el modelo PLDA (B-PLDA).

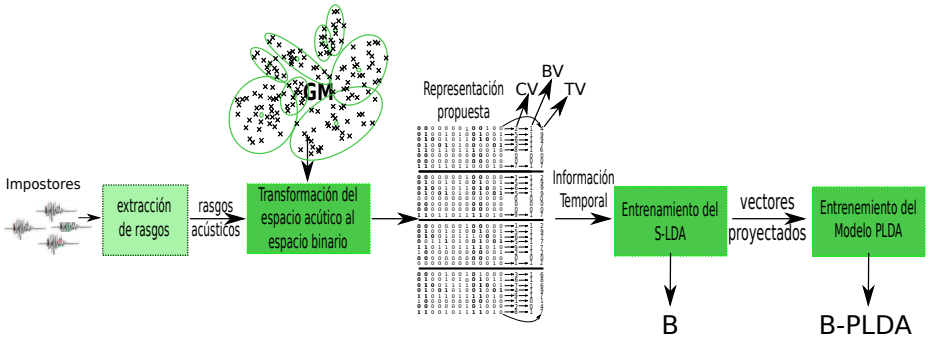


Figura 3.5: Parámetros para la compensación de la variabilidad

A partir de contar con los parámetros necesarios (fig. 3.5), se alcanza entonces el esquema del sistema de verificación del locutor independiente del texto.

Obsérvese en la fig. 3.6, que la configuración puede ser cambiada, se puede utilizar otra representación de la expresión de voz (CV o BV), si se desea utilizar la medida del coseno en la verificación, se sustituye el PLDA por la similitud por coseno o por otra función de similitud desarrollada. Además se puede sustituir el método S-LDA por el

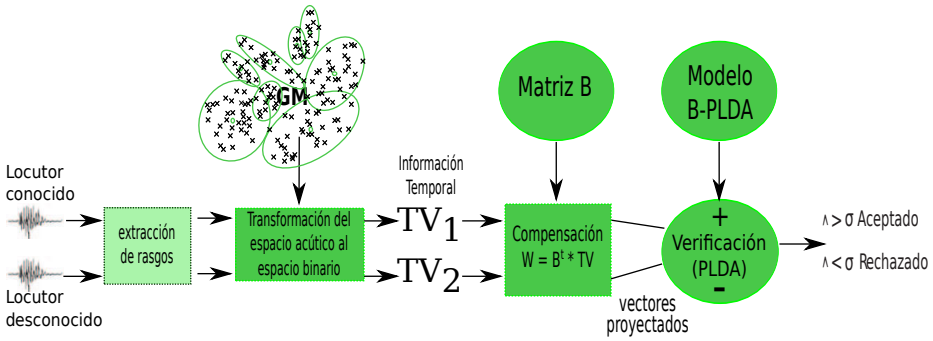


Figura 3.6: Sistema de verificación del locutor

método C&WCCN, el C&NAP o combinaciones de ellos, siempre con su respectiva proyección.

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

La presente investigación estuvo dirigida a fortalecer el desempeño de los algoritmos para el reconocimiento del locutor independiente del texto, obteniéndose como principales resultados:

- Una nueva representación binaria de la expresión de voz, capaz de contener los eventos discriminatorios desde los frecuentes hasta los pocos frecuentes.
- Se introdujo la información común en la matriz de dispersión intra-clase mejorando el desempeño de los algoritmos de compensación C&NAP, C&WCCN y S-LDA.
- Se desarrolló un método para obtener la información temporal de la expresión de voz y se incorporó en la verificación del locutor.

A continuación se describen los resultados obtenidos en la investigación.

Se desarrolló un método para evaluar la presencia de información redundante en el espacio de los super-vectores, que consistió en la reducción de la dimensión de dicho espacio, utilizando algoritmos basados en técnicas no lineales de reducción: Isomap y Laplaciano. Los resultados de este método mostraron claramente la información redundante presente en los super-vectores, utilizando la técnica Isomap se logró una reducción de dimensión en un factor de cuatro y no hubo prácticamente pérdidas en términos de eficacia. Además, utilizando la técnica del Laplaciano para reducir la dimensión en un factor de dos, se superó el resultado obtenido por una técnica lineal de reducción (PCA), lo que muestra la importancia de tener presente la naturaleza interna de los datos, como lo hace el enfoque topológico sobre la voz. Este resultado publicado en [37] demostró la gran cantidad de información redundante dentro del

espacio de los super-vectores GMM, lo cual confirmó la necesidad de buscar nuevas alternativas de representación menos redundantes, para el reconocimiento del locutor. Este resultado constituyó uno de los primeros pasos en el campo de reconocimiento del locutor, donde se utilizó la información topológica contenida en el espacio acústico.

Dentro de un nuevo enfoque basado en una representación binaria de la expresión de voz para reconocimiento del locutor, se propuso un método para la obtención del modelo generador y una nueva medida de similitud asociada con una representación global (vector acumulativo) de la información existente en la matriz binaria, dicha representación tiene en cuenta los eventos discriminatorios, desde los más frecuentes hasta los pocos frecuentes, en una señal de voz de un locutor. Ambos resultados fueron publicados en [11] y [36] y al compararlos con los métodos reportados en la literatura actual, se pudo comprobar que requieren considerablemente menos recursos de cómputo y memoria, mostrando un nivel de rendimiento comparable con el enfoque GMM-MAP. Este enfoque binario abrió nuevos caminos para continuar las investigaciones en el enfrentamiento a los problemas de variabilidad de sesión y a la explotación de la información temporal existente en la voz, para el reconocimiento del locutor.

Se propusieron nuevos métodos para la compensación de la variabilidad de sesión en el reconocimiento del locutor, asociada a la nueva representación binaria de la voz. Se propuso una nueva variante de la matriz de dispersión intra-clase para los vectores acumulativos e i-vector, que es capaz de contener no sólo los atributos no deseados, sino también la información común a los locutores, que tampoco es deseada. Se logró una disminución relativa del 47.7 % en el EER promedio de la verificación del locutor utilizando la representación binaria. Este resultado permitió comprobar la importancia de incluir la información común en la matriz de dispersión intra-clase, lo que confirmó el análisis de la varianza espectral realizado. También se propuso un método simple (máscara) basado en la varianza de los vectores acumulativos de los locutores, para seleccionar las especificidades más discriminatorias, capaz de eliminar un tercio de las especificidades del GM lográndose la mejor eficacia en el reconocimiento del locutor. Al aplicar los métodos propuestos en el marco i-vector, se logró una pequeña mejora en el rendimiento del 5 % en el EER promedio, porque este dominio presenta menos información común, a diferencia del dominio del vector acumulativo, donde su contenido se relaciona con las repeticiones de las especificidades. Se observó, al fusionarse las puntuaciones obtenidas, que ambos dominios contienen información complementaria. Estos resultados han sido publicados en [39] y [62].

Se propusieron métodos para capturar e incluir, la información temporal existente en las expresiones de voz, en el reconocimiento del locutor. Se propuso un método para extraer la información temporal a nivel de segmentos de la voz, logrando que el vector acumulativo de un segmento reflejara la distribución de las especificidades relativas al contenido fonético, lo cual es útil y aplicable en diferentes áreas de procesamiento de la voz y específicamente en el reconocimiento del locutor son características discriminatorias y robustas ante los efectos de ruido sobre la señal. Este resultado publicado en [11, 36] y [38] demostró un nivel de rendimiento comparable con el enfoque GMM-MAP, con menos recursos de computo. Se propuso un método para obtener una representación dinámica de la expresión de voz que refleje la información temporal a nivel de tramas, que demostró una mejora de la eficacia (4.09 % de EER) relativa a la obtenida con los vectores acumulativos (4.54 % de EER). Se aplicó a dichas representaciones temporales de la voz una nueva variante del método de compensación de variabilidad ya propuesto, logrando una mejora apreciable (14 %) en la eficacia. Se observó, al fusionarse las puntuaciones obtenidas, que ambos dominios contienen información complementaria [38]. La representación de la información temporal sobre el enfoque binario, constituye un nuevo paso en el reconocimiento del locutor el cual es imposible obtener en el enfoque i-vector.

Recomendaciones

Con los resultados obtenidos no se concluye el trabajo en esta temática. Del estudio realizado se derivan algunos trabajos futuros a modo de recomendaciones:

- Un inconveniente de estas técnicas de compensación, es que su rendimiento se ve afectado si el número de muestras para el entrenamiento de las matrices de proyección no es lo suficientemente grande. Los resultados pueden ser mejorados si se incrementan las bases de datos de entrenamiento.

- Proponer y evaluar un método de selección de los rasgos binarios más discriminatorios de la representación binaria a partir de los métodos presentados en [78], denominados Boosted Slice Classifiers.

- Realizar un análisis del algoritmo propuesto para obtener la implementación más eficiente que se pueda alcanzar y una posible versión paralela.

- Evaluar el desempeño de las representaciones temporales propuestas ante la presencia de ruido.

- Evaluar el comportamiento de la incorporación de la información común en los métodos de compensación de la variabilidad de sesión en ambientes ruidosos.

Bibliografía

[1] Anguera, X. and Bonastre, J. (2010). A novel speaker binary key derived from anchor models, in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2118–2121.

[2] Aronowitz, H. and Barkan, O. (2012). Efficient approximated i-vector extraction, in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012*, 4789–4792.

[3] Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, *The Journal of the Acoustical Society of America*, 55, no. 6.

[4] Auckenthaler, R., Carey, M. J., and Lloyd-Thomas, H. (2000). Score normalization for text-independent speaker verification systems, *Digital Signal Processing*, 10, no. 1-3, 42–54.

[5] Baker, B., Vogt, R., McLaren, M., and Sridharan, S. (2009). Scatter difference NAP for SVM speaker recognition, in *Advances in Biometrics, Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings*, 464–473.

[6] Belkin, M. and Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering, in *Advances in Neural Information Processing Systems: Natural and Synthetic, NIPS, December 3-8, 2001, Vancouver, British Columbia, Canada*, 585–591.

[7] Betser, M., Bimbot, F., Ben, M., and Gravier, G. (2004). Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted gmms, in *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*.

[8] Bimbot, F., Bonastre, J., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacrétaz, D., and Reynolds, D. A. (2004). A tutorial on text-independent speaker verification, *EURASIP J. Adv. Sig. Proc.*, 2004, no. 4, 430–451.

[9] Bishop, C. M. and Nasrabadi, N. M. (2007). Pattern recognition and machine learning in *J. Electronic Imaging*, 16, 049901.

[10] Bonastre, J., Bousquet, P., Matrouf, D., and Miró, X. A. (2011). Discriminant binary data representation for speaker recognition, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic*, 5284–5287.

[11] Bonastre, J., Miró, X. A., Hernández-Sierra, G., and Bousquet, P. (2011). Speaker modeling using local binary decisions, in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 13–16.

[12] Bonastre, J., Wils, F., and Meignier, S. (2005). Alize, a free toolkit for speaker recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005*, 737–740.

[13] Bousquet, P., Bonastre, J., and Matrouf, D. (2013). Identify the benefits of the different steps in an i-vector based speaker verification system, in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part II*, 278–285.

[14] Bousquet, P., Larcher, A., Matrouf, D., Bonastre, J., and Plchot, O. (2012). Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis, in *Odyssey 2012: The Speaker and Language Recognition Workshop, Singapore, June 25-28, 2012*, 157–164.

[15] Bousquet, P., Matrouf, D., and Bonastre, J. (2011). Intersession compensation and scoring methods in the i-vectors space for speaker recognition,

in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 485–488.

[16] Bricker, P., Gnanadesikan, R., Mathews, M., Pruzansky, S., Tukey, P., Wachter, K., and Warner, J. (1971). Statistical techniques for talker identification, *Bell System Technical Journal*, 4, 1427–1454.

[17] Brümmer, N. and de Villiers, E. (2010). The speaker partitioning problem, in *Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 28 - July 1, 2010*, 34.

[18] Burget, L., Fapso, M., Hubeika, V., Glembek, O., Karafiát, M., Kockmann, M., Matejka, P., Schwarz, P., and Cernocký, J. (2009). BUT system for NIST 2008 speaker recognition evaluation, in *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, 2335–2338.

[19] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Jones, D. A., and Leek, T. R. (2003). Phonetic speaker recognition with support vector machines, in *Advances in Neural Information Processing Systems 16, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada*.

[20] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., and Torres-Carrasquillo, P. A. (2006). Support vector machines for speaker and language recognition, *Computer Speech & Language*, 20, no. 2-3, 210–229.

[21] Campbell, W. M., Sturim, D. E., and Reynolds, D. A. (2006). Support vector machines using gmm supervectors for speaker verification, *IEEE Signal Process. Lett.*, 13, no. 5, 308–311.

[22] Cumani, S., Brummer, N., Burget, L., Laface, P., Plchot, O., and Vasilakakis, V. (2013). Pairwise discriminative speaker verification in the i-vector space, *IEEE Transactions on Audio, Speech & Language Processing*, 21, no. 6, 1217–1227.

[23] Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28, no. 4, 357–366.

[24] Dehak, N. and Chollet, G. June 2006 Support vector gmms for speaker verification, in *Proceedings of IEEE Odyssey: The Speaker and Language Recognition Workshop (Odyssey 2006)*, 1–4.

[25] Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., and Dumouchel, P. (2009). Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification, in *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, 1559–1562.

[26] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-end factor analysis for speaker verification, *IEEE Transactions on Audio, Speech & Language Processing*, 19, no. 4, 788–798.

[27] Fauve, B. G. B., Matrouf, D., Scheffer, N., Bonastre, J., and Mason, J. S. D. (2007). State-of-the-art performance in text-independent speaker verification through open-source software, *IEEE Transactions on Audio, Speech & Language Processing*, 15, no. 7, 1960–1968.

[28] Ferrer, L., Shriberg, E., Kajarekar, S. S., and Sönmez, M. K. (2007). Parameterization of prosodic feature distributions for SVM modeling in speaker recognition, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007, Honolulu, Hawaii, USA, April 15-20, 2007*, 233–236.

[29] Furui, S. (1981). Cepstral analysis technique for automatic speaker verification, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29, no. 2, 254–272.

[30] Furui, S. (1996). *An Overview of Speaker Recognition Technology*, 355 of *The Kluwer International Series in Engineering and Computer Science*. Springer US.

[31] Garcia-Romero, D. and Espy-Wilson, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems, in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 249–252.

[32] Glembek, O., Burget, L., Matejka, P., Karafiát, M., and Kenny, P. (2011). Simplification and optimization of i-vector extraction, in *Proceedings of*

the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic, 4516–4519.

[33] Gray C.H., K. G. (1944). Voiceprint identification, *Bell Telephone Laboratories Report, Bell Laboratories*.

[34] Hatch, A. O., Kajarekar, S. S., and Stolcke, A. (2006). Within-class covariance normalization for svm-based speaker recognition, in *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*.

[35] Heck, L. P. and Weintraub, M. (1997). Handset-dependent background models for robust text-independent speaker recognition, in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '97, Munich, Germany, April 21-24, 1997*, 1071–1074.

[36] Hernández-Sierra, G., Bonastre, J., and de Lara, J. R. C. (2012). Speaker recognition using a binary representation and specificities models, in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 17th Iberoamerican Congress, CIARP 2012, Buenos Aires, Argentina, September 3-6, 2012. Proceedings*, 732–739.

[37] Hernández-Sierra, G., Bonastre, J., Matrouf, D., and Calvo, J. R. (2010). Topological representation of speech for speaker recognition, in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2134–2137.

[38] Hernández-Sierra, G., Calvo, J. R., and Bonastre, J.-F. (2014). Temporal information in a binary framework for speaker recognition, to be published, in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (Bayro-Corrochano, E. and Hancock, E., eds.), Lecture Notes in Computer Science, Springer Berlin Heidelberg.

[39] Hernández-Sierra, G., Calvo, J. R., Bonastre, J., and Bousquet, P. (2014). Session compensation using binary speech representation for speaker recognition, *Pattern Recognition Letters*, 49, 17–23.

[40] Hernández-Sierra, G., Calvo, J. R., Reyes, F. J., and Fernández, R. (2009). Simple noise robust feature vector selection method for speaker recognition, in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, 14th Iberoamerican Conference on Pattern Recognition, CIARP 2009, Guadalajara, Jalisco, Mexico, November 15-18, 2009. Proceedings*, 313–320.

[41] Huang, X., Acero, A., and Hon, H. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR.

[42] Hume, J. (1997). Wavelet-like regression features in the cepstral domain for speaker recognition, in *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*.

[43] Jain, A., Flynn, P., and Ross, A. (2007). *Handbook of Biometrics*. Springer.

[44] Jansen, A. and Niyogi, P. (2005). A geometric perspective on speech sounds, *University of Chicago, Tech. Rep*.

[45] Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., and Mason, M. (2011). i-vector based speaker recognition on short utterances, in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 2341–2344.

[46] Karam, Z. N. and Campbell, W. M. (2007). A new kernel for SVM MLLR based speaker recognition, in *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, 290–293.

[47] Kenny, P. (2005). Joint factor analysis of speaker and session variability: Theory and algorithms, tech. rep., Montreal, CRIM, (2005).

[48] Kenny, P. (2010). Bayesian speaker verification with heavy-tailed priors, in *Odyssey 2010: The Speaker and Language Recognition Workshop, Brno, Czech Republic, June 28 - July 1, 2010*, 14.

[49] Kenny, P., Boulianne, G., and Dumouchel, P. (2005). Eigenvoice modeling with sparse training data, *IEEE Transactions on Speech and Audio Processing*, 13, no. 3, 345–354.

[50] Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P. (2005). Factor analysis simplified, in *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005*, 637–640.

[51] Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P. (2007). Speaker and session variability in gmm-based speaker verification, *IEEE Transactions on Audio, Speech & Language Processing*, 15, no. 4, 1448–1460.

[52] Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P. (2008). A study of interspeaker variability in speaker verification, *IEEE Transactions on Audio, Speech & Language Processing*, 16, no. 5, 980–988.

[53] Khoury, E., El Shafey, L., and Marcel, S. dec 2012 The idiap speaker recognition evaluation system at nist sre 2012, in *NIST Speaker Recognition Conference*, NIST.

[54] Kinnunen, T. Dec 2003 Spectral features for automatic text-independent speaker recognition, Tesis de Maestría, University of Joensuu, Department of Computer Science, P.O. Box 111, FIN-80101 Joensuu, Finland.

[55] Kinnunen, T. and Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors, *Speech Communication*, 52, no. 1, 12–40.

[56] Lee, K., You, C., Li, H., Kinnunen, T., and Zhu, D. (2008). Characterizing speech utterances for speaker verification with sequence kernel SVM, in *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, 1397–1400.

[57] Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models, *Computer Speech & Language*, 9, no. 2, 171–185.

[58] Li, K. and Hughes, G. W. (1974). Talker differences as they appear in correlation matrices of continuous speech spectra, *The Journal of the Acoustical Society of America*, 55, no. 4.

[59] Li, M., Tsiartas, A., Segbroeck, M. V., and Narayanan, S. S. (2013). Speaker verification using simplified and supervised i-vector modeling, in *IEEE*

International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013, 7199–7203.

[60] Mak, M., Hsiao, R. W., and Mak, B. (2006). A comparison of various adaptation methods for speaker verification with limited enrollment data, in *2006 IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP 2006, Toulouse, France, May 14-19, 2006*, 929–932.

[61] Martin, A. F., Doddington, G. R., Kamm, T., Ordowski, M., and Przybocki, M. A. (1997). The DET curve in assessment of detection task performance, in *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*.

[62] Martínez, M. A., Hernández-Sierra, G., and de Lara, J. R. C. (2013). Speaker verification using accumulative vectors with support vector machines, in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part II*, 350–357.

[63] Miró, X. A. and Bonastre, J. (2011). Fast speaker diarization based on binary keys, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic*, 4428–4431.

[64] Moreno, A., Comeyne, R., Haslam, K., van den Heuvel, H., Höge, H., Horbach, S., and Micca, G. (2000). SALA: speechdat across latin america. results of the first phase, in *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, 31 May - June 2, 2000, Athens, Greece*.

[65] Ortega-Garcia, J., Gonzalez-Rodriguez, J., and Marrero-Aguiar, V. (2000). AHUMADA: A large speech corpus in spanish for speaker characterization and identification, *Speech Communication*, 31, no. 2-3, 255–264.

[66] Oshaughnessy, D. oct 1986 Speaker recognition, *IEEE Acoustics, Speech, and Signal Processing, ASSP Magazine*, 3, 4–17.

[67] Prince, S. J. D. (2012). *Computer Vision: Models, Learning, and Inference*. New York, NY, USA: Cambridge University Press, 1st ed.

[68] Prince, S. J. D. and Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity, in *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, 1–8.

[69] Przybocki, M., Martin, A., and Le, A. June 2006 Nist speaker recognition evaluation chronicles - part 2, in *Odyssey 2006: The Speaker and Language Recognition Workshop, 2006.*, 1–6.

[70] Rabiner, L. R. and Juang, B. (1993). *Fundamentals of speech recognition*. Prentice Hall signal processing series, Prentice Hall.

[71] Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification, *Journal of the Royal Statistical Society - Series B*, 10, no. 2, 159–203.

[72] Reynolds, D. A. (1994). Experimental evaluation of features for robust speaker identification, *IEEE Transactions on Speech and Audio Processing*, 2, no. 4, 639–643.

[73] Reynolds, D. A. (1996). The effects of handset variability on speaker recognition performance: experiments on the switchboard corpus, in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, ICASSP '96, Atlanta, Georgia, USA, May 7-10, 1996*, 113–116.

[74] Reynolds, D. A. (2003). Channel robust speaker verification via feature mapping, in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03, Hong Kong, April 6-10, 2003*, 53–56.

[75] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models, *Digital Signal Processing*, 10, no. 1-3, 19–41.

[76] Reynolds, D. A. and Rose, R. C. (1995). Robust text-independent speaker identification using gaussian mixture speaker models, *IEEE Transactions on Speech and Audio Processing*, 3, no. 1, 72–83.

[77] Rosenberg, A. (1973). Listener performance in speaker verification tasks, *IEEE Transactions on Audio and Electroacoustics*, 21, no. 3, 221–225.

[78] Roy, A., Magimai-Doss, M., and Marcel, S. (2012). A fast parts-based approach to speaker verification using boosted slice classifiers, *IEEE Transactions on Information Forensics and Security*, 7, no. 1, 241–254.

[79] Scheffer, N., Ferrer, L., Graciarena, M., Kajarekar, S. S., Shriberg, E., and Stolcke, A. (2011). The SRI NIST 2010 speaker recognition evaluation system, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic*, 5292–5295.

[80] Solomonoff, A., Campbell, W. M., and Boardman, I. (2005). Advances in channel compensation for SVM speaker recognition, in *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005*, 629–632.

[81] Solomonoff, A., Quillen, C., and Campbell, W. M. (2004). Channel compensation for SVM speaker recognition, in *ODYSSEY 2004 - The Speaker and Language Recognition Workshop, Toledo, Spain, May 31 - June 3, 2004*, 57–62.

[82] Stevens, K. N. July 1998 *Acoustic Phonetics*. Cambridge, Mass.: MIT Press.

[83] Su, L., Fu, K., and ENGINEERING., P. U. L. I. S. O. E. (1973). *Automatic Speaker Identification Using Nasal Spectra and Nasal Coarticulation as Acoustic Clues*. Memo (Purdue University), School of Electrical Engineering, Purdue University.

[84] Tenenbaum, J. B., Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction, *Science*, 290, no. 5500, 2319–2323.

[85] Vogt, R. and Sridharan, S. (2006). Experiments in session variability modelling for speaker verification, in *2006 IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP 2006, Toulouse, France, May 14-19, 2006*, 897–900.

[86] Vogt, R. and Sridharan, S. (2008). Explicit modelling of session variability for speaker verification, *Computer Speech & Language*, 22, no. 1, 17–38.

[87] Wolf, J. J. (1972). Efficient acoustic parameters for speaker recognition, *The Journal of the Acoustical Society of America*, 51, no. 6B.

[88] Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C. Y., and Shamma, S. A. (2011). Linear versus mel frequency cepstral coefficients for speaker recognition, in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2011, Waikoloa, HI, USA, December 11-15, 2011*, 559–564.

Producción científica del autor sobre el tema de la tesis

Publicaciones

1. Hernández-Sierra, G., Bonastre, J.F. y otros: "Topological representation of speech for speaker recognition". Conference of the International Speech Communication (INTERSPEECH'10), pp. 2134-2137, 2010. (Conference Proceeding Citation Index-Science, **ISI Web of Science, Scopus**)

2. Bonastre, J.F., Miro, X.A., Hernández-Sierra y otros: "Speaker modeling using local binary decisions". Conference of the International Speech Communication, INTERSPEECH, pp. 13-16, 2011. (Conference Proceeding Citation Index-Science, **ISI Web of Science, Scopus**)

3. Hernández-Sierra, G., Bonastre, J.F. y otros: "Speaker recognition using a binary representation and specificities models". Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 17th Iberoamerican Congress, CIARP 2012, Lecture Notes in Computer Science, Springer Berlin Heidelberg, ISSN 0302-9743, pp. 732-739, 2012. (Conference Proceeding Citation Index-Science, **ISI Web of Science, Scopus**)

4. Martínez, M. A., Hernández-Sierra, G., y Calvo, J. R. "Speaker Verification Using Accumulative Vectors with Support Vector Machines". Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 18th Iberoamerican Congress, CIARP 2013, Lecture Notes in Computer Science, Springer Berlin Heidelberg, ISSN 0302-9743, pp. 350-357, 2013. (Conference Proceeding Citation Index-Science, **ISI Web of Science, Scopus**)

5. Hernández-Sierra, G., Calvo, J. R. y otros: "Session compensation using binary speech representation for speaker recognition". Pattern Recognition Letters,

ISSN 0167-8655, Volume 49, pp. 17-23, 2014. (**ISI Web of Science, Scopus**)

Aprobadas para Publicación

1. Hernández-Sierra, G., Calvo, J. R., y Bonastre, J.F.: "Temporal Information in a binary framework for Speaker Recognition". Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 19th Iberoamerican Congress, CIARP 2014, Lecture Notes in Computer Science, Springer Berlin Heidelberg, ISSN 0302-9743. (Conference Proceeding Citation Index-Science, **ISI Web of Science, Scopus**).

Asesor de tesis de maestría

1. Ing. Argel González Padilla: "Nueva representación del habla para el reconocimiento del locutor". Instituto Superior Politécnico José Antonio Echeverría (ISPJAE), tesis de maestría, durante julio 2009–julio 2010, 2010.

Asesor de tesis de diploma

1. Leandro Barak Carrasquel: "Compensación de la variabilidad de sesión para los vectores acumulativos". Tesis de Diploma Curso 2012-2013, Licenciatura en Ciencias de la Computación, 2013.

2. Manuel Aguado Martínez: "Aplicación de Máquinas de Vectores Soporte a los Vectores Acumulativos en el Reconocimiento del Locutor". Tesis de Diploma Curso 2012-2013, Licenciatura en Ciencias de la Computación, Asesor: Lic. Gabriel Hernández Sierra. 2013.

Reportes técnicos

1. Gabriel Hernández-Sierra, José Ramón Calvo De Lara: "Topological representation of speech for speaker recognition". Serie Azul CENATAV ISSN 2072-6287, 2010.

Otras publicaciones

1. Argel González Padilla, Hernández-Sierra, G. "Nueva representación del habla para el reconocimiento del locutor". En: VI Congreso Internacional de Telemáti-

ca y Telecomunicaciones, CITTEL 2010, La Habana, Cuba, 29 de noviembre al 3 de diciembre de 2010.

Glosario de acrónimos

ASI	Identificación Automática del Locutor
ASR	Reconocimiento Automático del Locutor
ASV	Verificación Automática del Locutor
BK	Matriz Binaria Dispersa
BV	Vector Binario
CV	Vector Acumulativo
C&NAP	Proyección de Atributos no Deseados y Comunes
C&WCCN	Atributos Comunes en la Normalización de la Covarianza Intra-clase
DTW	Distorsión Dinámica en el Tiempo
DV	Vector Dinámico
EER	Tasa de Error
EM	Maximización de la Esperanza
FA	Análisis de Factores
GM	Modelo Generador
GMM	Modelos de Mezclas Gaussianas
GMM-UBM	Modelos de Mezclas Gaussianas – Modelo Universal de Fondo
GSL	Núcleo Lineal con Supervectores GMM-UBM
G-PLDA	Análisis de discriminante lineal Probabilístico Gaussiano
HMM	Modelos Ocultos de Markov
HT-PLDA	Análisis de discriminante lineal Probabilístico de Cola Pesada
H-norm	Normalización del Auricular
IS	Similitud de la Intersección
ISDS	Similitud de la Intersección y la Diferencia Simétrica
i-vector	Vector de Identidad
JFA	Análisis de Factores Conjunto
KL	Kullback-Liebler
LDA	Análisis de Discriminante Lineal
LFCC	Coeficientes Cepstrales de Frecuencia lineales

LPCC	Coeficientes Cepstral de Predicción lineal
MAP	Máximo A-Posterior
MDS	Escalado Multidimensional
MFCC	Coeficientes Cepstrales de Frecuencia Mel
ML	Máxima Verosimilitud
MLLR	Regresión Lineal de Máxima Verosimilitud
NAP	Proyección de Atributos No Deseados
NIST	Instituto Nacional de Estándares y Tecnología
PCA	Análisis de Componentes Principales
PLDA	Análisis de Discriminante Lineal Probabilístico
SRE	Evaluaciones de Reconocimiento del Locutor
SV	Super-Vectores
SVM	Máquina de Vectores de Soportes
S-LDA	Análisis de Discriminante Lineal Desplazado
T	Espacio de Variabilidad Total
TM	Modelo de Trayectoria
T-norm	Normalización de la prueba
UBM	Modelo Universal de Fondo
VQ	Cuantificación Vectorial
WCCN	Normalización de la Covarianza Intra-clase

Glosario de términos

distancia geodésica	En geometría, la línea geodésica se define como la línea de mínima longitud que une dos puntos en una superficie dada, y está contenida en esta superficie.
eficacia	Es la capacidad de alcanzar el efecto que espera o se desea tras la realización de una acción.
eficiencia	Es el uso racional de los medios para alcanzar un objetivo predeterminado.
eigenvoices	Término en inglés por el cual se conoce la matriz de proyección constituida por las principales direcciones de varianza de las representaciones de los locutores.
eigenchannel	Término en inglés por el cual se conoce la matriz de proyección constituida por las principales direcciones de varianza de las sesiones.
estadísticas suficientes	Son la estadística básica requerida para calcular los parámetros deseados de un GMM, el peso, la media y la varianza.

ANEXO A

Listado de los sistemas de reconocimiento del locutor

En la tabla se presentan algunos de los principales sistemas de reconocimiento del locutor.

Sistema	Compañía	País	Descripción
KIVOX	Agnitio	EE.UU.	Se trata de un sistema de verificación del locutor para bancos y otras entidades comerciales. También presta servicios al gobierno, la policía, el mercado forense y es capaz de realizar la verificación del locutor en cualquier idioma y canal telefónico.
ETF Verifier	Authentify	EE.UU.	Software para la identificación de personas utilizando la línea telefónica. Sus soluciones incluyen acceso remoto a redes informáticas, protección de transacciones de alto riesgo y otras.
VoiceVerified	CSIdentity	EE.UU.	Software de verificación de identidad, protección contra fraude y otros.
Loquendo	NUANCE	Italia	Ofrece reconocimiento de voz, síntesis de voz, verificación de locutor y aplicaciones de identificación.
Identivox	Universidad Politécnica de Madrid	España	Sistema basado en los GMM para el reconocimiento del locutor independiente del texto, uso forense, aplicando el modelo Bayesiano para las conclusiones.

ANEXO B

Herramientas para el desarrollo de sistemas de reconocimiento del locutor

1. CMU Sphinx: Librería de Reconocimiento del habla (código abierto), http://cmusphinx.org/

2. CSLU Toolkit: Herramientas para la exploración, del conocimiento y la investigación en la voz e interacción hombre máquina, http://cslu.cse.ogi.edu/toolkit/

3. LNKnet MIT Lincoln Laboratories Pattern Classification Software. LNKnet es un paquete de software que integra más de 22 redes neuronales, clasificación con maquinas de aprendizaje, métodos de agrupamientos y algoritmos de selección de rasgos, máquinas de soportes vectoriales y sencillos clasificadores bayesianos. Tiene una versión para Linux y una para Windows usando el ambiente de Cygwin, todas estas herramientas están creadas en lenguaje C, http://www.ll.mit.edu/IST/lnknet/

4. Cambridge HTK: Biblioteca de algoritmos de Modelos Ocultos de Markov, es una librería portable para la construcción y manipulación de los Modelos Ocultos de Harkov. El ATK es un API en tiempo real para HTK, http://htk.eng.cam.ac.uk/, http://mi.eng.cam.ac.uk/sjy/software.htm

5. Julius es una aplicación para reconocimiento de voz publicado como software libre con licencia BSD. Es un programa de reconocimiento de habla continua,

está basado en trigramas (3 gramas) y en el Modelo oculto de Márkov. Julius ha sido desarrollado como parte de un kit de software libre desde 1977 y el trabajo ha sido continuado por el consorcio de reconocimiento de habla continua (CSRC) en Japón de 2000 a 2003. La versión más reciente de Julius es la 4.2.

6. LIA, ALIZE es un conjunto de herramientas de código abierto desarrollado, desde 2004, para el reconocimiento del locutor. La última versión (3.0) incluye métodos del estado del arte, tales como Análisis de Factores Conjunto, modelado i-vector y Análisis de Discriminante Lineal Probabilístico. ALIZE es un proyecto de desarrollo iniciado por el consorcio ELISA y presenta una documentación muy bien detallada para su uso, también tiene las siguientes características:

   - es simple y fácil de entender,
   - tiene un nivel de funcionamiento que corresponde con el estado del arte actual, en términos de error pero también en términos de recursos de cómputo necesarios,
   - facilita el desarrollo de demostraciones en aplicaciones prácticas,
   - está programado en C++.

   http://www.lia.univ-avignon.fr/heberges/ALIZE/