UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

Reconocimiento del hablante usando i-vectors y PLDA

TESIS

Que para obtener el título de

Ingeniero en Computación

P R E S E N T A

Arturo Rivera García

DIRECTOR DE TESIS

Dr. Abel Herrera Camacho

Ciudad Universitaria, Cd. Mx., 2020

AGRADECIMIENTOS

Gracias a יהוה principalmente quien me ha permitido la vida y ha dado sabiduría para poder concluir mis estudios y haber finalizado este trabajo de investigación, gracias a mi padres Abel y Mercedes quienes siempre me dieron el amor y el apoyo para salir adelante. A mi madre Rosa quien dio todo su esfuerzo y dedicación para mi bienestar y mi carrera. A mis hermanos quien siempre me han apoyado incondicionalmente. A mis tios y tias quienes siempre han visto por mí y mis necesidades y a todos las personas quienes siempre me han apoyado y han estado conmigo desde el inicio de este trabajo hasta el fin, les dedico este trabajo. Muchas gracias.

INTRODUCCIÓN

Con el paso del tiempo los sistemas biométricos han evolucionado más debido a el avance de la tecnología que hace posible que estos softwares sean más robustos y satisfactorios para los usuarios, sin embargo también la necesidad de las personas de utilizar sistemas de verificación cada vez más robustos y confiables ha provocado dentro del área de la inteligencia artificial que surjan algoritmos cada vez de mayor precisión, velocidad, seguridad, etc., esto ha hecho que conforme al tiempo se amplie el campo de investigación y desarrollo de estos sistemas para crear soluciones, algunos de estos sistemas están enfocados al reconocimiento facial, otros al reconocimiento dactilar, voz, iris, etc., que mejoran y crean una mayor confiabilidad frente a los métodos tradicionales de verificacion como lo es la firma, preguntas frecuentes, un NIP, imitar una frase clave y burlar un sistema de seguridad, etc.

Uno de lo mayores problemas que se han presentado en el área de forense es en el reconocimiento del hablante, debido a que cualquiera está expuesto a recibir una llamada de extorsión donde se podría aprovechar para grabar la llamada y extraer la voz del extorsionador, o bien también el caso donde se presenten pruebas de grabacion en un juicio donde se podría comprobar si una persona es dueña o no del contenido de la grabación, en éstas y más situaciones es donde se podría brindar soluciones para el área de reconocmiento del hablante que ayudarían a acelerar algun proceso juridico brindando una mayor confiabilidad atacando toda la variabilidad presente en las voces que hace difícil que un oido humano logre analizar y evaluar.

Es por esta razón que en esta tesis se pretende desarrollar un sistema de reconocimiento del hablante, donde el interés no es saber qué se pronuncia sino quién es el que habla. Para esto hoy en día existen una gran variedad de métodos para crear un sistema de reconocimiento del hablante, tales como las redes neuronales, modelos de mezclas gaussianas, i-vectors, etc., para este trabajo se desarrollará el sistema usando el método de i-vectors el cual es considerada una de las tecnicas del estado del arte recientemente creada en la decada del 2010 con el fin de analizar toda la variabilidad no deseada en una señal de voz, además logrando trabajar con pequeños segmentos de duración dentro un vector caracteristico haciendo que sea más rapido y eficiente frente a su antecesor del modelo de mezclas gaussianas y estando a la par frente algunas arquitecturas de redes neuronales.

JUSTIFICACIÓN

Con la tecnología de hoy en día se puede crear un sistema de reconocimiento del hablante capaz de reconocer con escasos segundos de grabación a una persona sin importar la variablidad encontrada tanto en la voz como en canal, sin tomar tampoco en cuenta si la fuente viene de un celular, teléfono fijo o teléfono público, trayendo beneficios al área forense donde es de gran importancia contar con sistemas confiables, precisos y actuales, pertenecientes al estado del arte que por supuesto mejoren a los métodos anteriores.

OBJETIVO GENERAL

Esta tesis tiene como objetivo hacer un sistema de reconocimiento del hablante basado en i-vectors junto con PLDA, las cuales son cosideradas técnicas del estado del arte. PLDA ayudará para la evaluación y para la reducción de la dimensión de los vectores, la combinación de estas técnicas según los autores prometen ser más precisas a la hora de trabajar con señales de voz de poca duración frente a una verificación de duración variable, para verificar esto se comparará el desempeño de i-vectors frente a su antecesor GMM, realizando distintas metodos de evaluacion, como i-vectors con distancia coseno, con PLDA y PLDA con reducción de la dimensión. El sistema a desarrollar en está tesis está pensado para trabajar con el español de la Ciudad de México utilizando para la etapa de entrenamiento y verificación el “Corpus Valquiria”, el cuál fue diseñado y creado por el “laboratorio de tecnologías del lenguaje” de la Facultad de Ingeniería y el grupo de ingeniería lingüística durante los años 2015 y 2016 y que cuenta con las grabaciones de hombres y mujeres originadas desde un teléfono público, fijo y de celular, por cual contiene un ambiente real para el desarrollo y comparación del sistema a desarrollar en esta tesis.

HIPÓTESIS

La combinación de i-vectors y PLDA para hacer el reconocimiento del hablante mostrará mejores resultados frente a su antecesor GMM y se espera mejor aún que el agregar una reducción de dimension a los vectores se presente un mejor desempeño frente a los otros métodos de evaluación de i-vectors y claro frente al sistema GMM debido principalmente a dos razones, la primera es que se trabajará con segmentos de voz de poca duración lo cual hace difícil contar con poca información de la voz de una persona y la segunda ligada a la primera es no contar con una base de datos más amplia en grabaciones.

Contenido

Listado de Figuras

Listado de tablas

Capítulo 1. Estado del Arte

La investigación y desarrollo sobre el área de reconocimiento del hablante ha mostrado grandes resultados con el avance del tiempo. Es importante mencionar que la evolución de la tecnología ha ayudado mucho a estos sistemas a mejorar en su manera de ser entrenados y en los distintos métodos para reconocer la voz de las personas. Acerca de estos avances que han existido en las últimas décadas, se mencionarán solo algunos de ellos con sus aportaciones más importantes que han logrado que se mejoren estos sistemas dando paso a crear y/o mejorar métodos de reconocimiento llegando hasta a i-vector el cual es considerado un método del estado del arte de esta última década y es el que se desarrollará dentro de esta tesis.

La investigación sobre el reconocimiento del hablante abarca alrededor de 5 décadas, con el objetivo de crear máquinas que puedan comunicarse con las personas de una manera natural, sin importar las condiciones en las que se requiera que esta máquina deba hablar, tal y como lo hacen los seres humanos.

1.1 Década de los 60

Empezando por la década de los 60 a los 70. Para el año de 1963 Pruzansky en los Laboratorios Bell fue uno de los primeros en iniciar la investigación para el desarrollo de un sistema de reconocimiento del hablante, usando bancos de filtros y correlacionando dos espectrogramas digitales para ver su similitud [1].

La función que tenían estos bancos de filtros era que separaban la señal entrada en múltiples componentes, donde cada componentes llevaba una solo frecuencia de sub-banda de la señal original, a este proceso de descomposición también se le llama *análisis*, que nos da como salida una señal de sub-banda con tantas sub-bandas, ya que hay filtros en dicho banco [2,6], posteriormente a esta señal filtrada se le aplicaba la transformada de Fourier para poder representar su espectrograma y ver la evolución de la señal a través del tiempo, una vez hecho esto se hacía la correlación entre dos espectrogramas y obtener su similitud para determinar quién es la persona.

Para 1964 Doddigton en Texas Instrument (TI) hizo la misma implementación de Prouzansky en los Laboratorios Bell, pero en su experimento él reemplazó los bancos de filtros por análisis de formantes [3], ya que en un formante se demostró que está la mayor intensidad de energía en el espectro de un sonido que se da en una determinada frecuencia, técnicamente los formantes son bandas de frecuencia donde se concentra la mayor parte de la energía sonora de un sonido, para posteriormente hacer la evaluación.

Más tarde, Texas Instruments systems (TI) construyó el primer sistema automatizado de verificación del hablante, este sistema se construyó de un conjunto de 16 palabras monosilábicas y para la prueba se utilizaban expresiones aleatorias formadas de 4 palabras, este sistema utilizaba bancos de filtros para poder hacer el análisis espectral de la señal, que era la oración formada de todas las posibles combinaciones. Se realizaron millones de pruebas durante un período de 6 años para cientos de hablantes [3].

1.2 Década de los 70

Para la década de los 70 fue donde surgió uno de más grandes problemas que ha existido en el reconocimiento de una persona y que actualmente sigue siendo, que son las variaciones del *intra-hablante*, para esto, Furui fue uno de los principales investigadores que se dieron a la tarea de analizar este problema a inicios de la década de los sesenta y crear una solución. En 1975 Furui en los laboratorios Bell [28] propone utilizar una nueva técnica para mejorar la robustez del sistema contra las distorsiones que puede haber en una llamada telefónica, propone utilizar la combinación de coeficientes cepstral. Estos coeficientes a final de cuentas son números que nos dan una representación valiosa del tracto vocal de la voz, dando solo la parte más representativa de la señal de interés, los cuales se convirtieron importantes no solo para sistemas de reconocimiento del hablante sino también para sistemas de reconocedores de voz [5]. Se hablará de estos coeficientes más a detalle a lo largo de esta tesis.

Para continuar con los siguientes avances que surgieron en las décadas posteriores es necesario definir dos enfoques muy importantes para el entendimiento de estos nuevos sistemas que surgieron en las siguientes décadas, ya que muchos de estos métodos fueron orientados a utilizar estos dos enfoques para resolver problemas que se presentaban en el reconocimiento de hablante, y son conocidos como: métodos dependientes del texto y métodos independientes del texto.

En el primer enfoque llamado *métodos dependientes del texto,* lo que se busca es que cada hablante de nuestro conjunto total a identificar grabe o nos proporcionen ciertas palabras claves u oraciones que posteriormente van a ser utilizadas para el entrenamiento del sistema. Estas frases que deben ser de corta duración, fáciles, sencillas y claras de pronunciar por los participantes para que ellos las graben, o bien, que ellos mismos proporcionen las palabras que ellos quieren grabar. Ahora bien, las mismas palabras que los hablantes han proporcionado deben de ser las mismas palabras utilizadas para el reconocimiento. Estos métodos generalmente se basa en técnicas de "*template-matching*" (técnicas de coincidencia de etiqueta), las cuales consisten en que los ejes de tiempo de una muestra de voz de entrada y cada modelo de referencia de los hablantes estén alineados, y la similitud entre ellos se acumula desde el principio hasta el final de la expresión [3], este tipo de enfoque puede ser muy bien utilizado para explotar la individualidad de la voz asociada con cada fonema o sílaba. En general este tipo de técnica alcanza un funcionamiento de rendimiento más alto de reconocimiento que los métodos independientes del texto, pero éste resulta no ser muy real, ya que no es aplicable a muchos casos de la vida real debido a la falta de confiabilidad que se puede presentar si otra persona reproduce la voz grabada de alguna persona dentro de la base de datos, por lo cual surge el siguiente enfoque llamado métodos independientes del texto.

Éste segundo enfoque surge debido a la escasez de aplicaciones en las que los métodos dependientes de texto no dan un buen desempeño y no son muy eficientes en áreas como lo forense y de vigilancia en los que se no pueden utilizar palabras claves predeterminadas debido a que una persona podría no ser detectada correctamente o bien, poder burlar muy fácilmente el sistema. Es por esto que se nace este enfoque, para hacer más robustos los sistemas de reconocimiento.

Los *métodos independientes de texto*, buscan ser una técnica similar a como reconoce el ser humano a una persona, ya que el ser humano es capaz de reconocer a una persona sin importar las palabras y expresiones que ésta pueda decir; por lo tanto el método independiente del texto ha tenido recientemente una atención mayor para aplicaciones de reconocimiento del hablante.

Otra ventaja de los métodos independientes del texto es que son secuenciales, esto quiere decir, que hasta que se alcance el nivel significativo deseado se reconocerá a la persona sin la necesidad de que el hablante tenga que repetir las palabras claves una y otra vez. En estos sistemas las palabras u oraciones no pueden ser predecidas por el sistema de reconocimiento, dado que es imposible modelar o coincidir los eventos de voz en el nivel de palabras u oraciones.

Con estos dos enfoques para poder reconocer al hablante se mejoraron los sistemas y los objetivos, pero también a la par los métodos para implementar estos enfoques. Esto logró una mayor robustez en los sistemas para tener una mejor precisión en los resultados, por lo tanto, se crearon métodos más eficientes para poder calcular vectores que nos puedan representar cada vez mejor la voz de una persona.

1.3 Década de los 80

En la década de 1980 se utiliza una técnica de reconocimiento del hablante dependiente de texto basado en Modelos Ocultos de Márkov (MOM)[28], debido a que los MOM no solo eran buenos para el reconocimiento de voz sino también para el hablante ya que se puede modelar eficientemente la variación estadística espectral que existe en la voz de una persona, estos modelos siguen siendo un gran apoyo para crear sistemas de reconocedores de voz y del hablante, ya que es un proceso por el cual se observa el comportamiento del sistema de manera indirecta donde los estados del mismo permanecen ocultos para el observador, de ahí su nombre de Modelo Oculto de Márkov. En este modelo nos dice que la probabilidad de la transición de un estado a otro solo depende del estado actual y no de los estados anteriores.

El principal objetivo de este método es el de establecer un modelo del habla el cual se pueda obtener de pequeñas cantidades de información de la voz y así obtener modelos para cada palabra individual incluido el silencio, de tal manera que estos modelos se combinan en una secuencia de estados donde cada estado puede ser una palabra o un silencio, de esta forma se podría generar una oración dada una gramática previa. Del mismo modo se puede hacer un modelo para un hablante dada expresiones de su voz y en la etapa de evaluación lo que se buscará es reconocer la voz de la persona analizando su modelo y sus patrones de pronunciación. Estos patrones que se almacenan son sus expresiones de voz las cuales puede contener diferentes palabras, tomando en cuenta también el patrón de silencio para que el modelo sea más satisfactorio. En la figura 1 se muestra un modelo oculto de markov típico de izquierda a derecha en el que la probabilidad de pasar al siguiente estado depende del estado actual y no del anterior.

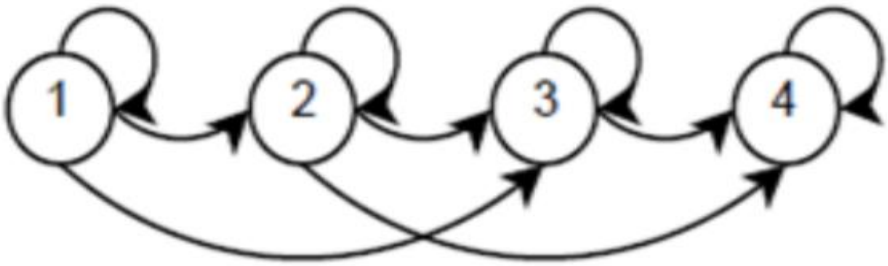

Figura 1. Modelo Oculto de Markov de izquierda a derecha [4].

Posteriormente en 1987 se crea otra técnica para el reconocimiento del hablante pero ahora para el método independiente del texto basado en Cuantización Vectorial (CV)[28] apoyándose en los MOM, esta técnica a diferencia de la primera es que dado

un conjunto de vectores de características de entrenamiento de corto tiempo de un hablante. Estos se pueden comprimir o reducir, para poder obtener solo algunos cuantos vectores que nos puedan representar al hablante de una manera más efectiva. El método basado CV trajo una gran ventaja sobre el sistema anterior, principalmente en el ahorro de memoria para los datos a guardar y por lo tanto también ahorra tiempo de procesamiento de los datos.

El método de CV realiza una etiqueta que corresponde a un vector código para cada muestra de la señal y mediante un algoritmo de entrenamiento se obtiene un vector final que representará a dicha palabra o fonema y la almacenará en un diccionario llamado “Libro de códigos”. Para cada conjunto de vectores característicos de entrenamiento de cada hablante y en la etapa del reconocimiento una expresión de entrada pasa por el mismo trato por el que se obtuvieron los vectores de entrenamiento, en la etapa de reconocimiento una vez que se obtiene un vector cuantificado se clasificará para saber a qué etiqueta corresponde, y se determine, de esta manera, a qué persona hace referencia. La distorsión acumulada puede llegar a ser esencial sobre toda una expresión de entrada en un sistema de CV y ser usada para hacer la determinación del reconocimiento [3], ver figura 2.

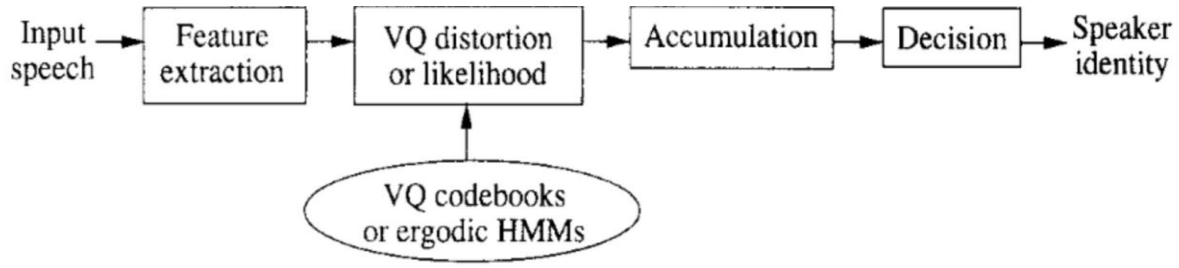

Figura 2. Modelo de un sistema de reconocimiento del hablante basado en Cuantización Vectorial [3].

Por otra parte, durante esta misma década de los 80 Poritz [9] propuso usar un MOM ergódico, esto quiero decir que todas las posibles transiciones entre los estados están permitidas con la finalidad de agilizar y robustecer las arquitecturas de los MOM tradicionales, véase figura 3.

Fue a inicios de la década de los 90 cuando Tishby [10] expandió la idea de Poritz’s usando un MOM autorregresivo ergódico de 8 estados representado por funciones de densidad de probabilidad continua con 2 a 8 componentes de mezclas guassianas por estado el cual tuvo un espectro más alto resolución que el modelo de Poritz. La diferencia que radica entre un MOM de izquierda a derecha como el de la figura 1 y en un MOM ergódico es que en un modelo ergódico cada estado del modelo puede ser alcanzado en un sólo paso desde cualquier otro estado del modelo. Posteriormente, Rose en [11] propuso utilizar un MOM de un solo estado, que ahora se llama Modelo de Mezcla Gaussiana (GMM, de Gaussian Mixture Model), como un modelo paramétrico robusto.

1.4 Década de los 90

La búsqueda por hacer más robustos los sistemas de reconocimiento del hablante se convertía en un tema de suma importancia. A inicios de esta década, Matsui en [12,28] compara los métodos basados en CV con los basados en MOM desde un punto de vista de robustez contra las variaciones de los enunciados y encontró que los métodos de MOM continuos ergódicos son más rápidos que los discretos o los MOM de izquierda a derecha y que los modelos continuos son tan robustos como los basados en CV cuando los datos de entrenamiento son suficientes y están disponibles.

Además, investigaron las tasas de identificación del hablante usando los MOM continuos como una función del número de estados y transiciones.

Por otro lado demostró que las tasas de reconocimiento del hablante estaban fuertemente correlacionadas con el número total de mezclas, independientemente del número de estados. Esto significa que el uso de información sobre las transiciones entre diferentes estados no es efectivo para el reconocimiento del hablante independiente del texto y, por lo tanto, GMM logra casi el mismo rendimiento que el MOM ergódico de múltiples estados

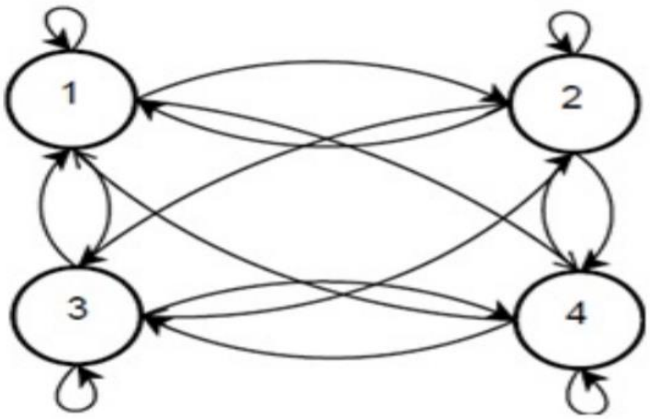

Figura 3. Modelo Oculto de Márkov ergódico [4]

En 1993 Matsui propone un método llamado Método de texto solicitado en que las oraciones claves cambian por completo cada vez que se usa el sistema [7]. El sistema acepta a la persona registrada solo si esta persona pronunciaba a la oración dada por el sistema. Este método presentó mejoras significativas en el reconocimiento del hablante las cuales consistían en que un impostor no puede saber de antemano qué expresión se le pedirá decir. Por lo tanto, si el impostor llegaba con una grabación de alguna persona registrada que dijera alguna palabra clave u oración, no iba a funcionar contra este método y, debido a que el vocabulario es ilimitado los impostores no podían saber de antemano la oración o palabra que usarían.

También el sistema puede rechazar una expresión cuyo texto no es idéntico al texto solicitado incluso aun si este es emitido por una persona que está registrada. De esta manera una voz grabada y reproducida podía ser rechazada.

Para esta década también se implementó técnicas de Normalización de Puntuación para reducir las probabilidades que el sistema arrojaba, ya sean probabilidades más cercanas a cero o más cercanas a uno. Esto ayudó mucho a los sistemas de reconocimiento debido a que se presentaron muchas variaciones de los valores de probabilidad. Esto era debido por las variaciones que una misma persona presentaba en su voz, ya que es casi imposible que una persona reprodujera una palabra u oración exactamente igual, por lo que también se crearon técnicas de probabilidad a posteriori (*likelihood rate*) [8] para reducir el costo computacional, y hacer predicciones más precisas sobre quién puede ser la persona que habla.

1.5 Década del 2000

Para la década 2000 nuevas técnicas de normalización fueron implementadas: *Znorm, Hnorm, Tnorm, Htnorm, Cnorm* y *Dnorm*, [28] con el fin de mejorar el desempeño de un sistema cuando se va a entrenar y se empieza además a realizar pruebas con diferentes micrófonos tanto para la fase de entrenamiento como para la fase prueba.

Dentro de lo sucedido durante esta década para el área de procesamiento de voz se mencionará solamente a GMM, ya que esté método será la base para el desarrollo del objetivo de esta tesis.

Reynolds, fue quien desarrolló un sistema de reconocimiento del hablante que actualmente es de lo más utilizados e implementados para hacer esta tarea y dicho sistema hace uso de la variación de modelos de mezclas Gaussianas. Agregándole dichas variaciones se le conoce al sistema como Gaussian Mixture Model-Universal Background Model (GMM-UBM) haciendo uso de la normalización Hnorm [13] que mejora el desempeño del sistema cuando se entrena y se realizan las pruebas con diferentes micrófonos.

Hablar de GMM-UBM es de suma importancia en esta tesis porque es el sistema base para el cual se va a implementar la técnica i-vectors y contra el cual se comparará en la etapa de experimentos. GMM es un sistema de verificación del hablante, cuyo objetivo es el de determinar la similitud que existe entre dos modelos dados. El primer modelo es el modelo de entrada, es decir, se da una oración o expresión y se crea su modelo que lo va a representar durante la fase de verificación. El segundo modelo es nuestro hablante que se toma como hipótesis y del cuál pensamos que es él mismo el que ha pronunciado el modelo de prueba. Una vez calculados estos dos modelos se comparará contra un umbral establecido para determinar si se trata o no de la misma persona. Si el resultado es mayor al umbral, se acepta, de lo contrario, se rechaza.

Para este sistema de GMM se utilizan matrices diagonales de covarianza en lugar de matrices completas porque se demostró en [13] que pueden alcanzar el mismo nivel de desempeño y son más eficientes computacionalmente.

1.6 Década del 2010

A inicios de esta década se crea una nueva técnica para los sistemas de reconocimiento del hablante, y que será además la técnica sobre la que se emplee en el desarrollo de esta tesis. Se trata del tema central, llamada *i-vector*, y fue propuesta por Dehak [22]. Esta técnica surgió debido a los problemas que se presentan al trabajar con grandes dimensiones de datos de entrenamiento. Además, debido a los problemas que ya no solo se presentan en la variabilidad de una persona sino también ahora en la variabilidad del canal, que es el medio en el que se graba la voz de una persona.

I-vector es una técnica que considera perteneciente al *state of the art (Estado del arte)*, ya que en ella se basan y se recopila todos los conocimeintos sobre este tema de reconocimiento del hablante para dar inicio a una nueva técnica que mejore a los sistemas anteriores y actuales. La manera en la que trabaja i-vectors y por la cual es que fue elegida a implementar en esta tesis es por las razones siguientes:

- Define un solo espacio que contiene la variabilidad de una persona y del medio en que se obtiene su voz, llamado Espacio de Variabilidad Total.

- Trabaja con la reducción de la dimensionalidad de los datos, esto genera grandes ventajas al utilizar muchos datos de entrenamiento.

- Define un tamaño de vector igual para todos sin importar la duración de cada expresión.

- La misma importancia que le da a expresiones largas se la da a las expresiones cortas

- Utiliza el enfoque de Método de Texto Independiente.

- Considerada una técnica más en el Estado del Arte.

- El utilizar esta técnica da pie a crear nuevos sistemas que mejoren al actual.

- Actualmente es una técnica poco desarrollada e implementada, en la UNAM.

En el capítulo 4 se habla a detalle de cómo trabaja i-vector, la manera en que procesa una expresión de voz y los algoritmos que utiliza para su implementación.

Capítulo 2. Reconocimiento del hablante como parte de un sistema biométrico

El objetivo de esta tesis es desarrollar un sistema de verificación del hablante usando la técnica de i-vectors con PLDA. Este sistema de verificación también se clasifica como un sistema de reconocimiento biométrico de voz. A continuación, se explicará qué es un sistema biométrico y por qué su implementación con voz trae consigo grandes beneficios, poniendo solo como ejemplo uno de ellos, su aplicación al área forense.

Un sistema de reconocimiento biométrico se basa en algunas características biológicas o de comportamiento de una persona que sea totalmente único e inigualable en otra persona, es decir, que sea univoco.

Las características biométricas empleadas deben tener las siguientes propiedades:

- Universalidad: todos los individuos las tienen
- Singularidad o univocidad: distinguen a cada individuo
- Permanencia en el tiempo y en distintas condiciones ambientales
- Medibles de forma cuantitativa

Y las tecnologías para medir estas características deben proporcionar:

- Rendimiento: nivel de exactitud

- Aceptación: por parte del usuario
- Resistencia al fraude y usurpación

Existen varios tipos de sistemas de reconocimiento biométricos, entre los más empleados, son:

- Huella dactilar
- Iris del ojo
- Simetría de la cara
- Firma electrónica
- Voz

En esta tesis se emplea la voz como característica única de cada persona para poder hacer el sistema de reconocimiento del hablante.

Un sistema biométrico consta de dos etapas las cuales son: identificación y verificación. La diferencia que radica entre estas dos etapas es que en la identificación el sistema puede arrojar una salida binaria para decir si una persona está o no está registrada en la base de datos, mientras que, en la verificación la intención del sistema no solo es saber si una persona existe o no dentro de la base de datos sino también determinar si la persona es quien dice ser, es decir, es la comparación de la muestra recogida del usuario frente a una base de datos de rasgos biométricos registrados previamente.

No se requiere de una identificación inicial por parte del usuario, es decir, el único dato que se recoge de él en ese momento es su voz como muestra biométrica. Este método requiere de un proceso de cálculo complejo, puesto que se ha de comparar esta muestra con cada una de las anteriormente almacenadas para buscar una coincidencia.

En la verificación, se toma como referencia el sistema de identificaron para comprobar que la personas es quien dice ser y se lleve a cabo la verificación de una manera más rápida, sin embargo, el primer paso del proceso es la identificación del usuario mediante algún nombre de usuario, tarjeta o algún otro método. De este modo se selecciona de la base de datos el patrón que anteriormente se haya registrado para dicho usuario. Es un proceso simple, al tener que comparar únicamente dos muestras, en el que el resultado es positivo o negativo [15] con esto entonces se le puede dar acceso a la información, la entrada a alguna área restringida, manipulación de información, etc. En esta tesis lo que se busca es que el sistema no tenga el paso de identificación previamente, sino que sin saber quién es la persona, se pueda determinar cuál es su modelo con el que i-vector lo está comparando

2.1 Comparativa de un sistema biométrico frente a sistemas de autenticación e identificación automática.

Las tecnologías biométricas surgen como alternativa o complemento a las técnicas de identificación y autenticación existentes. Por ello es posible establecer una comparación directa entre ambas, destacando beneficios que resultan del uso de biometría junto con aspectos en los que las técnicas tradicionales son superiores [16].

Se han de considerar los siguientes aspectos:

1. Necesidad de secreto: las contraseñas han de ocultarse y las tarjetas no deben de estar al alcance de terceros, mientras que la biometría no requiere de estas medidas de protección que son exclusivamente dependientes del usuario.

2. Posibilidad de robo: las tarjetas y contraseñas pueden ser robadas. Sin embargo, robar un rasgo biométrico es extremadamente complejo.

3. Posibilidad de pérdida: las contraseñas son fácilmente olvidables y las tarjetas se pueden perder. Los rasgos biométricos permanecen invariables salvo en contadas excepciones y siempre están con el sujeto a quien identifican.

4. Registro inicial y posibilidad de regeneración: la facilidad con la que se puede enviar una contraseña o tarjeta nueva contrasta con la complejidad que supone el registro en un sistema biométrico, ya que requiere de la presencia física del individuo en esta fase. Hay que añadir que los rasgos biométricos son por definición limitados, mientras que la generación de contraseñas es ilimitada, lo cual es una ventaja.

5. Proceso de comparación: la comparación de dos contraseñas es un proceso sencillo. Sin embargo, comparar dos rasgos biométricos requiere de mayor capacidad computacional.

6. Comodidad del usuario: el usuario ha de memorizar una o múltiples contraseñas y, en el caso de que use una tarjeta, ha de llevarse siempre consigo. Utilizando tecnología biométrica no se necesita realizar estos esfuerzos.

7. Vulnerabilidad ante el espionaje: una discreta vigilancia hacia nuestra actividad podría servir para obtener nuestra contraseña o robar nuestra tarjeta. Ese método no es válido ante los sistemas biométricos.

8. Vulnerabilidad a un ataque por fuerza bruta: las contraseñas tienen una longitud de varios caracteres que pueden ser encontrados mediante distintas combinaciones. Por su parte, una muestra biométrica emplea cientos de bytes, lo que complica mucho a los ataques por fuerza bruta.

9. Medidas de prevención: los ataques contra sistemas protegidos por contraseña o tarjeta se producen desde hace años, y las medidas de prevención contra ellos ya se encuentran muy sofisticadas. Por el contrario, los ataques a los sistemas biométricos son un área en la que estas medidas de prevención se están generando en estos momentos.

10. Autenticación de usuarios “reales”: la autenticación de usuarios mediante contraseña o tarjeta y su efectividad, dependen absolutamente de la voluntad del usuario a la hora de hacerlas personales e intransferibles. La biometría está altamente relacionada con el propio usuario pues no puede ser prestada ni compartida.

11. Costo de implantación: en el momento de la implantación, el hecho de instaurar un sistema de contraseñas tiene un costo bajo, mientras que en el caso de un sistema basado en muestras biométricas es más costoso.

12. Costo de mantenimiento: el costo de mantenimiento de un sistema biométrico, una vez está implantado con éxito, es menor al de un sistema de contraseña o tarjeta ya que no conlleva gastos de gestión asociados a la pérdida u olvido de credenciales.

En la tabla 1 se identifican los aspectos en los que destaca cada método de autenticación:

ASPECTO	BIOMETRÍA	CONTRASEÑAS/ TARJETAS
Necesidad de secreto	✔	
Posibilidad de robo (baja)	✔	
Posibilidad de pérdida (baja)	✔	
Registro inicial y posibilidad de regeneración		✔
Proceso de comparación (fácil)		✔
Comodidad del usuario	✔	

Vulnerabilidad ante el espionaje (baja)	✔	
Vulnerabilidad a un ataque de fuerza bruta (baja)	✔	
Medidas de prevención		✔
Autentificación de usuarios “reales”	✔	
Coste de implementación (bajo)		✔
Coste de mantenimiento (bajo)	✔	

Tabla 1. Comparativa entre sistemas de autenticación.

Sin embargo, como se mencionó en el capítulo 1, en los sistemas de reconocimiento de voz y del hablante se han manifestado a lo largo de los años una serie de problemas que dificultan el rendimiento del sistema, estos problemas pueden verse también como vulnerabilidades del reconocimiento o debilidades.

En las tablas 2 y 3 se muestran las vulnerabilidades en común que se encuentran en los sistemas biométricos. La tabla 3 es más específica en mostrar las vulnerabilidades de un sistema de reconocimiento de voz y del hablante [16]

Sistema Biométrico	Vulnerabilidades
Vulnerabilidades comunes en los sistemas biométricos	- Calidad baja de los dispositivos de captura de datos. - Ubicación inadecuada del dispositivo de captura. - Falta de conocimientos técnicos del personal. - Falta de recursos (tanto de personal como económicos). - Escasa concienciación en materia de seguridad. - Percepción de ausencia de privacidad por parte de los usuarios. - Desconocimiento de la calidad de productos a conseguir.

Tabla 2. Vulnerabilidades en un sistema biométrico de voz.

Sistema Biométrico	Vulnerabilidades
Reconocimiento de voz y del hablante	- Enfermedades de la voz: bronquitis, faringitis, gripe, laringitis, afonías, etc. - Variación entre el dispositivo de registro y el usado en la captura de muestras - Variación entre entornos de registro y captura de muestras (por ejemplo: interior y exterior) - Volumen del habla

Tabla 3. Descripción de las vulnerabilidades en un sistema biométrico de voz.

La técnica i-vectors junto con PLDA ayuda a resolver estas vulnerabilidades que tiene la mayoría de los sistemas reconocimiento del hablante, utilizando para ello el Corpus Valquiria, el cual cuenta con grabaciones de larga duración para modelar sus variaciones intra-hablante, esto también resuelve el volumen del habla, y siendo un corpus generado con expresiones de voz de distintos canales, se puede resolver el problema de variación del dispositivo junto con la de variación del entorno del registro debido a que se graba en las variaciones externas que viene siendo el ambiente y ésta última es clave de usar porque i-vectors puede modelar también la variación total del canal.

Un sistema biométrico se puede emplear en una gran cantidad de aplicaciones. En esta tesis se pretende también mostrar como la ayuda de un sistema de reconocimiento del hablante podría beneficiar en ámbitos forenses, y entiéndase por forense el área de juicios legales para la determinación de si una persona es o no culpable de lo que se le acusa, utilizando para ello la voz como evidencia que es dejada en un caso delictual, esto puede ser posible gracias a los avances tecnológicos y los algoritmos de hoy en día para hacer sistemas más sofisticados y robustos en su implementación, sin embargo, como se explicó en el capítulo anterior, uno de los grandes problemas que existen en el reconocimiento de voz es la intravariabilidad que existe en las personas. Es un gran problema porque fácilmente una persona puede modificar su forma de hablar intencionalmente para así hacer que el sistema falle y por ende se pueda tomar la decisión errónea, aunque se han desarrollado técnicas que nos permiten resolver este problema, tales como los Modelos de Mezclas Gaussianas para incrementar la discriminación de la persona mediante el uso de una función de densidad de probabilidad y con la ayuda de clasificadores que sean adecuados como Redes Neuronales Artificiales (RNA), Cuantización Vectorial (CV), Maquinas de Vectores de Soporte (MSV), etc., se han mostrado buenos resultados ante esta dificultad de la intravariabilidad. Generalmente, estos sistemas terminan la etapa de prueba con un umbral de decisión sobre la discriminación de la clase a la cual puede pertenecer la persona. Esto es adecuado para aplicaciones de control de

acceso, pero no para la inferencia forense, ya que el resultado no puede estar determinado por un umbral de identificación, más bien por una probabilidad de reconocimiento.

En ese sentido, se necesitará una mayor robustez del sistema para tener mejor confiabilidad de los que da como resultado el sistema, para esto se pretende utilizar un método junto a i-vectors llamado *Probabilistic Linear Discriminant Analysis* que nos permite migrar de un sistema de discriminación a uno de aproximación probabilística, que es más susceptible para ser usado como evidencia en una corte.

La manera en que opera un sistema de reconocimiento biométrico es la siguiente, dada una entrada, que viene siendo toda una serie de atributos de una persona desconocida la tarea es identificarla dentro de alguna de las clases existentes que son conocidas por el reconocedor.

Para lograr esto el sistema consta de dos etapas, la primera etapa es llamada entrenamiento, la cual se encarga de caracterizar las distintas clases con sus peculiaridades. La segunda etapa es la verificación, que es el proceso donde se evalúan características desconocidas frente a este universo entrenado donde durante el entrenamiento se le dan al sistema todas las grabaciones que se tiene sobre cada una de las personas de las cuales queremos que el sistema aprenda. Mediante un algoritmo de entrenamiento será que se obtenga un modelo para caracterizar a cada una de las personas. Una vez obtenido el modelo es almacenado para ser utilizado en la etapa de prueba. Se desea que estas grabaciones dadas tengan una representación valiosa de cada persona para hacer más robusto el sistema, es decir, grabaciones con voz clara, sin ruido, que se mencionen una gran cantidad de palabras, variabilidades que tenga la persona al hablar para obtener un buen modelo de cada persona, para que en la etapa de prueba cuando se verifique una señal de voz se pueda entonces determinar una decisión más certera. Para poder hacer el procesamiento de las señales de entrenamiento y poder crear los modelos se debe pasar por una serie de bloques en el sistema donde cada bloque se encarga de una tarea en específico. Se hablará más a detalle de esto en el siguiente capítulo.

Finalmente, el reconocedor del hablante puede ser dependiente o independiente de texto, pero, como se mencionó en el capítulo anterior, un reconocedor dependiente de texto tiene sus limitaciones al tener un espacio contable de palabra a usar tanto para el entrenamiento como en las pruebas, mientras que un sistema independiente de texto la persona no tiene un límite de palabras a usar tanto para su entrenamiento como para la etapa de prueba lo que hace que el sistema sea tanto más robusto como más real.

Capítulo 3. Gaussian Mixture Models

Los sistemas de verificación del hablante basados en modelos de mezclas gaussianas (GMM de Models Mixtures Gaussians) son actualmente una herramienta dominante para aplicaciones de identificación y verificación del hablante de texto independiente. En particular es común que estos sistemas basados en GMM utilicen una adaptación bayesiana en los modelos de los hablantes utilizando un modelo de fondo universal llamado UBM (de Universal Background Model) con el fin de saber si la persona que se va a evaluar comparte características fonéticas con alguna otra persona previamente entrenada. Esto hace que el sistema sea más robusto y más apegado a la vida real porque muchas personas llegan a tener similitudes en voz. Estos sistemas son conocidos como sistemas de detección/verificación del hablante GMM-UBM, del cual es necesario hablar y profundizar en su metodología porque será la base para crear los i-vectors del sistema de verificación de esta tesis

3.1 Procesamiento de la voz

Para poder hacer un sistema de reconocimiento es necesario que todos nuestros conjuntos de grabaciones a utilizar, sean sometidos a una serie de pasos antes de entrar a la etapa de modelado, ver figura 4. Es necesario que esto sea primero porque no sabemos en qué condiciones están grabados, y no podemos asegurar que todo el conjunto que será utilizado para entrenamiento está en condiciones adecuadas, es decir, sin ruido de fondo, claridad en la voz, micrófonos de alta calidad, palabras entendibles, etc., por lo que se realizan un conjunto de pasos por el cual cada expresión de voz tanto de entrenamiento como de prueba debe ser procesado antes de crear los modelos de cada persona.

Dentro del bloque de procesamiento ocurren varias etapas para el análisis de la señal:

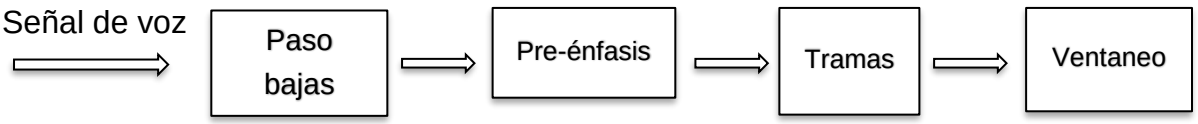

Figura 4. Análisis de una señal de voz.

Cuando entrenamos y reconocemos voz con un conjunto de diferentes micrófonos o canales el hacer entonces este paso de Procesamiento es un paso crucial para tener un reconocimiento preciso.

En el primer bloque la señal original pasa primero por un filtro paso bajas donde lo que se busca es reducir las frecuencias altas producidas por el ruido ambiental o bien por un mal equipo de grabación, dando como salida una señal de voz más clara.

Posteriormente la señal de voz pasa por un segundo filtro llamado pre-énfasis donde se busca ahora es resaltar las frecuencias altas de la señal entregada del filtro pasa bajas. Es lógico pensar que durante el filtro anterior algunos componentes de la voz también fueron eliminados. Con el filtro pre-énfasis compensamos este evento. El cálculo de dicho filtro se realiza mediante la ecuación:

$$y(n) = x(n) - ax(n-1) \qquad (1)$$

Donde a es una constante entre los valores 0.9 y 1.0.

Es difícil trabajar con toda la señal completa y querer obtener un buen modelo, debido a la gran cantidad de datos que posee, por lo que lo más correcto y óptimo es trabajarla por segmentos o tramas de N muestras de toda la señal, de esta manera

será más fácil trabajar con toda la señal de voz y hacer su análisis trama por trama individualmente, por lo que el siguiente paso es segmentar nuestra señal. Un punto clave de realizar este paso es que, al segmentar la señal, se pueden presentar discontinuidades entre los inicios y finales de las tramas de la señal original al procesarlas, una forma de solucionar este problema es hacer uso del traslape.

El uso de traslape es hacer que cada trama contenga un segmento de la trama anterior, típicamente se utiliza 30% o 50% de la trama anterior. Lo que se logra es evitar pérdidas en las tramas, ya que al haber segmentos con partes de la anterior se puede mantener la señal original sumando los traslapes entre las divisiones de las tramas, como lo muestra la figura 5.

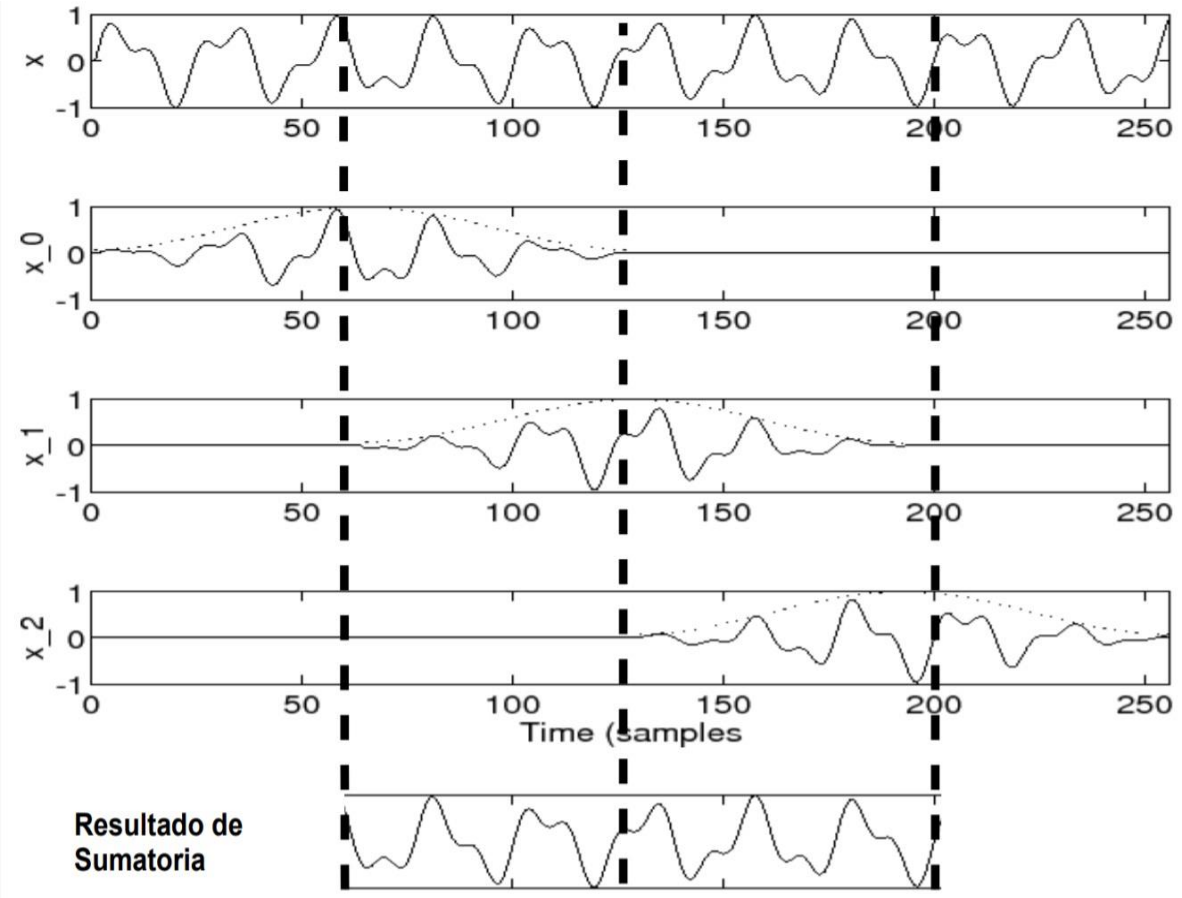

Figura 5. Resultado de proceso de segmentar una señal de voz, haciendo uso del traslape y multiplicándola por una ventana [18].

El último paso a realizar dentro del bloque de procesamiento antes de empezar con la extracción de las características es aplicar una ventana a las tramas resultantes. Debido a que la voz no es una señal periódica en el tiempo, puede entonces presentar dificultades para su análisis. Además, esto se hace porque las ventanas son funciones matemáticas usadas con frecuencia en el análisis y el procesamiento de señales para evitar las discontinuidades al principio y al final de cada trama como se ha

mencionado. Lo que se hace entonces es que cada trama traslapada será multiplicada por una ventana llamada Hamming.

Existen varios tipos de ventanas que se trabajan con señales digitales, pero la ventana más utilizada para voz comúnmente, y que será utilizada en esta tesis es la de Hamming, ya que esta ventana nos permite eliminar discontinuidades, lo que nos da como resultado una señal periódica y sin distorsiones para sistemas de reconocimiento de voz y de hablante, además el hacer este paso nos beneficiará más adelante para tener un mejor análisis espectral de la voz por medio de la Transformada Rápida de Fourier. La ventana Hamming se define como:

$$w(n) = 0.54 - 0.46\,cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \leq n \leq N-1 \qquad (2)$$

Después de haber realizado todos los pasos anteriores se puede proceder a la extracción de características de la señal de voz, la cual es fundamental en el desarrollo; y utilizando una técnica adecuada, puede generar mejores resultados. Siguiendo los pasos mostrados en la figura 6 en donde para cada ventana obtenida de la etapa anterior se le realiza una extracción de características que representen a cada hablante de manera particular.

Los coeficientes cepstral en frecuencia Mel o MFCC (de Mel-Frequency Spaced Cepstral Coefficients) es una manera de caracterizar la voz basados en la percepción auditiva humana. Este método ha sido ampliamente utilizado las últimas dos décadas para reconocimiento de voz y verificación de locutor [21].

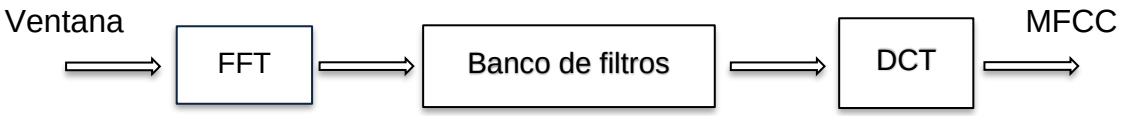

Figura 6. Extracción de características de la señal de voz.

Los pasos a realizar se describirán a continuación:

1.- Transformada rápida de Fourier: Después de recibir las ventanas del audio se procede aplicar la transformada rápida de Fourier (FTT) para obtener un espectro de magnitud y un espectro de fase de la ventana. Pero la que nos interesa es el espectro de magnitud porque es el que contiene la envolvente de la señal y que contiene, a su vez, información sobre las propiedades de resonancia del tracto vocal y se ha establecido que ésta es la parte más informativa del espectro para la tarea de verificación del hablante [22].

2.- Banco de filtros: La información de dicha envolvente del espectro de magnitud pasa por el un banco de filtros triangulares, el objetivo de este filtrado es aproximar la

resolución espectral a la respuesta del oído humano usando la frecuencia mel. Ver figura 7.

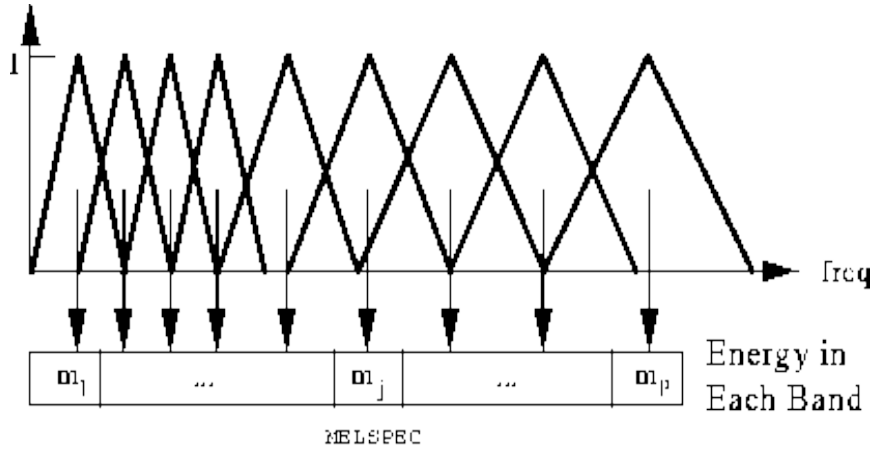

Figura 7. Banco de filtros esparcidos por la frecuencia Mel [19].

El número de filtros depende de la implementación, pero debe ser menor a la cantidad de elementos de la FFT.

Los centros de los filtros triangulares están espaciados de acuerdo a la escala de frecuencias mel, definida como:

$$f_{MEL} = 2595\, log\, (1 + \frac{f_{LIN}}{700}) \qquad (3)$$

donde f_{LIN} es la frecuencia en escala lineal. Las salidas del banco de M filtros se denota como Sm con m=1, ..., M.

3.- Transformada de coseno: Como último paso se transforman la salida del banco de filtros Sm con la transformada de coseno discreta o DCT (por Discrete Cosine Transform), con el objetivo de comprimir la información en pocos coeficientes, dando como salida los coeficientes cepstrales, que están dados por:

$$c_n = \sum_{m=1}^{M} log\, (S_m) cos[\frac{\Pi n}{M}(m - \frac{1}{2})] \qquad (4)$$

donde n es el índice del coeficiente cepstral. El vector final de coeficientes cepstral se obtiene manteniendo sólo los primeros 10 a 15 coeficientes de la DCT.

3.2 Likelihood Ratio

Recordando del capítulo 2 que un sistema de reconocimiento del hablante busca verificar que una persona es quien dice ser, pero para esto es necesario comparar su modelo de voz contra el resto de los modelos conocidos para tomar la decisión de que no es el modelo de otra persona.

Para lograr esta tarea se calcula una razón de probabilidad para tomar dicha decisión, conocida como Likelihood Ratio (LR), la cual funciona de la siguiente manera.

Dado un segmento de voz (Y) y una hipótesis del hablante (S), la tarea de detectar al hablante también conocida como verificación es determinar si (Y) fue dicha por el hablante (S). Una suposición implícita usada a menudo es que (Y) contiene la voz de solo una persona; si no se cuenta con información previa de que (Y) contiene a solo una persona entonces la tarea se convierte en detectar a múltiples hablantes o bien se necesitará emplear técnicas de separación de fuentes para encontrar a nuestra persona de interés, ver figura 8.

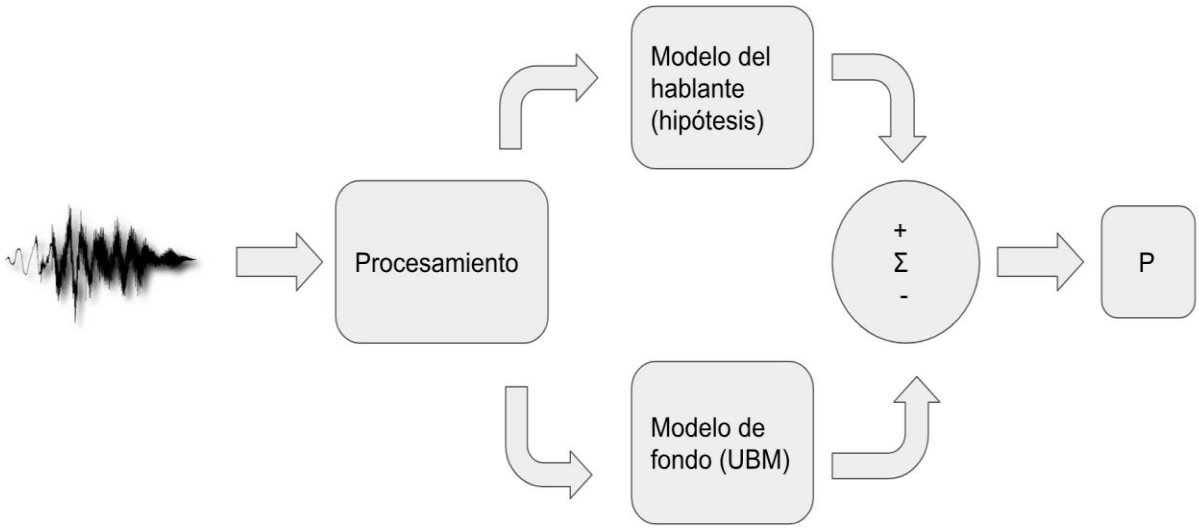

Figura 8. Sistema de reconocimiento del hablante usando Likelihood Ratio.

La idea de cómo verifica Likelihood a una persona por su voz junto con un modelo de hipótesis puede ser representado de la siguiente manera:

H_0: (Y) viene de la hipótesis del hablante (S)

Y

H_1: (Y) NO viene de la hipótesis del hablante (S)

para tomar obtener un resultado óptimo de esta prueba para decidir entre estas 2 hipótesis se necesita de probabilidad (verosimilitud) dado por:

$\frac{P(Y|H0)}{P(Y|H1)}$ >= θ se acepta, caso contrario se rechaza

Donde $p(Y|H_i)$ con $i = 0\, y\, 1$ es la función de densidad de probabilidad para las hipótesis Hi evaluadas para el segmento de voz observado (Y) también conocido como la probabilidad de la hipótesis H_i dado un segmento de voz. El objetivo básico de un sistema de verificación del hablante es determinar técnicas para calcular los valores para las 2 probabilidades $p(Y|H_0)$ y $p(Y|H_1)$[13].

La importancia de tener una etapa de procesamiento es porque permite extraer de la señal de voz las características que nos den información representativa de la persona; en dicha etapa se pueden aplicar técnicas para reducir algunos efectos que nos permiten trabajar de manera más limpia la señal de voz y obtener una mejor extracción de características, como filtros para el ruido. La salida de esta etapa es típicamente una secuencia de vectores característicos representantes del segmento de prueba $X = \{x_1, \ldots, x_T\}$ donde x_T es un vector característico en tiempo discreto $t \in \{1,2, \ldots, T\}$, esta secuencia de vectores característicos son los que se usan para calcular las probabilidades H_0 y H_1

Ahora bien, matemáticamente H_0 es representada por λ_h y caracteriza la hipótesis del hablante (S) en el espacio de características de x. Uno puede asumir que una distribución gausiana representa mejor la distribución de vectores para H_0 así que λ_h estaría denotado por el vector de medias y los parámetros de la matriz de covarianza de la distribución gaussiana. H_1 es representada por el modelo $\lambda_{\overline{h}}$ por lo que la relación de estadísticas quedaría $P(X|\lambda_h)/P(X|\lambda_{\overline{h}})$. A menudo se usa el logaritmo para esta relación y obtener finalmente la razón log-probabilística.

$$P(X) = log\, P(X|\lambda_h) - log\, P(X|\lambda_{\overline{h}}) \qquad (5)$$

3.3 Gaussian Mixture Models

La implementación de GMM se debe a la importante necesidad de seleccionar una función de probabilidad $p(X|\lambda)$ que nos garantice el mejor desempeño del sistema y la elección de esta función es bastante dependiente de las características que son usadas y de las especificaciones de la aplicación.

Para reconocedores de texto independiente, donde no hay información previa de lo que la persona va a decir, la función de probabilidad más exitosa ha sido GMM.

La función de densidad de probabilidad para un vector de características (X) de D dimensione se define como:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i \, p_i(X) \qquad (6)$$

A esta función de densidad de probabilidad se le llama también como una combinación lineal de sumas ponderadas de M densidades gaussianas, $p_i(x)$, cada una de ellas parametrizada por un vector de medias μi de dimensión $Dx1$ y una matriz de covarianza Σi de dimensión DxD.

$$p_i(x) = \frac{1}{(2\pi)^{1/2}|\Sigma_i|^{1/2}} exp\{-\frac{1}{2}(x-\mu_i)'(\Sigma_i)^{-1}(x-\mu_i)\} \qquad (7)$$

Los pesos mixtos w_i deben satisfacer que su suma de sus elementos sea igual a uno $\sum_{i=1}^{M} w(i) = 1$.
Por lo tanto, los parámetros del modelo de densidad de un hablante estarán representados como $\lambda = \{w_i, \mu_i, \Sigma_i\}$ donde $i = 1, \ldots, M$.

Ahora bien, por lo general no se usa toda la matriz de covarianza sino solo su diagonal esto se hace por 3 razones principales:

1.- El modelar la densidad de orden M_n de toda la covarianza GMM puede ser igualada usando un orden grande de la diagonal de la covarianza GMM.

2.- La matriz diagonal de todos los modelos de GMMs son más eficientes computacionalmente que toda la covarianza GMMs para su entrenamiento, y además el repetir matrices inversas DxD ya no es requerido.

3.- Empíricamente se observa que la diagonal de la matriz GMM supera y nos representa mejor los GMMs que todas las matrices GMMs.

La manera de entrenar a un modelo GMM y de obtener estos parámetros que nos representan o identifican a cada persona es por un algoritmo de iteración llamado Expectation Maximization (EM), este algoritmo mejora los parámetros de GMM en cada iteración, con la restricción de que en cada iteración la probabilidad actual debe ser mayor que la probabilidad anterior.

De igual forma los parámetros para el UBM son entrenados usando también el algoritmo EM, pero por lo general a este modelo se le agrega una adaptación bayesiana para darle otro tipo de entrenamiento a los modelos de hablantes, es decir, con cada modelo se analiza si comparte características fonéticas con respecto al modelo UBM de los hablantes de fondo.

Usualmente los vectores característicos de X se asumen independientes, de tal forma que la probabilidad logarítmica de un modelo λ para una secuencia de vectores característicos $X = \{x_1, x_1, \ldots, x_T\}$ se calcula como:

$$log\ p(X|\lambda) = \sum_{i=1}^{T} log\ p(x_t|\lambda) \qquad (8)$$

donde $p(x_t|\lambda)$ se calcula igual como en la ecuación 6.

Las ventajas de usar GMM como una función de probabilidad son:

- Es computacionalmente barato.
- Se basa en un modelo estadístico bien entendido
- Para tareas de texto independiente, es insensible a los aspectos temporales de la voz, modelando solamente la distribución fundamental de observaciones acústicas de un hablante.

Esto último podría verse como una desventaja porque los niveles más altos de una señal de voz también poseen información sobre lo que el hablante a transmitido pero no se utiliza, o no se usa esa información; porque el extraer esos niveles altos de la señal de voz podría beneficiar más a un sistema de reconocimiento de voz para identificar palabras dependientes o palabras claves, así que solo se utilizan como promedios para cálculos de valores de probabilidad sin usar los niveles altos de información.

3.4 Universal Background Model

La implementación de este método se da debido a que mientras un modelo H_0 está bien definido y puede ser estimado usando entrenamiento de expresiones de voz de la persona (S). El modelo para $\lambda_{\overline{h}}$ es menos bueno, puesto que potencialmente debe representar un espacio entero de posibles alternativas para la hipótesis del hablante. Existen 2 enfoques para tener estos modelos de hipótesis alternativas [13], la primera es usar un conjunto de otros modelos de hablantes para cubrir el espacio de hipótesis alternativas, llamados cohortes, hablantes de fondo. Dado un conjunto de N modelos de hablantes de fondo el modelo de hipótesis alternativo es representado por:

$$p(X|\lambda_{\overline{h}}) = F(p(X|\lambda_1), \ldots, p(X|\lambda_N)) \qquad (9)$$

donde F es una función que actúa como el promedio o el máximo de los valores de probabilidad del conjunto de hablantes de fondo. Se ha encontrado que se puede obtener un mal rendimiento con este enfoque, ya que se requiere usar varios conjuntos de hablantes de fondo de hablantes específicos y esto puede ser un problema en la aplicaciones debido a que se requiere usar un gran número de hipótesis de hablantes y cada uno requiere su propio modelo y esto se vuelve más complicado de realizar porque a nuestra muestra de voz no solo la comparamos contra el modelo hipótesis sino que también habría que compararla contra todos los modelos existentes y si estos van aumentando en número puede hacer que el sistema ya no sea tan óptimo, se vuelve caro computacionalmente. Además de que entre más modelos de hablantes tengamos más memoria ocupamos lo cual también no beneficia mucho hacerlo con este primer enfoque.

El segundo enfoque es mejor, este explica que se puede agrupar voz de varios hablantes y entrenar un solo modelo llamado Universal Background Model (UBM). Dado una colección de muestras de voz de un largo número de hablantes representativos de una población durante el reconocimiento se puede tener un solo modelo, que es entrenado para representar la hipótesis alternativa, la principal ventaja de este enfoque es que un solo modelo del hablante independiente puede ser entrenado una solo vez para una tarea en particular y luego ser usado para todas las hipótesis de hablantes en esa misma tarea.

El UBM puede ser visto como un gran GMM [4] ya que sigue la misma estructura para ser modelado y de igual forma se entrena por el algoritmo EM. Específicamente, queremos seleccionar la voz que refleja a una persona alternativa, para ser encontrada durante el reconocimiento, de ahí que un sistema se llame GMM-UBM.

A continuación, se mencionan algunos puntos claves e importantes por el cual es más óptimo aplicar UBM:

- No hay una medida objetiva para determinar el número correcto de hablantes o la cantidad de voz para usar en el entrenamiento de UBM, se ha observado que no se pierde el rendimiento del entrenamiento de UBM usando una hora de voz, comparado con un entrenamiento donde se usan 6 horas de voz

- El UBM se entrena por el algoritmo EM.

- Se debe de tener cuidado de que el juntar los datos para entrenar estos deben estar balanceados, de lo contrario el modelo final estará sesgado hacia la subpoblación dominante (más masculinos que femeninos o viceversa).

- Otro enfoque es entrenar UBM's de forma individual, uno para masculino y uno para femenino y después juntar los modelos de subpoblación en uno solo, esto tiene una gran ventaja, debido a que uno puede usar los datos desbalanceados y no importaría porque se puede compensar con los datos del otro modelo y tener el control del modelo final UBM. En la figura 9 se muestra una descripción gráfica de esto.

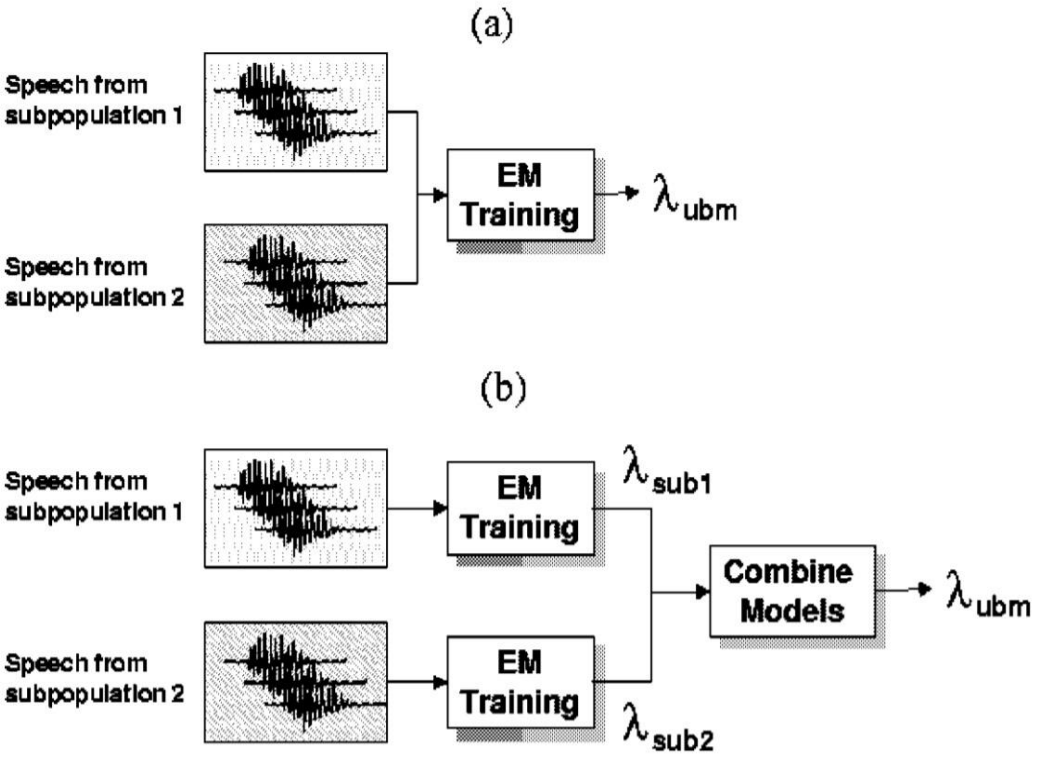

Figura 9. En el índice (a) los datos de las subpoblaciones primero se agrupan para obtener el modelo UBM por medio de EM, en el índice (b) los modelos agrupados primero se entrenan de forma individual y después se combinan para obtener el modelo UBM final [13].

Para la obtención de los parámetros de los modelos GMM y UBM se utiliza igualmente el algoritmo EM, el cual parte de un modelo aleatorio λ_0, llamado también modelo inicial y con cada iteración el algoritmo mejora sus parámetros, de manera que el likelihood del modelo con los datos aumenta, es decir, para las iteración k y $k + 1$, $P(X_{Bkg}|\lambda_{k+1}) > p(X_{Bkg}|\lambda_k)$. Posteriormente el nuevo modelo se vuelve el inicial para la siguiente iteración y el algoritmo se detendrá hasta que haya una convergencia en las probabilidades o se alcance el límite en el número de iteraciones especificadas. En cada una de las iteraciones del algoritmo se llevan a cabo dos pasos, el primero es el E y después se realiza el M.

Siguiendo el entrenamiento de [20] en el paso E se calcula $\gamma_i(x_j)$, que corresponde a la probabilidad posterior o a posteriori de que el vector x_j siga la distribución de la gaussiana i, y por el teorema de Bayes, se calcula como la probabilidad a priori multiplicada por la función de likelihood de la gaussiana i, y normalizada:

$$\gamma_i(x_j) = \frac{w_i \ * \ p(x_j)}{\sum_{k=1}^{M} \ w_k \ * \ p_k(x_j)} \qquad (10)$$

En el paso M se hace uso de los pesos obtenidos en el paso E para obtener los nuevos valores de los parámetros necesarios para el modelo, las ecuaciones de actualización son las siguientes:

Para los pesos de las componentes

$$\underline{w_i} = \frac{1}{N} \sum_{j=1}^{N} \gamma_i(x_j) \qquad (11)$$

Para las medias

$$\underline{\vec{\mu}_i} = \frac{\sum_{j=1}^{N} \gamma_i(x_j) \, x_j}{\sum_{j=1}^{N} \gamma_i(x_j)} \qquad (12)$$

Las matrices de covarianza

$$\underline{\Sigma_i} = \frac{\sum_{j=1}^{N} \gamma_i(x_j) \left(x_j - \underline{\mu_i}\right)\left(x_j - \underline{\mu_i}\right)'}{\sum_{j=1}^{N} \gamma_i(x_j)} \qquad (13)$$

Capítulo 4. Sistema I-vectors

I-vectors es el método a aplicar en esta tesis para lograr el desarrollo de un sistema de verificación del hablante. Este método se conoce también como *front end factor analysis* y fue propuesta primero por Dehak [21,22] para proporcionar una representación de varios hablantes en varios canales entre el vector GMM de alta dimensión y las representaciones de características de MFCC tradicionales de baja dimensión.

La extracción de los tamaños que tendrían los vectores o i-vectors fue motivado por la existencia de los vectores basados en *Joint Factor Analisys* (JFA) [22] lo cuales eran de gran dimensión. El enfoque de JFA modela a el hablante y la variabilidad del canal por separado y de igual manera los trabaja por separado, mientras que el método de i-vectors está hecho para modelar un solo espacio de baja dimensión que contiene y cubre a toda la variabilidad total. Se ha demostrado que parte de la información discriminante del hablante puede perderse en el espacio del canal en el enfoque de JFA [22].

Como los i-vectors están basados en un solo espacio de variabilidad que contiene tanto la información de hablante como la variabilidad de canal, se requiere de técnicas de compensación para limitar los efectos de la alta variabilidad presentada en el canal de los i-vectors que representan al hablante. La compensación del canal juega un rol importante en los sistemas de verificación del hablante. Cuando el método de los i-vectors fue introducido, técnicas de compensación del canal también se incluyeron, tales como Within-class Covariance Normalization (WCCN), Linear Discriminant Analysis (LDA), Nuisance Attribute Projection (NAP) y Scatter Difference NAP (SD-NAP) [21,22], estas fueron usadas para la compensación de la variación del canal en el i-vector para sistemas de verificación del hablante. Poco después Kenny al observar que cada grabación puede ser representada por un i-vector de baja dimensión, introdujo PLDA [26] para modelar la variabilidad del canal dentro del espacio de i-vectors.

4.1 I-vectors

Este método se ha convertido en el estado del arte en los sistemas de verificación del hablante, debido a que aplica la reducción de la gran dimensión de los datos de entrada reteniendo la información más relevante. Los i-vectors es una técnica que define un solo espacio que contiene tanto la variabilidad del hablante como la variabilidad de la sesión o del canal, llamado espacio de variabilidad total, esto permite que al trabajar se pueda darle un mejor trato a los i-vectors para extraer la información más importante de una persona en su totalidad, ésta es una gran ventaja frente a sistemas basados en JFA el cual trabaja por separado ambos espacios, tratándolos de forma independiente. Ahora bien, la necesidad de crear un solo espacio, es que se ha demostrado que existe información de la voz de una persona en el espacio de variabilidad del canal [22] por esto es que surge el método de i-vectors que junta ambos espacios para tener la mayor información sobre los hablantes.

Dejando más claro, lo qué es un i-vector, se puede decir que, un i-vector es una proyección en un espacio de variabilidad total generado por dos fuentes de variabilidad: el canal y el hablante. El método para modelar un hablante usando i-vectors se define de la siguiente manera. Dado un conjunto de grabaciones X_s de entrada el nuevo modelo dependiente del hablante y del canal es definido como:

$$M_s = m + Tw \qquad (14)$$

Donde M es el modelo independiente del hablante y el canal, m es un vector de medias de los hablantes de fondo, el cual puede ser tomado del modelo UBM, de ahí de la necesidad de explicar el método de GMM-UBM, T es una matriz rectangular la cual contiene tanto las variabilidad del hablante como del canal, y w es un vector aleatorio que tiene una distribución normal estándar $N(0,1)$, los componentes del vector w son los factores totales. los cuales son llamado como vectores de identidad (i-vector). Ver figura 10.

En este modelo, M se asume para ser normalmente distribuido con un vector de medias (m) y una matriz de covarianza Σ, como si fuera en GMM.

La manera para entrenar a la matriz T es exactamente la misma a como se entrena la matriz V en el método JFA [2], excepto por una diferencia muy importante, en el entrenamiento de la matriz V la cual contiene los eigenvoices, todas la grabaciones de una persona son consideradas para ser la misma persona, sin embargo aquí en i-vectors para la matriz de variabilidad total, se considera que todo el conjunto de grabaciones de un hablante dado, han sido producidas por diferentes hablantes, de

aquí que este método se conoce también como un método de análisis factorial porque a cada grabación de voz nos permite proyectarla sobre el espacio de variabilidad total de baja dimensión.

El factor total *w* es una variable oculta, la cual puede ser definida por una distribución posterior condicionada por las estadísticas de Baum Welch para una expresión dada. Esta distribución posterior es una distribución gaussiana y el promedio de esta distribución corresponde exactamente para nosotros los i-vector. De manera similar para las estadísticas de Baum Welch son extraídas usando el UBM [23]

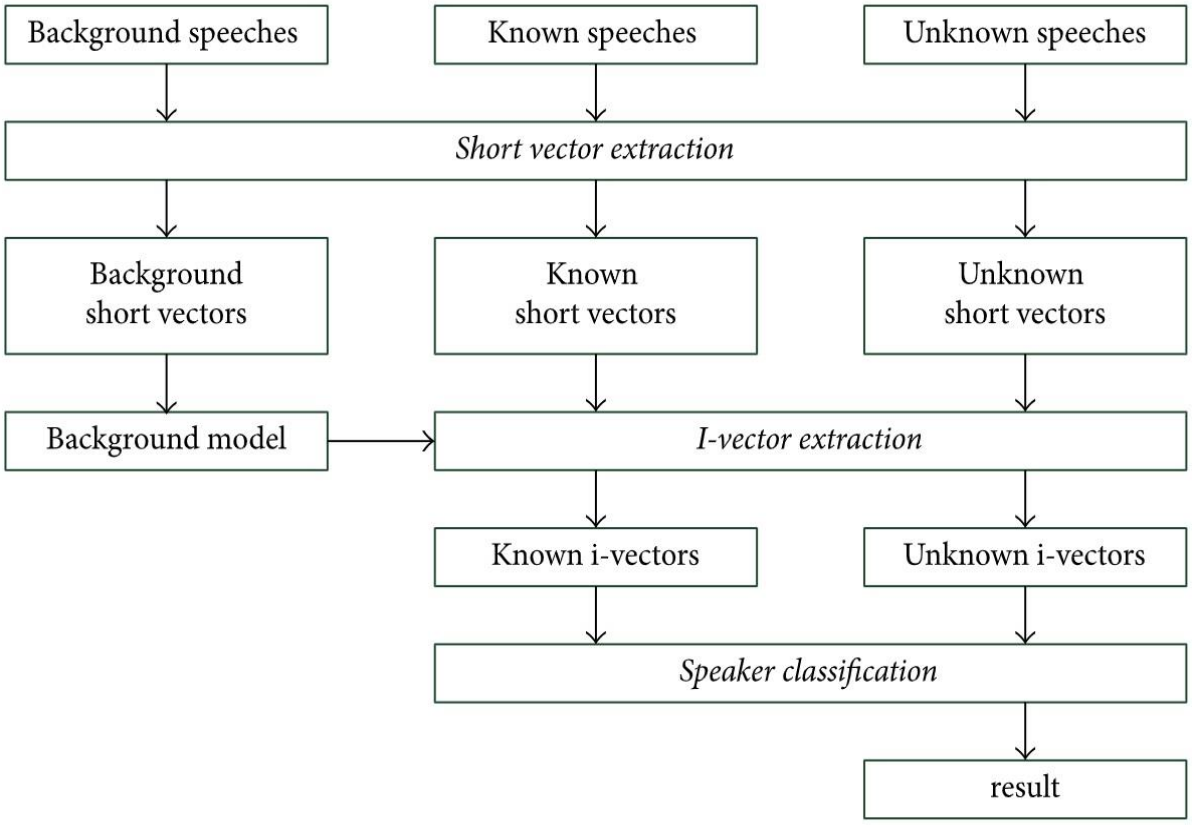

Figura 10. Sistema de verificación del hablante basado en i-vector [24].

En la figura 11 se muestra un ejemplo de la obtención de los primeros 5 i-vectors de entre 139 hablantes, cada i-vector contiene 50 características que son las que representan al hablante dentro de todo el espacio de variabilidad, se pueden representar dentro de una matriz de i-vectors donde el número de filas representa el tamaño del i-vector y el número de columnas es igual al número de personas con las que el sistema será entrenado.

50x139 double

	1	2	3	4	5
1	0.0011	-4.3983e-06	1.0008e-06	-3.3730e-04	2.5593e-04
2	-9.2703e-04	-1.2820e-04	6.7361e-04	-5.0812e-06	3.0758e-04
3	-6.9475e-04	2.5052e-05	-9.3283e-05	3.5324e-05	7.6799e-04
4	2.0873e-05	6.1273e-04	2.1886e-06	-2.0849e-04	-2.3674e-04
5	-5.3952e-04	-5.9516e-04	2.5819e-04	1.5922e-04	-4.0860e-04
6	-2.8905e-05	4.9162e-04	4.1515e-04	-6.9828e-04	-7.1998e-04
7	-4.9435e-04	0.0013	1.5723e-05	1.7988e-04	-1.0540e-04
8	5.7871e-05	-6.1892e-04	-1.9210e-04	3.0510e-04	4.8257e-04
9	9.8029e-04	2.8000e-04	-4.9878e-05	3.0917e-04	3.7300e-04
10	3.1728e-04	6.0733e-04	-6.3799e-05	8.2532e-04	1.0684e-04
11	-7.1259e-05	4.3959e-04	8.3630e-04	-1.1130e-05	-1.6076e-04
12	-1.4765e-05	1.0084e-04	7.5986e-04	9.9256e-04	3.9879e-05
13	-4.5592e-04	0.0010	-1.0879e-04	6.0349e-04	4.5389e-04
14	-4.9054e-04	1.4180e-05	-3.8788e-04	1.0730e-04	-6.9995e-05
15	-3.6745e-04	-6.5810e-04	-8.1993e-04	-4.7793e-04	-4.0274e-04
16	-2.9311e-04	-8.0800e-04	-6.0655e-04	-2.8273e-04	-3.4499e-05
17	-1.9450e-04	-4.0220e-04	-1.7665e-04	8.0156e-05	-4.8406e-04
18	4.2758e-04	-3.0120e-05	-2.4855e-04	-4.8459e-05	-8.3416e-04
19	-2.4160e-05	-3.7447e-04	-0.0016	-3.5297e-05	3.4712e-04
20	-5.4043e-05	5.7581e-04	6.5369e-04	-7.4487e-04	-1.4385e-04

Figura 11. Fragmento de la matriz de i-vectors de los hablantes.

4.2 Entrenamiento de la matriz T

La matriz de variabilidad total T se entrena con un proceso iterativo de tipo EM. Los datos de entrenamiento consisten en grabaciones de todos los hablantes que participarán de la verificación.

Luego de inicializar la matriz T aleatoriamente, se calculan las estadísticas de Baum-Welch de orden cero y uno para cada grabación u de entrenamiento. A diferencia del entrenamiento de la matriz V de JFA, se calculan las estadísticas para cada grabación porque para el entrenamiento de la matriz T se asume como si cada grabación fuera pronunciada por una persona diferente.

Luego, para una grabación u definida por una secuencia de N vectores de características $x_1, \ldots\ldots, x_N$ pertenecientes al hablante s donde cada vector es de tamaño D, y un UBM $\Lambda^{(b)} = \left\{\lambda_i^{(b)}, \mu_i^{(b)}, \Sigma_i^{(b)}\right\}_{i=1}^{C}$ con C componentes de mezclas, se calculan las estadísticas de Baum-Welch de orden cero y de primer orden.

$$N_c = \sum_t x_k \;\; ; \qquad F_c = \sum_t (x_k) * x_k$$

Dado que la alineación de los vectores de características con los componentes de mezcla no es un hecho dado, es decir, no se puede garantizar de que ciertas componentes gaussianas nos siguen generando los mismos vectores de características; es por esto que se procede al cálculo de las estadísticas de Baum-Welch, estas estadísticas son el resultado de un algoritmo iterativo donde en cada iteración se aumente la probabilidad de generar una nueva una expresión. El algoritmo de Baum-Welch pertenece a la familia de métodos EM, solo que aquí los pasos de expectativa y maximización son realizados de forma simultánea.

$$N_c\,(u) = \sum_{k=1}^{N} \gamma_k(c) \qquad (15)$$

$$F_c\,(u) = \sum_{k=1}^{N} \gamma_k(x_k) \qquad (16)$$

$$\vec{F}_s\,(u) = \sum_{k=1}^{N} \gamma_c(x_k)(x_k - \mu_C) = F_c\,(u) - N_c\,(u) * \mu_c \qquad (17)$$

donde las estadísticas de orden cero N_c nos indican la probabilidad posterior $(\gamma_k(c))$, esto nos indica la probabilidad de que una expresion se encuentre en cierto punto de estimación o estado en el tiempo t, y las estadísticas de primer orden F_c nos indican la probabilidad posterior de que x_t sea generado por un componente de mezcla de $i = 1, \ldots, C$ y el cálculo de estas estadisticas siguen viendo que el siguiente vector siga siendo generado por la misma componente c en el tiempo t [27]

Una vez obtenido las estadísticas de una grabación u se puede continuar a calcular su i-vector w_s el cual se calcula de la siguiente manera:

$$w_s = L_s^{-1} T^T \Sigma^{(b)-1} \vec{F}_s \qquad (18)$$

donde

$$L_s = I + T^T \Sigma^{(b)-1} N_s T \qquad (19)$$

es una matriz de precisión, I es la matriz identidad, N_s es una matriz diagonal de dimensión $CDxCD$ donde sus elementos de la diagonal son bloques dados por $N_c(u)I\ (C = 1, \ldots, C)$, $\vec{F}_s$ es un vector de dimensión $CDx1$ formado por la concatenación de las estadísticas de primer orden centradas de Baum-Welch para una expresión dada u, $\vec{F}_s = \left[\vec{F}_1(u); \ldots; \vec{F}_C(u)\right]$, $\Sigma^{(b)}$ es la matriz de covarianza que modela la variabilidad residual no capturada por la matriz de variabilidad total T de dimensión $CDxR$. En la práctica se sustituye esta matriz por las matrices de covarianza del modelo UBM $\Sigma^{(b)} = diag\left\{\Sigma_1^{(b)}, \ldots\ldots, \Sigma_C^{(b)}\right\}$. La w_s es la media posterior que viene siendo el i-vector calculado en las ecuaciones 18 y 19 que nos representa al hablante s y se calculará para cada grabación del conjunto de entrenamiento.

Posteriormente una vez calculado el i-vector w_s se procede a actualizar la matriz de variabilidad total T dadas las siguientes ecuaciones, claro que esto es para hacer más robusta la matriz en cada iteración. Su actualización se define de la siguiente manera

$$C_i = \sum_s F_{s,i} w_s^T \qquad (20)$$

$$A_i = \sum_s N_{s,i} \left(L_s^{-1} + w_s w_s^T\right) \qquad (21)$$

$$T_i = C_i A_i^{-1} \qquad i = 1, \ldots\ldots, C \qquad (22)$$

Donde T_i es una submatriz de T de dimensión $D\ x\ R$. En resumen, la matriz de variabilidad total puede ser obtenida de igual forma por el algoritmo iterativo EM, donde el paso E equivale al cálculo de las ecuaciones 18 y 19 y el paso M al cálculo de las ecuaciones 20, 21 y 22.

En la figura 12 está la representación de una matriz de variabilidad total la cual contiene tanto las características de cada persona como las características de las fuentes no deseadas, así como el ruido del canal. El número de filas es igual al tamaño del i-vector y las columnas de la matriz es igual al número de coeficientes MFCC trabajados por el numero de componentes gaussianas implementadas en el modelo UBM, para este ejemplo son 13 coeficientes MFCC y 512 gaussianas.

50x6656 single

	1	2	3	4	5	6	7	8
1	1.2986e+03	304.3454	482.2730	-450.1388	2.5091e+03	1.6371e+03	925.3516	-2.5767e+03
2	-2.2594e+03	2.0181e+03	-2.0221e+03	-1.0075e+03	-325.7390	-637.2549	-704.2054	3.7384e+03
3	107.7325	293.6382	-495.1253	-279.9082	1.1055e+03	-614.4387	-232.7239	-590.6366
4	232.4364	-1.6373e+03	1.4847e+03	-282.5307	421.1766	-520.6458	-130.8683	-1.0864e+03
5	547.6380	-1.9757e+03	-226.6517	939.8602	2.3722e+03	1.0711e+03	1.4484e+03	-754.9155
6	35.8838	296.5356	-681.2608	-187.7014	-507.7151	-372.5364	-188.0252	145.1895
7	-597.7451	-957.7278	719.1306	-946.6612	3.3825e+03	-755.8773	614.7466	82.6918
8	-312.9457	-791.7435	-437.4660	505.4991	2.8066e+03	80.8423	-66.2261	993.4421
9	811.0988	-1.4177e+03	-284.4975	3.1892e+03	-621.3872	825.9953	-364.6298	-1.3057e+03
10	34.1245	-2.0970e+03	295.4169	-1.7331e+03	677.2914	-365.7692	-550.8615	300.4455
11	1.2385e+03	696.1030	344.8557	2.0244e+03	-3.4723e+03	1.3271e+03	-22.4182	-1.6238e+03
12	482.4647	301.6622	1.2305e+03	1.0507e+03	-2.4190e+03	-589.4385	-815.5968	-614.7121
13	555.9335	-907.1628	-1.3490e+03	1.0678e+03	377.1248	2.3540e+03	707.7141	-718.0334
14	-877.5187	1.5299e+03	1.0785e+03	-2.0920e+03	2.2447e+03	-1.3019e+03	-53.5703	340.3623
15	-438.3336	433.4825	-548.5379	-1.4573e+03	1.1031e+03	327.4803	154.2576	205.7384
16	1.1012e+03	-3.9596e+03	2.0743e+03	-322.6945	-1.2305e+03	-78.8346	651.8399	-2.2738e+03
17	-383.5022	2.8259e+03	-462.9336	-1.0883e+03	-1.3240e+03	177.9285	800.0921	1.0088e+03
18	36.2419	682.0635	-180.4642	-194.1682	-2.5457e+03	35.4790	-390.9987	-38.2686
19	-415.0982	-447.1557	761.6611	-1.1500e+03	-1.5562e+03	-1.5044e+03	-508.9327	536.8196
20	-974.0378	-2.7732e+03	190.6499	-1.6524e+03	1.4940e+03	-147.2917	-40.1475	2.4879e+03

Figura 12. Fragmento de la matriz de variabilidad total T.

4.3 Extracción del i-vector

Una vez entrenada la matriz de variabilidad total T se procede a calcular la adaptación de los modelos de los hablantes que participarán de la verificación. Observamos de la ecuación (14) que adaptar el modelo del hablante se reduce a estimar el i-vector correspondiente a cada hablante que participará en la verificación.

Para un hablante h y sus datos de entrenamiento, representados por una secuencia de N vectores de características acústicas $x_1, \ldots\ldots, x_N$, se calculan las estadísticas Baum-Welch de orden cero y uno, de acuerdo a las ecuaciones previas. Por lo que finalmente la ecuación para extraer un i-vector que nos represente a un hablante h, se define de la siguiente manera:

$$w(h) = (I + T^T \Sigma^{-1} N_h T)^{-1} T^T \Sigma^{-1} F_h \qquad (23)$$

4.4 Puntaje de Distancia Coseno

La simple métrica de distancia coseno ha sido aplicada satisfactoriamente en el espacio de variabilidad total para comparar dos vectores y tomar una decisión de detectar a un hablante. Dado dos i-vectors la distancia coseno se calcula como:

$$DC = \frac{\ll w_{obj} \gg \ll w_{test} \gg}{|w_{obj}||w_{test}|} \geq \theta \qquad (24)$$

Donde θ es un umbral de decisión, en el que si es mayor o igual al umbral se determina que son la misma persona, en caso contrario se rechaza. Esta función de puntuación es considerablemente menos compleja que las operaciones de puntuación del cociente de probabilidad de registro (Log Likelihood) en GMM [2].

Es importante destacar que el score distancia coseno considera el ángulo entre los dos i-vectors por lo que si el reultado de la distancia el coseno es 1, significa que los vectores son exactamente iguales; si es -1, son exactamente opuestos, si el coseno es 0, significa que los vectores son independientes. Ahora bien, si existe información no relacionada con el locutor (como el canal o la sesión) que esté afectando a la magnitud de los i-vectors no tenerla en cuenta en la etapa de obtención de puntuaciones puede afectar la robustez del sistema, de que ahí la necesidad de que ambos i-vector pasen por una etapa de compensación de la variabilidad del canal.

Capítulo 5. LDA/PLDA

Después de haber realizado el paso de la extracción de los i-vector viene un punto muy importante, el cual muchos sistemas de reconocimiento del hablante lo utilizan para dos cosas. La primera es complementar la información escasa de las personas debido a que podría no tenerse suficientes datos para representar por completo a una persona; y la segunda es porque existen muchas variabilidades en la voz. Esto último es uno de los mayores problemas que se presentan para la implementación en este tipo de sistemas debido a que es muy difícil identificar a un hablante con diferentes variabilidades porque éstas pueden ser intencionales o no intencionales. Es por esto por lo que se implementará PLDA el cual es un método derivado de haber implementado previamente la técnica llamada Linear Discriminant Analysis (LDA).

PLDA se puede implementar junto con i-vectors para poder tener mejores resultados en la evaluación ya que nos permite crear modelos generativos a partir de los i-vectors ya extraídos previamente, estos modelos generativos llamados así en [32] nos permiten modelar tanto la variabilidad de la persona como la del canal que están presentes en los i-vector, debido a no se tiene el control de los sistemas de grabación y los entornos donde se recoge la grabación para poder obtener una buena información de la persona que habla.

Trabajar con estas 3 dificultades, escases de información, variabilidad de la persona y del canal, nos lleva a utilizar PLDA para poder tener un sistema más robusto que pueda tener un mejor desempeño en su tarea de evaluación y pueda entregar mejores resultados.

Como se mencionó PLDA viene después de haber implementado LDA, esta técnica de LDA ayudará a cubrir 2 aspectos importantes del sistema que son, la reducción de la dimensión de los datos y normalización de los datos. El objetivo de LDA es buscar dentro de las matrices de dispersión unos nuevos ejes que nos indican la posición del i-vector maximizando la variabilidad dentro de las grabaciones de la persona y minimizando la variabilidad en el canal, de esta manera lograr tener una menor dimensión en los vectores resultantes que nos representen mejor a cada clase.

LDA utiliza la técnica de normalización de fuente, esto ayuda para mejorar la estimación de las matrices de dispersión del conjunto de datos de entrenamiento,

donde existen insuficiente variabilidad de las expresiones del hablantes de diferentes fuentes [30], haciendo esto se puede hacer frente a uno de los 3 problemas, la escases de la información, donde al normalizar los datos LDA puede encontrar unos mejores ejes que puedan hacer la discriminación entre los diferentes individuos de una mejor manera.

Una vez hecho estos dos pasos en LDA se puede implementar PLDA, que utiliza como base el nuevo espacio creado por LDA para así realizar una mejor discriminación con los nuevos datos proyectados en el espacio de baja dimensión, y además, nos ayuda a cubrir con el último problema que es la variabilidad tanto de la persona como la del canal. Para el espacio de la persona se genera una matriz con los nuevos datos dados por LDA mediante el algoritmo EM y para el espacio del canal modela todas aquellas fuentes de variabilidad no deseables dentro de una sola matriz de covarianza.

Al utilizar LDA y PLDA nos permiten reducir la variabilidad no deseada que sabemos que ésta es introducida por factores tales como la transmisión del canal, ruido de fondo y características de la persona (edad, salud, idioma, etc.) en los i-vector que puedan hacer que para el sistema sea más difícil realizar la tarea de reconocimiento al tener información que es poco informativa de cada persona.

Para obtener buenos resultados mediante la evaluación con PLDA se necesitan una gran cantidad de datos para cada clase. Para un enfoque forense contar con demasiada información podría incluso alejarse de la realidad, donde no se cuenta con mucha información sobre algún individuo, por esta razón es que se emplea primero LDA para la normalización de los datos, y para tener una mejor representación de la información al reducir la dimensión [30].

En los experimentos hechos para esta tesis se comparan también los resultados de PLDA al reducir la dimensión y sin reducir la dimensión para poder ver de una manera más clara las limitaciones del algoritmo, al ir variando la dimensión a la cual se quiere reducir y tener un mejor control del sistema al momento de variar el tiempo de prueba y saber entonces hasta qué punto el sistema debe obtener una buena reducción de la dimensión.

5.1 Linear Discriminant Analysis

Utilizamos esta técnica para poder aplicar la reducción de la dimensionalidad [34] para así con nuestra matriz de datos originales proyectarla a una espacio de menor dimensión, esto se realiza para evitar tener información no relevante de las personas en nuestros i-vectors, reteniendo solo la información más representativa, y además quitando toda posible redundancia entre los datos de una gran matriz. En la figura 13 se muestra de manera gráfica cómo sigue LDA el proceso para obtener las matrices de dispersión entre clases y dentro de las clases para después reducir la dimensión con los eigenvectores que mejor nos describan a los hablantes en el nuevo espacio reducido.

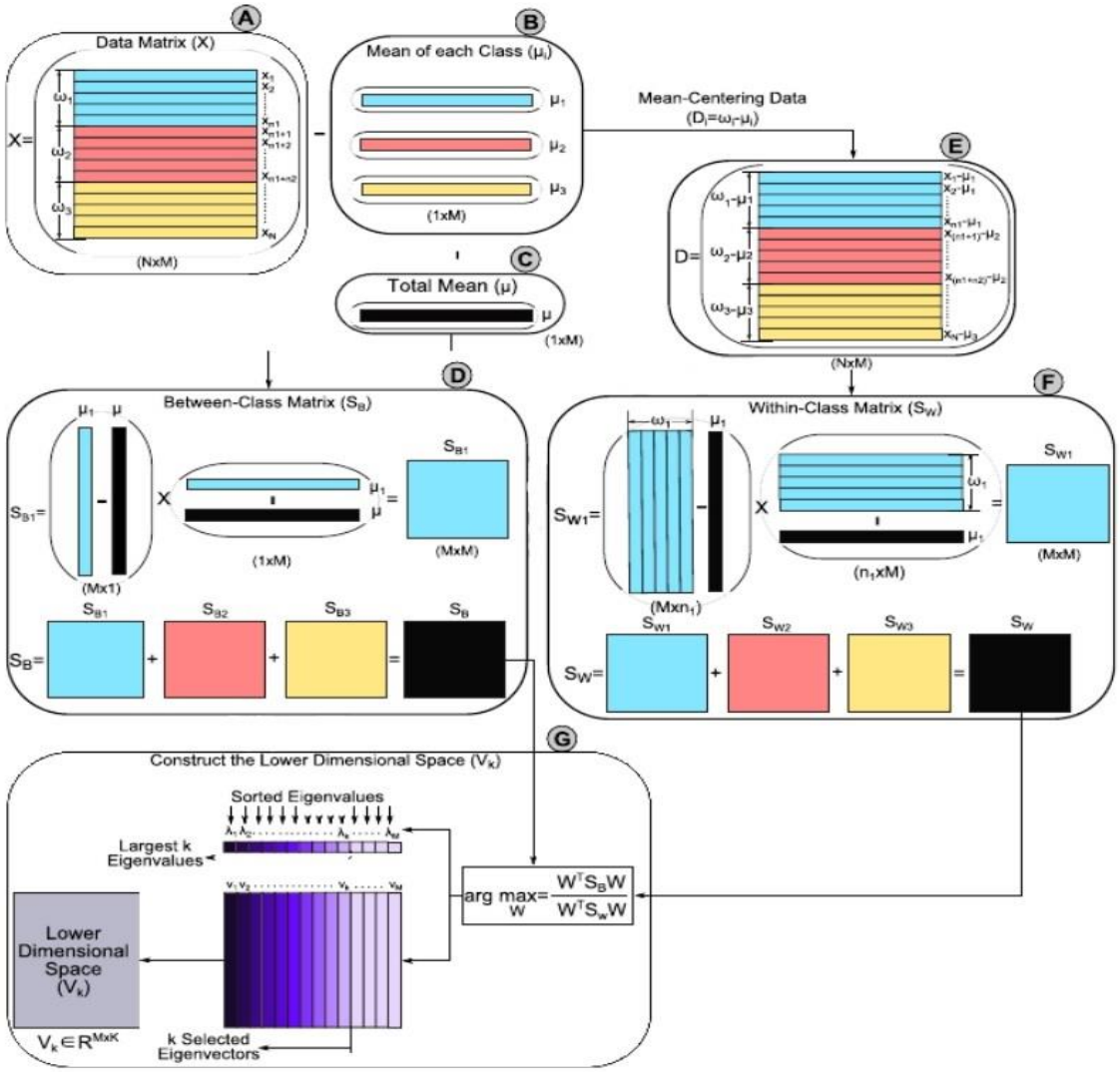

Figura 13. Etapas del algoritmo LDA para el cálculo de varianza entre y dentro de las clases y la construcción de un nuevo espacio de baja dimensionalidad [31].

Como primer paso consiste entonces en el cálculo de encontrar las matrices S_B y S_W, donde S_B es una matriz de dispersion que nos indica la variabilidad que existe entre cada individuo tratando de maximizar esa diferencia entre los hablantes de tal manera que sea más notoria para obtener mejores resultados y S_W es la otra matriz de dispersión que trabaja con los datos presentes dentro de cada clase tratando de minimizar la separación entre ellos de tal manera que representen mejor al individuo al tener más agrupados los datos.

Para poder obtener la matriz S_B se debe realizar el cálculo del vector de media de cada clase y generar con los vectores de medias un vector de media global, para después calcular la separación entre clases. Para S_W el proceso es semejante ya que se debe calcular la distancia entre la media de una clase y los ejemplos que existen en esa clase, y como tercer paso se debe construir un nuevo espacio de baja dimensión que nos va a permitir maximizar la varianza entre las clases y minimizar la varianza dentro de la clase.

5.2 Cálculo de la varianza entre clases (S_B)

La varianza entre clases (S_{Bi}), que es el paso D de la figura 13 representa la distancia entre la media de una clase i (μi) y la media total (μ). La manera de obtener la varianza entre clases o la matriz entre clases (S_B) se hace de la siguiente manera. Dada la matriz de datos original (que viene siendo nuestra matriz T), que contiene a los i-vectors, y cada i-vector (x_I) representa la muestra, el patrón o la observación con la que se entrenó la matriz, y N es el número total de muestras y cada muestra está representada por M características. En otras palabras, cada muestra se representa como un punto en el espacio M-dimensional. Considerando que nuestro conjunto de entrenamiento tiene N ejemplos $\{x^i, \ldots., x^N\}$ donde cada ejemplo x^i es un i-vector. Cada ejemplo de entrenamiento pertenece a una de las K clases. Sea C_K el conjunto de todas las muestras de la clase k y sea $n_k = |C_K|$ el número de muestras en la clase $k = 1 \ldots K$

$$S_B = \sum_{k=1}^{K} n_k (m_k - m)(m_k - m)^T \qquad (25)$$

Donde

$$m_k = \frac{1}{n_k} \sum_{\mathrm{i} \in C_k} x^i \qquad (26)$$

es la media de la clase extraída en el paso B y

$$\mathrm{m} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}} x^i \qquad (27)$$

es la media global de todo el conjunto de entrenamiento extraído en el paso C de la figura 13.

En el paso D de la figura 13 se muestra primero cómo se calcula la matriz entre clases de la primera clase (S_{B1}) y luego cómo se calcula la matriz total entre clases (S_B) sumando todas las matrices entre clases de todas las clases.

5.3 Cálculo de la varianza dentro de la clase (S_W)

La varianza dentro de la clase de la clase k (S_{Wk}) representa la diferencia entre la media y los ejemplos de esa clase (los i-vectors) [2]. LDA busca minimizar la diferencia entre la media proyectada (m_k) y las muestras proyectadas de cada clase (x^i) o simplemente minimiza la varianza dentro de la clase. La varianza dentro de las clases de cada clase (S_W) es calculada como en la ecuación 28.

$$S_w = \sum_{i=1}^{K} \sum_{i \in C_k} n_k (x^i - m_k)(x^i - m_k)^T \qquad (28)$$

Donde K es el numero total de clases, m_k es la media de la clase k y se calcula de la misma forma que en la ecuación 26. En el paso F de la figura 13 se puede ver que primero se calcula la matriz de dispersión para una clase utilizando los i-vectors de esa clase centralizados y la media global, posteriormente suma todas las matrices S_{wk} para obtener S_w.

5.4 Reducción de la dimensión

Después de calcular la variancia entre las clases (S_B) y la varianza dentro de la clase (S_W), la matriz de transformación (W) de la técnica LDA es llamada criterio de Fisher [31] ecuación (29) y puede ser reformulada como en la ecuación 31.

$$arg\ max_w \frac{W^T S_B W}{W^T S_W W} \qquad (29)$$

$$S_w W = \lambda S_B W \qquad (30)$$

donde λ representa los eigenvalores de la matriz de transformación (W) que contiene los eigenvectores. Por lo tanto, para obtenerlos se calcula los eigenvalores ($\lambda = \{\lambda_1, \lambda_2, \ldots..,\ \lambda_M\}$) y los eigenvectores ($V = \{v_1, v_2, \ldots..,\ v_M\}$) de la siguiente ecuación.

$$W = S_W^{-1} S_B \qquad (31)$$

Los eigenvalores son valores escalares, mientras que los eigenvectores son vectores diferentes a cero, lo que nos permite aplicar la ecuación (30) y nos proporciona la información sobre el espacio LDA. Los eigenvectores representan las direcciones del nuevo espacio, y los eigenvalores correspondientes representan el factor de escala, el tamaño o la magnitud de los eigenvectores.

Por lo tanto, cada eigenvector representa un eje del espacio LDA, y el eigenvalor asociado representa la robustez de ese eigenvector. La robustez del eigenvector refleja su capacidad para discriminar entre diferentes clases, es decir, aumenta la varianza entre clases y disminuye la varianza dentro de la clase de cada clase; por lo tanto, cumple con el objetivo LDA. Entonces, los eigenvectores con los k eigenvalores más altos se utilizan para construir un espacio dimensional inferior (V_k), mientras que los otros eigenvectores ($\{v_{k+1}, v_{k+2}, v_M\}$) se descartan como se muestra en la figura 13 en el paso G.

La siguiente ecuación (32) nos describe cómo se proyecta la matriz de datos originales $(X\ \epsilon\ R^{N\ x\ M})$ a un espacio de baja dimensión $(V_k\ \epsilon\ R^{M\ x\ k})$. La dimensión de la proyección de datos es k; por lo tanto, las características de $M - k$ son ignoradas o eliminadas de cada muestra. Entonces, cada muestra (x^i) que fue representada como un punto en un espacio M-dimensional será representará en un espacio k-dimensional de baja dimensión (V_k) como sigue, $y_i = x^i V_k$ [2]

$$Y = XV_k \qquad (32)$$

5.5 Probabilistic LDA

PLDA es un método poderoso para distinguir variabilidad entre los hablantes y que genera información característica del hablante de entre todas las demás fuentes de variabilidad no deseada que caracterizan a las distorsiones [30], surge en [32], donde se propone un modelo generativo para distinguir rostros entre un conjunto de personas. Pero también se ha comprobado que puede trabajar para modelar la voz de las personas en las tareas de reconocimiento del hablante [14, 17]

Existe varios enfoques del algoritmo PLDA como el estándar, simplificado o gaussiano y de dos colas. Para esta tesis se hace uso únicamente del enfoque simplificado o también llamado PLDA gaussiano el cual será explicado más adelante. La diferencia de usar este enfoque frente a los otros que este no trabaja con el espacio del canal, ya que se ha demostrado en [17] que los i-vectors al ser de baja dimensión y utilizando un enfoque estándar o de dos colas no muestran una gran ventaja cuando el espacio del canal no es eliminado.

El aplicar esta técnica junto a los i-vector nos permite discriminar mejor las voces de las personas descomponiendo los i-vectors en 3 partes que son, un vector de medias global, una matriz que contiene el espacio de variabilidad entre las clases y una matriz de residuo la cual contiene la variabilidad no capturada junto con la variabilidad dentro de las clases y la variabilidad dentro las personas, estos tres componentes son el modelo generativo que PLDA obtiene al trabajar con los i-vectors. Según los investigadores PLDA es un algoritmo que muestra un buen desempeño al trabajar con suficientes datos etiquetados bajo diferentes distorsiones o variabilidad de las personas. En esta tesis de hace uso de la técnica reducción de la dimensión con la intención de esperar mejores resultados incluso cuando no se cuentan con suficientes datos representativos de cada persona.

Un modelo PLDA gaussiano se define de la siguiente manera, dado un hablante (H_s) y una colección de i-vectors $X = \{i_1, \dots, i_N\}$ pertenecientes a esa persona, el modelo PLDA gaussiano se expresa como:

$$x_{si} = \mu + Fh_s + \varepsilon_{si} \qquad (33)$$

Donde x_{si} es el i-vector de la persona s, μ es el vector de medias global de todos los i-vectors de entrenamiento, F es una matriz de dimension PxR y sus columnas representan una base donde se representa la variabilidad entre clases, conocida también como matriz de eigenvoices, h_s es un vector de $RX1$ con una distribución normal estándar y se le conoce como variable latente que es la que identifica al i-vector x_{si} dentro del espacio de variabilidad F, y ε_{si} es la matriz residual que contiene la variabilidad no capturada y que representa la variabilidad dentro de las clases que

es equivalente al espacio del canal, la matriz ε_{si} es representada con una distribución normal con media cero y una matriz de covarianza completa Σ.

La primera parte de la ecuación comprende a $\mu + Fh_s$ los cuales representan en el modelo a las personas y definen a la identidad de cada individuo y depende solo de la persona no de una grabación en particular, en otras palabras, describe la variación entre los hablantes. Para encontrar los valores óptimos de los parámetros del modelo para cada hablante se utiliza igualmente el algoritmo iterativo EM para entrenar a PLDA. [33].

5.6 Entrenamiento mediante el algoritmo EM

De la ecuación 33 del modelo PLDA, se observa que descompone la voz en tres partes: un vector global de medias μ, un espacio del hablante F con su factor hs y un espacio de variabilidad no deseada ε_{si} el cual contiene la variabilidad dentro de la clase y el ruido residual, por lo cual los parámetros a calcular son $\{\mu, F, \Sigma\}$, para obtener estos parámetros se hace uso del algoritmo EM. Donde el paso E calcula la distribución posterior de la variable latente h_s con los parámetros iniciales y el paso M calcula los nuevos parámetros.

Paso E:

$$(h_s|X) = L_s^{-1} F^T \Sigma^{-1} \sum_{i=1}^{H_s} (x_{si} - \mu) \qquad (34)$$

$$\left(h_s {h_s}^T \middle| X\right) = L_s^{-1} + (h_s|X)(h_s|X)^T \qquad (35)$$

$$L_s = I + H_i F^T \Sigma^{-1} F \qquad (36)$$

Paso M:

$$\mu' = \frac{\Sigma_{si} x_{si}}{\Sigma_s H_s} \qquad (37)$$

$$F' = \left[\sum_{si} (x_{si} - \mu')(h_s|X)^T\right]\left[\sum_{si} \left(h_s {h_s}^T \middle| X\right)\right]^{-1} \qquad (38)$$

$$\Sigma' = \frac{1}{\sum_{s=1}^{N} H_s}\left\{\sum_{s=1}^{N}\sum_{i=1}^{H_s}[(x_{si} - \mu')(x_{si} - \mu')^T F'(h_s|X)(x_{si} - \mu')^T]\right\} \qquad (39)$$

En las siguientes figuras 14 y 15 se observa como el método PLDA ayuda notablemente a la separación de las características de las personas de toda información de fuentes no deseadas, conteniendo dentro una matriz las características que son dependientes del hablante y en otra la mayor parte de las características que forman parte y que son originarias de fuentes no deseadas tanto del ruido ambiental como del ruido del canal.

50x138 double

	1	2	3	4	5	6	7
1	0.1164	-0.0667	-0.1549	0.0092	0.0211	-0.0735	0.0291
2	-0.2109	-0.0991	0.1194	0.0328	0.0821	-0.0754	0.0918
3	0.0030	-0.1171	0.2140	-0.0316	-0.0244	0.0262	0.1329
4	-0.0491	-0.1781	0.1314	0.0439	-0.0368	-0.0720	-0.0343
5	0.0119	0.1454	-0.0465	-0.0563	0.0229	-0.0841	-0.0167
6	-0.1777	0.0193	0.1567	-0.1001	-0.1959	0.0893	-0.0864
7	0.0972	-0.0855	0.0920	-0.0356	-0.1123	0.0085	0.1208
8	-0.1024	0.1514	0.1296	0.0163	0.0069	-0.0178	-0.1847
9	-0.0762	-0.0887	-0.0290	0.1914	-0.0693	0.0578	-0.1220
10	0.0560	0.0375	0.0751	0.0330	-0.0829	-0.0034	0.1149
11	0.0875	0.1298	0.0933	0.1169	0.1764	0.0480	0.1409
12	0.0826	0.0219	0.1492	-0.1125	-0.0102	0.0533	-0.0684
13	0.0604	-0.1140	0.0681	0.0406	0.0341	-0.0959	-0.1162
14	0.1054	0.0197	-0.0571	0.0547	-0.1111	0.1611	-0.0642
15	0.0123	-0.1299	0.0552	-0.1369	0.0751	-0.0015	-0.1767

Figura 14. Matriz F obtenida por el modelo generativo de PLDA.

La figura 14 es la representación de la matriz F de la ecuación 33 y la figura 15 es toda la información no deseada modelada dentro de la matriz ε de la ecuación 33, se puede observar que PLDA cumple su objetivo de realizar un modelo generativo donde separa los i-vectors de la información del canal que vienen siendo todas las características no deseadas que pueden perjudicar los resultados cuando se quiere hacer una evaluación más precisa y cuando no se tiene suficientes datos, los números obtenidos del modelo PLDA para la matriz F son mucho más limpios y representativos que los obtenidos para la matriz ε de la figura 15 y los de la matriz T de la figura 12.

50x50 double

	1	2	3	4	5	6	7
1	1.4873e-04	5.0847e-05	-6.9247e-04	-2.2893e-05	-9.6836e-04	-7.9823e-05	-4.3253e-04
2	5.0847e-05	1.7383e-05	-2.3673e-04	-7.8264e-06	-3.3105e-04	-2.7289e-05	-1.4787e-04
3	-6.9247e-04	-2.3673e-04	0.0032	1.0659e-04	0.0045	3.7164e-04	0.0020
4	-2.2893e-05	-7.8264e-06	1.0659e-04	3.5237e-06	1.4905e-04	1.2286e-05	6.6575e-05
5	-9.6836e-04	-3.3105e-04	0.0045	1.4905e-04	0.0063	5.1970e-04	0.0028
6	-7.9823e-05	-2.7289e-05	3.7164e-04	1.2286e-05	5.1970e-04	4.2840e-05	2.3213e-04
7	-4.3253e-04	-1.4787e-04	0.0020	6.6575e-05	0.0028	2.3213e-04	0.0013
8	8.0943e-04	2.7672e-04	-0.0038	-1.2459e-04	-0.0053	-4.3441e-04	-0.0024
9	-2.1712e-05	-7.4228e-06	1.0109e-04	3.3420e-06	1.4136e-04	1.1653e-05	6.3141e-05
10	-2.1612e-04	-7.3886e-05	0.0010	3.3266e-05	0.0014	1.1599e-04	6.2851e-04
11	6.7221e-04	2.2981e-04	-0.0031	-1.0347e-04	-0.0044	-3.6076e-04	-0.0020
12	2.5723e-05	8.7939e-06	-1.1976e-04	-3.9593e-06	-1.6748e-04	-1.3805e-05	-7.4805e-05
13	3.0570e-04	1.0451e-04	-0.0014	-4.7053e-05	-0.0020	-1.6406e-04	-8.8900e-04
14	5.1772e-04	1.7699e-04	-0.0024	-7.9686e-05	-0.0034	-2.7785e-04	-0.0015
15	4.9108e-04	1.6788e-04	-0.0023	-7.5586e-05	-0.0032	-2.6355e-04	-0.0014

Figura 15. Matriz ε obtenida por el modelo generativo de PLDA.

Las filas de la matriz F representan el tamaño del i-vector y sus columnas representan el tamaño del nuevo espacio reducido, el ejemplo de la figura 14 no representa todavía

un espacio que sea representativo, debido a que el tamaño de sus columnas como mínimo debe cumplir ser menor que el número de los hablantes por lo cual al menos se garantiza que una dimensión fue quitada.

Continuando con el mismo ejemplo, en las siguientes dos figuras 16 y 17 se hace ya un uso notable de la técnica de reducción de la dimensionalidad encontrando un mejor espacio representativo, al pasar de 138 columnas a 40 columnas y reduciendo además el tamaño del i-vector a 40 características para ayudarle al modelo generativo de PLDA a obtener incluso aún mejores resultados que son más representativos para cada persona para escenarios donde no se cuenta con suficientes datos de entrenamiento de cada hablante, el uso de esta técnica ayudó notablemente en los resultados de los experimentos mostrados en el capítulo 6.

40x40 double

	1	2	3	4	5	6	7
1	-0.1124	-0.2244	-0.1736	0.0907	-0.0458	0.0495	-0.1825
2	-0.1944	0.0842	-0.2366	0.0424	-0.1450	0.0419	0.0491
3	0.1311	-0.0874	0.0717	-0.1259	-0.1588	0.0055	0.0273
4	-0.0497	-0.0687	-0.0027	0.0757	-0.1221	0.0702	-0.0469
5	0.1017	-0.0224	0.0450	-0.2292	-0.0119	0.0892	-0.0734
6	0.0225	0.0279	0.0567	-0.0757	0.0932	0.1446	0.0290
7	-0.2071	0.0581	0.0783	0.1281	0.1729	0.4188	0.0412
8	0.1905	0.1325	-3.6061e-04	-0.0744	0.1642	0.0891	-0.2673
9	-0.0126	-0.3074	0.1441	-0.0124	0.0492	-0.0235	-0.0289
10	-0.3327	-0.0234	0.1668	0.1858	0.0403	0.0890	-0.2499
11	-0.0218	-0.0492	0.1208	-0.0276	-0.2138	-0.0162	0.0171
12	-0.1503	-0.0534	0.1227	-0.3495	0.1433	-0.1671	-0.2768
13	0.0646	0.1389	0.0563	0.1068	0.1512	-0.1163	0.1583
14	-0.0992	0.0582	-0.0667	-0.0768	-0.1908	0.0319	0.0927
15	0.1398	-0.0384	0.0601	0.1243	-0.0910	-0.1652	-0.0958

Figura 16. Matriz F obtenida por el modelo generativo de PLDA después de aplicar la reducción de la dimensionalidad.

40x40 double

	1	2	3	4	5	6	7
1	1.4519e-05	-9.9276e-06	8.1715e-06	-7.9841e-07	-1.1866e-05	5.3781e-06	5.2196e-06
2	-9.9276e-06	6.7884e-06	-5.5875e-06	5.4594e-07	8.1138e-06	-3.6774e-06	-3.5691e-06
3	8.1715e-06	-5.5875e-06	4.5991e-06	-4.4937e-07	-6.6785e-06	3.0269e-06	2.9377e-06
4	-7.9841e-07	5.4594e-07	-4.4937e-07	4.3907e-08	6.5254e-07	-2.9575e-07	-2.8704e-07
5	-1.1866e-05	8.1138e-06	-6.6785e-06	6.5254e-07	9.6981e-06	-4.3955e-06	-4.2659e-06
6	5.3781e-06	-3.6774e-06	3.0269e-06	-2.9575e-07	-4.3955e-06	1.9922e-06	1.9335e-06
7	5.2196e-06	-3.5691e-06	2.9377e-06	-2.8704e-07	-4.2659e-06	1.9335e-06	1.8765e-06
8	4.5256e-06	-3.0946e-06	2.5471e-06	-2.4888e-07	-3.6988e-06	1.6764e-06	1.6270e-06
9	-4.9429e-06	3.3799e-06	-2.7820e-06	2.7182e-07	4.0398e-06	-1.8310e-06	-1.7770e-06
10	-2.0453e-05	1.3986e-05	-1.1512e-05	1.1248e-06	1.6717e-05	-7.5765e-06	-7.3532e-06
11	5.8459e-06	-3.9974e-06	3.2902e-06	-3.2148e-07	-4.7779e-06	2.1655e-06	2.1017e-06
12	3.7904e-06	-2.5918e-06	2.1334e-06	-2.0845e-07	-3.0979e-06	1.4041e-06	1.3627e-06
13	-1.8064e-05	1.2352e-05	-1.0167e-05	9.9337e-07	1.4764e-05	-6.6913e-06	-6.4941e-06
14	-2.0437e-05	1.3975e-05	-1.1503e-05	1.1239e-06	1.6703e-05	-7.5705e-06	-7.3474e-06
15	2.4066e-06	-1.6456e-06	1.3545e-06	-1.3234e-07	-1.9669e-06	8.9145e-07	8.6518e-07

Figura 17. Matriz ε obtenida por el modelo generativo de PLDA después de aplicar la reducción de la dimensionalidad.

5.7 Evaluación PLDA

Después de la estimación de sus parámetros, PLDA nos permite hacer la evaluación entre dos personas, para cada dos i-vector de prueba $\boldsymbol{x_1}$y $\boldsymbol{x_2}$ el puntaje de verificación será calculado usando el log-likelihood ratio de la hipótesis H_s, contra la hipótesis H_d donde los dos i-vectors pertenecen a diferentes personas [32]. Si ambos i-vectors pertenecen a la misma persona significa entonces que los i-vector tienen la misma variable de identidad hi, en caso contrario los dos i-vector pertenecen a diferentes personas y tienen diferentes variables de identidad. Los resultados para la puntuación de PLDA se indican como PLDA-Gaussian.

$$score\ (x_1, x_2)\ = \frac{p(x_1, x_2|\ H_s)}{p(x_1, x_2|\ H_d)} \qquad (40)$$

Dado la suposición de que los i-vector manejan una distribución gaussiana se puede calcular el puntaje de evaluación log-likelihood planteando la siguiente formula [33]

$$score\ (x_1, x_2) = \ [\boldsymbol{x_1}^T\ \boldsymbol{x_t}^T] \begin{bmatrix} \Sigma + FF^T & FF^T \\ FF^T & \Sigma + FF^T \end{bmatrix}^{-1} [\boldsymbol{x_1}\ \ \boldsymbol{x_2}]$$
$$-\ \boldsymbol{x_1}^T[\Sigma + FF^T]^{-1}\boldsymbol{x_1} -\ \boldsymbol{x_t}^T[\Sigma + FF^T]^{-1}\boldsymbol{x_2} \qquad (41)$$

Capítulo 6. Experimentos

6.1 Corpus Valquiria

Este corpus es el que fue utilizado tanto para la etapa de entrenamiento como la etapa de prueba para el sistema. Las características principales con las que cuenta este corpus es tener una colección de llamadas telefónicas desde un teléfono público a celular, de público a teléfono fijo, celular a celular, celular a fijo, de fijo a fijo. Por lo tanto, son grabaciones en entornos reales que es lo que buscamos para comprobar el desempeño de i-vector con PLDA a pesar de no ser grabaciones en un ambiente controlado sin ruido. Además, son grabación hechas con el español del centro de la Ciudad de México.

Las grabaciones con las que cuenta el corpus son de personas con los siguientes rangos de edades: de 18 a 30, de 31 a 45, de 46 a 60 y de 60 en adelante. En cada uno de estos rangos se grabaron a 7 mujeres y 7 hombres diferentes, tomando en cuenta los 5 diferentes tipos de llamadas realizadas, dan un total de 280 grabaciones en el corpus, todas de diferentes personas.

El “Corpus valquiria” fue diseñado y creado por el “laboratorio de tecnologías del lenguaje” de la Facultad de Ingeniería y el grupo de ingeniería lingüística durante los años 2015 y 2016. Este corpus aún no ha sido liberado al público, ya que se están realizando sus transcripciones fonéticas; las transcripciones antes referidas no son utilizadas en esta tesis.

Cabe mencionar que si se contara con un corpus grabado en mejores condiciones tanto ambientales como con mejores canales de grabación se esperaría que el sistema desarrollado en esta tesis mostrara mejores resultados debido a que se obtendría mejores características de las voces de las personas, pero como ha sido resaltado en esta tesis el contar con un corpus deseado puede alejarse de la realidad, además en la mayoría de los casos prácticos las grabaciones no van a ser obtenidas en un entorno controlado.

Los nombres de los archivos tienen un formato que permite identificar las características del hablante como lo son el género, rango de edad, el tipo de llamada y al final un índice que va del 0001 al 0007. A continuación, se listan las diferentes formas de identificar cada una de estas.

- Género: 01 para hombres, 02 para mujeres
- Rango de edad: “a” corresponde de 18 a 30 años, “b” de 31 a 45, “c” de 46 a 60 y “d” de 60 en adelante.
- Tipo de llamada: 01 calle a celular, 02 calle a fijo, 03 celular a celular, 04 celular a fijo y 06 fijo a fijo

Por ejemplo, si el nombre del archivo es el siguiente 02-a-04-0003svr.wav, significa que se trata de una mujer, de edad entre 18 y 30 años, la llamada se grabó de celular a fijo y es la tercera persona en esta categoría de las 7 presentes.

6.2 Condiciones

Para la fase de los experimentos fue utilizado un entrenamiento de 5 y 10 segundos variando el tiempo de prueba hasta llegar a toda la duración de la grabación ya que i-vector puede trabajar con expresiones cortas y/o largas sin importar el tiempo que dura una expresión.

Tanto para los entrenamientos de 5 segundos,10 segundos y los tiempos de audio de prueba fueron utilizados los mismos valores de los parámetros. En el procesamiento de la señal de voz se utilizaron 13 MFCC, un análisis de 25 milisegundos para el ventaneo con un traslape de 10 milisegundos y un filtro preénfasis con $\alpha = 0.97$.

Para los i-vector se trabajó un i-vector de tamaño 50, una reducción de dimensión de 10, modelos UBM de 256 gaussianas para el entrenamiento de 5 segundos y 512 gaussianas para el entrenamiento de 10 segundos.

Los valores de los parámetros utilizados en esta tesis son valores establecidos por los programas y sugeridos por los investigadores, puede modificarse su valor de todos los parámetros si es requerido, para esta tesis se decidió trabajar con los parámetros establecidos, modificando únicamente el tamaño del i-vector, la reducción de la dimensión y el número de iteraciones en los entrenamientos.

Durante todos los experimentos se trabajó diferentes dimensiones de i-vector como 400, 200 100 y 50, se observó en los resultados que un i-vector de dimensión 50 presentaba mejores resultados que los que son de mayor dimensión llegando a la conclusión de que debido a que no se está trabajando con un corpus amplio con un mayor número de hablantes y un mayor número de grabaciones por persona las dimensiones mayores a 50 presentaban un menor desempeño porque se captura más variabilidad no deseada en el i-vector, es por esto que todos los experimentos mostrados a continuación fueron trabajados con un i-vector de dimensión 50.

De igual la forma el sistema mostró un mejor desempeño al trabajar con una reducción de dimensión de 10 al intentar con diferentes tamaños de reducción. En las siguientes secciones se muestra los resultados obtenidos para mujeres y hombres.

6.3 Mujeres

N° Audios Entrenamiento	N° Audios Prueba	Entrenamiento (seg)	i-vector	N° Gaussianas	N° MFCC	R.D.
140	30	5	50	256	13	40

Tabla 4. Parámetros utilizados para las pruebas de mujeres con un entrenamiento de 5 segundos.

Se tomó una muestra aleatoria de tamaño 30 para las voces de la fase de prueba. Las siguientes tablas muestran el resultado de cada sistema al enfrentarse con diferentes tiempos de prueba de cada audio.

Tiempo de prueba (seg)	PLDA	Precisión PLDA	CDS	Precisión CDS	GMM	Precisión GMM	PLDA-RD	Precisión PLDA-RD
5	30/30	100%	30/30	100%	30/30	100%	30/30	100%
10	24/30	80%	28/30	93%	30/30	100%	29/30	96%
20	19/30	63%	26/30	86%	27/30	90%	27/30	90%
30	15/30	50%	26/30	86%	24/30	80%	27/30	90%
40	13/30	43%	22/30	73%	24/30	80%	26/30	86%
50	12/30	40%	23/30	76%	23/30	76%	26/30	86%
60	10/30	33%	21/30	70%	18/30	60%	26/30	86%

Tabla 5. Resultados de los sistemas y su precisión con 5 segundos de entrenamiento y variando el tiempo de prueba utilizando las voces de mujeres.

Aumentando el entrenamiento a 10 segundos y el número de gaussianas a 512.

N° Audios Entrenamiento	N° Audios Prueba	Entrenamiento (seg)	i-vector	N° Gaussianas	N° MFCC	R.D.
140	30	10	50	512	13	40

Tabla 6. Parámetros utilizados con un entrenamiento de 10 segundos para voces de mujeres.

Tiempo de prueba (seg)	PLDA	Precisión PLDA	CDS	Precisión CDS	GMM	Precisión GMM	PLDA-RD	Precisión PLDA-RD
30	30/30	100%	30/30	100%	30/30	100%	30/30	100%
60	28/30	93%	30/30	100%	30/30	100%	30/30	100%
90	28/30	93%	30/30	100%	30/30	100%	30/30	100%
120	28/30	93%	30/30	100%	27/30	90%	30/30	100%
Audio completo	26/30	86%	30/30	100%	24/30	80%	30/30	100%

Tabla 7. Resultados de los sistemas y su precisión con 10 segundos de entrenamiento y variando el tiempo de prueba utilizando las voces de mujeres.

Se recolectó una segunda muestra para la fase prueba con audios que cumplieran con una duración mayor a 3 minutos dentro del corpus Valquiria, con la finalidad de observar el desempeño los sistemas al enfrentarse con los audios de mayor duración

con tan solo un entrenamiento de 10 segundos. En totalidad se encontraron 22 audios que cumplían con una duración mayor a los 3 minutos.

N° Audios Entrenamiento	N° Audios Prueba	Entrenamiento (seg)	i-vector	N° Gaussianas	N° MFCC	R.D.
140	22	10	50	512	13	40

Tabla 8. Parámetros utilizados con un entrenamiento de 10 segundos para la segunda muestra de prueba para voces de mujeres.

Tiempo de prueba (seg)	PLDA	Precisión PLDA	CDS	Precisión CDS	GMM	Precisión GMM	PLDA-RD	Precisión PLDA-RD
Audio completo	18/30	73%	22/22	100%	18/22	73%	22/22	100%

Tabla 9. Resultados de los sistemas y su precisión frente a la muestra de mayor variabilidad de las voces de mujeres.

6.4 Hombres

N° Audios Entrenamiento	N° Audios Prueba	Entrenamiento (seg)	i-vector	N° Gaussianas	N° MFCC	R.D.
140	30	5	50	256	13	40

Tabla 10. Parámetros utilizados para las pruebas de hombres con un entrenamiento de 5 segundos.

De igual forma para la fase de prueba en voces de hombres se tomó una muestra aleatoria de tamaño 30. Los resultados obtenidos de cada sistema son mostrados en las siguientes tablas.

Tiempo de prueba (seg)	PLDA	Precisión PLDA	CDS	Precisión CDS	GMM	Precisión GMM	PLDA-RD	Precisión PLDA-RD
5	30/30	100%	30/30	100%	30/30	100%	30/30	100%
10	22/30	73%	27/30	90%	30/30	100%	30/30	100%
20	14/30	46%	23/30	76%	28/30	93%	27/30	90%
30	12/30	40%	16/30	53%	24/30	80%	25/30	83%
40	11/30	36%	14/30	46%	22/30	73%	21/30	70%
50	11/30	36%	13/30	43%	19/30	63%	20/30	66%
60	10/30	33%	11/30	36%	17/30	56%	20/30	66%

Tabla 11. Resultados de los sistemas y su precisión con 5 segundos de entrenamiento y variando el tiempo de prueba utilizando las voces de hombres.

Aumentando el entrenamiento a 10 segundos y el número de gaussianas a 512.

N° Audios Entrenamiento	N° Audios Prueba	Entrenamiento (seg)	i-vector	N° Gaussianas	N° MFCC	R.D.
140	30	10	50	512	13	40

Tabla 12. Parámetros utilizados para las pruebas con un entrenamiento de 10 segundos para las voces de hombres.

Tiempo de Prueba (seg)	PLDA	Precisión PLDA	CDS	Precisión CDS	GMM	Precisión GMM	PLDA-RD	Precisión PLDA-RD
30	24/30	80%	30/30	100%	30/30	100%	30/30	100%
60	14/30	46%	30/30	100%	28/30	93%	30/30	100%
90	13/30	43%	29/30	96%	28/30	93%	29/30	96%
120	15/30	50%	29/30	96%	26/30	86%	29/30	96%
Audio completo	17/30	56%	28/30	93%	24/30	80%	29/30	96%

Tabla 13. Resultados de los sistemas y su precisión con 10 segundos de entrenamiento y variando el tiempo de prueba.

Al igual que en los experimentos con las voces de mujeres, con los hombres se pretende también observar el comportamiento de cada sistema al enfrentarse a la máxima durabilidad que pueden tener los audios en el corpus con voces de hombres.

Esta segunda muestra recolectada tiene un tamaño de 17 audios de personas diferentes donde cada uno tiene una durabilidad mayor a 3 minutos. Las siguientes tablas muestran los parámetros trabajados y los resultados obtenidos por los sistemas.

N° Audios Entrenamiento	N° Audios Prueba	Entrenamiento (seg)	i-vector	N° Gaussianas	N° MFCC	R.D.
140	17	10	50	512	13	40

Tabla 14. Parámetros utilizados para las pruebas con un entrenamiento de 10 segundos para la segunda muestra de prueba en voces de hombres.

Tiempo de prueba (seg)	PLDA	Precisión PLDA	CDS	Precisión CDS	GMM	Precisión GMM	PLDA-RD	Precisión PLDA-RD
Audio completo	6/17	35%	17/17	100%	13/17	76%	17/17	100%

Tabla 15. Resultados de los sistemas y su precisión frente a la muestra de mayor variabilidad.

Capítulo 7. Conclusiones

El sistema desarrollado en esta tesis de reconocimiento del hablante basado en i-vectors ha demostrado tener un mejor desempeño al enfrentarse contra una gran variabilidad de voz frente a otro sistema como GMM. El apoyarse junto a otro método como PLDA fortalece al sistema a clasificar de una mejor manera las voces de las personas, ya que este método divide las voces en dos partes, una parte enfocada en la personalidad del hablante y otra enfocada a la variabilidad no deseada lo cual hace que se le dé un mejor trato de análisis a las voces. Además, cuando no se cuenta con grandes cantidades de datos para el entrenamiento PLDA con reducción de dimensión es una buena opción para hacer nuestro sistema más robusto.

El hacer usos de los parámetros recomendados por los investigadores brindó mejores resultados tanto para una evaluación con distancia coseno como con PLDA que fueron los que más se mantuvieron con un mejor rendimiento a diferencia de GMM. Durante los experimentos el puntaje de evaluación de distancia coseno mostró siempre buenos resultados y un gran desempeño a la hora de enfrentarse a una mayor variabilidad, por lo cual se puede concluir que usar esta métrica no sería una mala opción si se cuenta con mejores datos y una mayor cantidad para la etapa de entrenamiento. Además si nos interesara el tiempo de ejecución de los métodos de evaluación, la distancia coseno a diferencia de la evaluación con Maximum Likelihood de GMM y la evaluación de PLDA presentó durante los experimentos un menor tiempo de procesamiento debido a que su fórmula de cálculo es mucho menos compleja que la de los otros métodos lo cual la hace mucho más rápida, aunque en este caso PLDA resultó ser un método de evaluación superior a los demás métodos.

Se observó durante los experimentos que el sistema i-vectors con PLDA puede brindar malos resultados al no ser tratado con un buen entrenamiento, ya que al brindarle poca información como en el entrenamiento de 5 segundos y recibir un entrenamiento de la matriz T con un numero de iteraciones mayor a 50 el sistema logra un sobreajuste (*overfitting* en inglés) en los datos de entrenamiento que hace que en la evaluación el sistema presente un mal desempaño, por lo cual el no hacer un análisis previo de los datos con los que se cuenta y enfocarse en pretender robustecer al sistema en la fase de entrenamiento puede lograr un mal desempeño del sistema. En esta tesis se analizó el uso de los datos con los que se contaba y probando un entrenamiento no mayor a 50 iteraciones ni menor a 10 iteraciones se

logró obtener mejores resultados que cuando se hacía con un mayor número de iteraciones y que cuando se hacía con un menor número de iteraciones, ya que también no generalizaba de manera correcta y por lo tanto el sistema no aprendía y no se ajustaba de una mejor manera a los datos. Se realizaron experimentos en condiciones difíciles de enfrentamiento, razón por la cual los sistemas no logran mejorar su desempeño al aumentar el tiempo de prueba. El sistema GMM usado en esta tesis ya se usó en condiciones mejoradas y sí obtiene los resultados esperados [19].

El uso de técnicas de reducción de dimensión ayuda notablemente al sistema cuando no se cuenta con los suficientes datos para representar a cada persona y se busca un nuevo espacio que maximice la separación entre clases y minimice la variabilidad interna que hay dentro de cada clase, esto ayudó al sistema a obtener mejores resultados a la hora de la evaluación, ya que presenta tanto mejores resultados como un mejor rendimiento cuando el tiempo de prueba seguía variando. Un punto importe en estas conclusiones es que no se requiere siempre el uso de técnicas reducción de la dimensión cuando se cuenta con un corpus robusto en grabaciones de cada persona, cuando se cuenta con grabaciones que son de mejor calidad que hacen que se pueda extraer mejores características de las personas y notablemente cuando se tiene un mayor entrenamiento con datos más puros y/o de mayor duración, ya que en los resultados mostrados en esta tesis y durante la fase de los experimentos i-vectors junto a PLDA sin reducción de la dimensión lograba un desempeño favorable e igualable frente a GMM y frente a la reducción de la dimensión e incluso en los experimentos de mujeres con entrenamiento de 10 segundos obtuvo mejores resultados que GMM en algunas evaluaciones.

El objetivo planteado para esta tesis se logró cumplir debido a que se desarrolló un sistema de reconocimiento del hablante basado en i-vector y PLDA los cuales son considerados métodos del estado del arte, además fue un sistema que trabajó con el español de la Ciudad de México que logró tener un buen rendimiento. En la parte de la hipótesis se obtuvieron los resultados esperados, aunque hubo también comportamientos muy destacables por los otros métodos de evaluación como lo es distancia coseno, donde hacer directamente la evaluación de los i-vectors sin PLDA puede traer buenos resultados competitivos y una mayor rapidez del sistema, si se contará con una mejor calidad de los datos de entrenamiento y una mayor cantidad posiblemente la parte de PLDA podría omitirse. Para esta tesis el hacer uso de la reducción de la dimensión sí mostró un mejor desempeño frente a GMM y los demás métodos de valuación, como se esperaba y como los investigadores también mencionan y por lo cual el sistema i-vectors con PLDA es mejor frente a su antecesor GMM.

El desarrollar sistemas basados en inteligencia artificial abren las puertas a un gran campo de investigación y de desarrollo de nuevas estrategias y enfoques para diversas tareas y aplicaciones en las que se podría implementar esta nueva

tecnología, ya que la inteligencia artificial dentro de sus propósitos está el procesar grandes masas de información y ayudar a los humanos a tener una mejor herramienta de mayor confiabilidad a la hora de tomar una decisión importante y que para los humanos les sería imposible procesar tanta información a gran velocidad, como es el sistema de reconocimiento del hablante, por lo cual el hacer uso de esta tecnología de inteligencia artificial nos beneficia en gran manera, ya que sabiéndola utilizar se convierte es una herramienta poderosa para mejorar nuestras actividades, como fue mencionado anteriormente nos proporciona nuevas áreas de trabajo.

Referencias

[1] Pruzansky, Sandra. “Pattern-Matching Procedure for Automatic Talker Recognition”. *The Journal of the Acoustical Society of America*, 1963, vol. 35, no 3, p. 354-358.

[2] Clot, Jaume; Hernandos Pericás, Francisco Javier; Nadeu Camprubí, Climent. “Estudio comparativo y nuevas propuestas de técnicas de parametrización de la señal de voz para el reconocimiento del habla”. En *URSI 1994: Unión Científica Internacional de Radio: IX Symposium Nacional: Las Palmas de Gran Canaria: 21-23 de septiembre de 1994*. 1994. p. 1199-1203.

[3] Furui, Sadaoki. “Recent advances in speaker recognition”. *Pattern recognition letters*, 1997, vol. 18, no 9, p. 859-872.

[4] Macas, Macas; Xismena, Diana; Padilla Pineda, Willian Alfonso. “*Estudio de los Modelos Ocultos de Markov y desarrollo de un prototipo para el reconocimiento automático del habla”*. 2012. Tesis de Licenciatura.

[5] Furui, Sadaoki. “Cepstral analysis technique for automatic speaker verification”. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1981, vol. 29, no 2, p. 254-272.

[6] Hernando Pericás, Francisco Javier. “Técnicas de procesado y representación de la señal de voz para el reconocimiento del habla en ambientes ruidosos”. 1993.

[7] Matsui, Tomoko; Furui, Sadaoki. “Concatenated phoneme models for text-variable speaker recognition”. En *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1993. p. 391-394.

[8] Higgins, Alan; Bahler, L.; Pporter, J. “Speaker verification using randomized phrase prompting". *Digital Signal Processing*, 1991, vol. 1, no 2, p. 89-106.

[9] Poritz, A. “Linear predictive hidden Markov models and the speech signal”. En *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1982. p. 1291-1294.

[10] Tisby, Naftali Z. “On the application of mixture AR hidden Markov models to text independent speaker recognition”. *IEEE Transactions on Signal Processing*, 1991, vol. 39, no 3, p. 563-570.

[11] Rose, Richard C.; Reynolds, Douglas A. “Text independent speaker identification using automatic acoustic segmentation”. En *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1990. p. 293-296.

[12] Matsui, Tomoko; Furui, Sadaoki. “Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's”. *IEEE Transactions on speech and audio processing*, 1994, vol. 2, no 3, p. 456-459.

[13] Reynolds, Douglas A.; Quatieri, Thomas F.; DUNN, Robert B. “Speaker verification using adapted Gaussian mixture models”. *Digital signal processing*, 2000, vol. 10, no 1-3, p. 19-41.

[14] Senoussaoui, Mohammed, et al. “Mixture of PLDA models in i-vector space for gender-independent speaker recognition”. En *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

[15] INCIBE, “Tecnologías biométricas aplicadas a la ciberseguridad”, https://www.incibe.es/sites/default/files/contenidos/guias/doc/guia_tecnologias_biometricas_aplicadas_ciberseguridad_metad.pdf, 2016

[16] Ochoa, Felipe; San Martín, César; Carrillo, Roberto. “Identificación biométrica de locutores para el ámbito forense: Estado del arte”. En *VI Congreso Iberoamericano de Acústica-FIA 2008*. 2008.

[17] Kenny, Patrick. “Bayesian speaker verification with heavy-tailed priors”. En *Odyssey*. 2010. p. 14.

[18] Rascón, C. “Transformada de Fourier y la librería de FFTw3”, http://calebrascon.info/AR/Topic4/04.3-Fourier.pdf

[19] Herrera, A.; Zúñiga, A. “Reconocimiento automático de hablantes en el ámbito forense usando MFCC’sy GMM’s”. Memorias de la 26° Reunión de Otoño de Comunicaciones, Computación, Electrónica y Exposición Industrial, ROC&C’2016 de la IEEE Sección México, memoria USB, noviembre del 2016.

[20] Reynolds, Douglas A.; Rose, Richard C. “Robust text-independent speaker identification using Gaussian mixture speaker models”. *IEEE transactions on speech and audio processing*, 1995, vol. 3, no 1, p. 72-83.

[21] Kanagasundaram, Ahilan, et al. “I-vector based speaker recognition on short utterances”. En *Proceedings of the 12th Annual Conference of the International Speech Communication Association*. International Speech Communication Association (ISCA), 2011. p. 2341-2344.

[22] Dehak, Najim, et al. “Front-end factor analysis for speaker verification”. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, vol. 19, no 4, p. 788-798.

[23] Fredes Sandoval, Josué Abraham. “Estudio comparativo de técnicas para robustez de sistemas de verificación de locutor texto independiente”. Universidad de Chile, 2015.

[24] Lei, Lei; KUN, She. “Speaker Recognition Using Wavelet Packet Entropy, I-Vector, and Cosine Distance Scoring”. *Journal of Electrical and Computer Engineering*, 2017, vol. 2017.

[25] Shum, Stephen, et al. “Unsupervised Speaker Adaptation based on the Cosine Similarity for Text-Independent Speaker Verification”. En *Odyssey*. 2010. p. 16.

[26] Mak, Man-Wai. “Fast scoring for mixture of PLDA in i-vector/plda speaker verification”. En *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2015. p. 587-593.

[27] León, Ricardo Antonio Mendoza. “Métodos de inferencia estadística para entrenamiento de modelos ocultos de Markov”. *Elementos*, 2011, vol. 1, no 1, p. 57-70.

[28] Furui, Sadaoki. “Fifty years of progress in speech and speaker recognition”. *The Journal of the Acoustical Society of America*, 2004, vol. 116, no 4, p. 2497-2498.

[29] Kenny, Patrick. “A small footprint i-vector extractor”. En *Odyssey 2012-The Speaker and Language Recognition Workshop*. 2012.

[30] Rodríguez Fonollosa, José Adrián; VIDAL, Josep. “Identificaión ciega adptativa basada en el algoritmo de Baum-Welch”. En *URSI 1994: IX Simposium Nacional de la Unión Científica Internacional de Radio: Las Palmas de Gran Canaria: 21-23 de Septiembre de 1994*. 1994. p. 513-517.

[30] Khosravani, Abbas; Homayounpour, Mohammad M. “A PLDA approach for language and text independent speaker recognition. *Computer Speech & Language”*, 2017, vol. 45, p. 457-474.

[31] Tharwat, Alaa, et al. “Linear discriminant analysis: A detailed tutorial”. *AI communications*, 2017, vol. 30, no 2, p. 169-190.

[32] Prince, Simon JD; Elder, James H. “Probabilistic linear discriminant analysis for inferences about identity”. En *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007. p. 1-8.

[33] Rajan, Padmanabhan, et al. “From single to multiple enrollment i-vectors: Practical PLDA scoring variants for speaker verification”. *Digital Signal Processing*, 2014, vol. 31, p. 93-101.

[34] Matéjka, Pavel, et al. “Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification”. En *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011. p. 4828-4831.