Universidad Autónoma de Madrid
Escuela politécnica superior

Proyecto fin de carrera

ANÁLISIS DE COMPENSACIÓN DE VARIABILIDAD EN RECONOCIMIENTO DE LOCUTOR APLICADO A DURACIONES CORTAS

Ingeniería de Telecomunicación

Rubén Zazo Candil

Julio 2014

ANÁLISIS DE COMPENSACIÓN DE VARIABILIDAD EN RECONOCIMIENTO DE LOCUTOR APLICADO A DURACIONES CORTAS

AUTOR: Rubén Zazo Candil
TUTOR: Javier Gonzalez Dominguez

Área de Tratamiento de Voz y Señales
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Julio 2014

Resumen

Resumen

En este proyecto se estudian, implementan y evalúan sistemas automáticos de reconocimiento de locutor en presencia de locuciones de duración corta. Para llevarlo a cabo se han utilizado y comparado diversas técnicas del estado del arte en reconocimiento de locutor así como su adaptación a locuciones de corta duración.

Como punto de partida del proyecto se ha realizado un estudio de las diferentes técnicas que han ido marcando el estado del arte, destacando las que han conseguido una mejoría notable en evaluaciones promovidas por el *National Institute of Standards and Technology (NIST)* de reconocimiento de locutor durante la última década.

Una vez entendido el estado del arte desde el punto de vista teórico el siguiente paso se define la tarea sobre la que se evaluarán las diferentes técnicas. Históricamente, la tarea principal en evaluaciones *NIST* consiste en entrenar el modelo de locutor con una conversación, de aproximadamente 150 segundos, y realizar la verificación de usuario frente a una locución de la misma duración. En la tarea que se desarrolla durante la realización de este proyecto disponemos de locuciones con una duración mucho más limitada, aproximadamente 10 segundos, provenientes de evaluaciones *NIST* de reconocimiento de locutor.

Para la parte experimental se llevaron a cabo dos fases de experimentos. Durante la primera fase el objetivo ha sido comparar y analizar las diferencias entre dos técnicas del estado del arte basadas en *Factor Analysis (FA)*, *Total Variability (TV) y Probabilistic Linear Discriminant Analysis (PLDA)*, evaluando principalmente el rendimiento de éstas técnicas sobre nuestro entorno experimental que seguirá el protocolo de las evaluaciones *NIST*. En la segunda fase se hace un ajuste de los parámetros de dichas técnicas para comprobar el impacto de los mismos en presencia de duraciones cortas y mejorar el rendimiento de los sistemas con escasez de datos. Para ello evaluamos el sistema en base a dos medidas, la tasa de error y la función de coste que suele emplearse en dicha evaluación, que será detallada en los siguientes capítulos.

Finalmente, se presentan las conclusiones extraídas a lo largo de este trabajo, así como las líneas de trabajo futuro.

Parte del trabajo llevado a cabo durante la ejecución de este Proyecto Final de Carrera ha sido publicado en la conferencia de caracter internacional IberSpeech 2012 [1]:

Javier Gonzalez-Dominguez, Ruben Zazo, and Joaquin Gonzalez-Rodriguez. "On the use of total variability and probabilistic linear discriminant analysis for speaker verification on short utterances".

Palabras Clave

Sistema biométrico, reconocimiento de patrones, reconocimiento automático de locutor, *NIST*, *Factor Analysis*, duraciones cortas, *i-vector*, *PLDA*.

Abstract

This project is focused on automatic speaker verification (SV) systems dealing with short duration utterances (∼10s). Despite the enormous advances in the field, the broad use of SV in real scenarios remains a challenge mostly due to two factors. First, the session variability; that is, the set of difference among utterances belonging to the same speaker. Second, the system performance degradation when dealing with short duration utterances.

As an starting point of this project, an exhaustive study of the state-of-the-art speaker verification techniques has been conducted. This, with special focus on those methods, which achieved outstanding results and open the door to better SV systems. In that sense, we put particular emphasis on the recent methods based on Factor Analysis (FA) namely, Total Variability (TV) and Probabilistic Linear Discriminant Analysis (PLDA). Those methods have become the state of the art in the field due to their ability of mitigating the session variability problem

In order to assess the behaviour of those systems, we use the data and follow the protocol defined by the US National Institute of Standards and Technology (NIST) in its Speaker Recognition Evaluation series (SRE). Particularly, we follow the SRE2010 protocol, but adapted to the short durations problems. Thus, instead of using 150s duration utterances as defined in the core task of SRE2010, we experiment with 10s duration utterance in both training and testing.

The experiments conducted can be divided in two phases. During the first phase we study, compare and evaluate the use of TV and PLDA as effective methods to perform SV. Second phase is then devoted to adapt those methods to the short duration scenarios. We analyse in this point the effect and importance of the multiple parameters of the systems when facing to limited data for both training and testing. Conclusions and future lines of this work are then presented.

Part of this work has been published on the international conference IberSpeech 2012 [1]:

Javier Gonzalez-Dominguez, Ruben Zazo, and Joaquin Gonzalez-Rodriguez. "On the use of total variability and probabilistic linear discriminant analysis for speaker verification on short utterances".

Key words

Biometric system, pattern recognition, automatic speaker recognition, NIST, factor analysis, short durations, ivector, PLDA.

Agradecimientos

Quiero agradecer en primer lugar a mi ponente, Joaquín González, la oportunidad que me ha brindado de colaborar con el ATVS y su apoyo desde mis inicios en el universo de los papers.

Desde que empecé a frecuentar el C-109, además de iniciarme en el mundo de la investigación, he conocido a grandísimas personas con quién he compartido mucho más que trabajo. En especial quiero agradecer a mi tutor, Javier González, su apoyo y guía desde mi entrada a la investigación. Gracias por las horas de pizarra y rotulador que convirtieron un mundo oscuro en algo apasionante, siempre dispuesto a remangarse y zambullirse en código codo con codo. Sería ingrato no mencionar al resto de compañeros que tanto han ayudado con Fer, JaviF, Alicia, Álvaro, Eslava, Sandra, Dani, Doroteo, Iván, Vera, Pedro, Sara, Ester y Marta a la cabeza, habéis hecho de un laboratorio un sitio al que acudir con una sonrisa cada mañana.

Con este documento pongo punto y final a una etapa de 5 años que ha supuesto mucho más que formación en mi vida. Durante estos años he conocido grandes personas, y aunque son demasiadas para nombrarlas una a una me gustaría destacar a Borja y Pablo, por lo que son, y por supuesto a Tolosana, dueño indiscutible de al menos la mitad de mi título.

De este último año de carrera me vienen deudas en muchos idiomas que resumiré con un insuficiente gracias a mi equipo de Otaniemi y sobretodo a mi familia de Timpurinkuja y vecinos cercanos, siempre preparados para presentarse a cualquier chilling, fiesta o café que se presentase.

Cuando de agradecimientos de toda una vida se trata quiero mencionar a la gente que ha hecho de mi lo que soy, a Ainhoa, Edu, Aitor, María, Tony, Magas, Guille, Jose y compañía, representando al Chan-Chan y a Álvaro, Alberto, Iván, Pablo, Jenny, Belen, Blanca, Coral, Hugo, David y demás, haciendo lo propio con El Cerrito. Porque simplemente soy trocitos de ellos.

Por último no puedo no dedicar este proyecto que pone el broche final a unos cuantos años a mi familia, destacando a mis padres y a mi hermana. Me habéis proporcionado la oportunidad, la confianza y todo lo que he necesitado en cualquier proyecto que me haya planteado apoyando cada una de mis decisiones, no sería lo mismo si no pudiese compartirlo con vosotros.

Rubén Zazo Candil
Julio 2014

ésas... ¡no volverán!
Gustavo Adolfo Bécquer

Índice general

Índice de figuras

Índice de tablas

Preámbulo

Notaciones utilizadas

Los *términos* pertenecientes al ámbito de las disciplinas tratadas han sido por norma general traducidos al castellano, salvo en dos casos: cuando el término en inglés es de uso común en la literatura en castellano y cuando, aun no estando extendido, su traducción no resulta directa.

El significado de los *acrónimos* aparecidos en el trabajo se incluye en el Glosario. Se han utilizado las siguientes *abreviaturas*: Cap. (capítulo), Sec. (sección), Fig. (figura) y Ecu. (Ecuación).

Para la *bibliografía* se ha optado por la notación **unsrt** de LaTeX por considerarla la más adecuada para este tipo de documentos por legibilidad y elegancia. Ejemplos de esta notación son [2] para referencias entre corchetes y para referencias *inline*.

Herramientas utilizadas

El presente trabajo ha sido redactado por el autor usando LaTeX. El formato del texto es Computer Roman Modern a tamaño 11pt.

Nota sobre el copyright ®

Los derechos de cualquier marca comercial o registrada mencionada en el presente documento son propiedad de sus respectivos titulares.

1

Introducción.

1.1. Motivación del proyecto.

Las necesidades de mejorar la seguridad en diferentes contextos de la sociedad actual dan lugar a un nuevo concepto de identidad de cada individuo. Los sistemas tradicionales de identificación tienen carencias que contrastan con dichas necesidades: están basados o requieren información adicional que el sujeto **recuerda** como claves de acceso o que el sujeto **posee** como documentos de identidad o tarjetas de acceso que pueden ser perdidas, robadas e incluso falsificadas fácilmente provocando brechas de seguridad. Este hecho ha motivado el desarrollo de varias técnicas de reconocimiento automático basadas en rasgos biométricos, que permiten la identificación de individuos de forma segura y rápida en función de lo que el sujeto **es**, puesto que no se precisa mayor información que la contenida en dichos rasgos inherentes a las personas como la voz, la cara, el iris, la huella, las manos, etc. [5].

El reconocimiento de personas a través de la voz o reconocimiento de locutor es una de las técnicas de reconocimiento biométrico más extendidas en este momento debido: a que es una técnica no invasiva y cuya adquisición no supone grandes costes; a la facilidad de ser aplicada en entornos móviles sin costes adicionales; y, especialmente, a la fiabilidad y robustez de este rasgo, así como a la cantidad de información que contiene la voz como rasgo biométrico: identidad, idioma, estado de ánimo, género, etc. [6].

En el área de reconocimiento de locutor la tarea gira en torno a la extracción de la identidad del locutor a partir de una o varias locuciones. Además esta tarea se divide a su vez en dos grandes áreas:

1. Dependiente de texto, cuando la tarea de extracción de la identidad se realiza conociendo la información lingüística que el locutor ha pronunciado.

2. Independiente de texto, donde el mensaje pronunciado por el locutor es libre y no fijado.

El presente proyecto se basa en el área de reconocimiento de locutor independiente de texto.

La señal de voz contiene información de identidad en distintos niveles que extraemos en forma de características y que pueden clasificarse en los siguientes niveles: nivel lingüístico, nivel fonético, nivel prosódico y nivel acústico, cuyos detalles se explicarán más adelante. Actualmente los sistemas que obtienen una mejor precisión y rendimiento se basan en el nivel acústico/espectral, utilizando diferentes técnicas de modelado: Máquinas de soporte de vectores *(SVM)*, vectores de identidad *(i-vectors)* y *Probabilistic Linear Discriminant Analysis (PLDA)*.

A pesar de los continuos avances en el desarrollo de la tecnología del reconocimiento de locutor su implantación en entornos reales sigue siendo limitada por diferentes aspectos. Si bien el reconocimiento de locutor puede alcanzar unas tasas de error muy bajas en condiciones de laboratorio éste rendimiento se ve rápidamente decrementado debido a la variabilidad de sesión, donde destaca la variabilidad introducida debido a los diferentes canales utilizados (telefónico, microfónico, etc.). Por otro lado los sistemas típicos de reconocimiento de locutor se han basado en locuciones de larga duración (superiores a 200 segundos) pero en entornos reales donde la duración se ve sesgada a fragmentos inferiores a 10 segundos los sistemas se degradan rápidamente.

Con el objetivo de paliar la variabilidad de sesión haciendo especial hincapié en la variabilidad de canal han surgido varias técnicas generativas basadas en *Factor Analysis (FA)* que modelan en distintos subespacios la variabilidad de sesión y la variabilidad de locutor obteniendo grandes avances en el rendimiento de los sistemas. Este proyecto se centra en el estudio de los sistemas de nivel acústico/espectral basados en estas técnicas y su adaptación a entornos donde las locuciones con que se cuenta son de una duración significativamente más corta. Por este motivo la tarea que se evaluará en este trabajo consiste en el reconocimiento de locutor con locuciones de duración en torno a 10 segundos tanto para el entrenamiento de los modelos como para la realización de la verificación.

1.2. Objetivos y enfoque.

El proyecto enfoca los siguientes cuatro objetivos:

1. Estudio de las diferentes técnicas del estado del arte que han producido los mejores resultados en la última década en reconocimiento de locutor.

2. Prueba de sistemas con distintas técnicas del estado del arte en reconocimiento de locutor.

3. Evaluación de los sistemas mediante una serie de experimentos utilizando diferentes bases de datos utilizadas en evaluaciones NIST de reconocimiento de locutor, debido a que es el protocolo de evaluación más utilizado en publicaciones de referencia. Se pretende además encontrar la mejor configuración de los sistemas analizados para enfrentarse a locuciones de corta duración.

4. Entender en profundidad el desafío de las duraciones cortas y analizar la problemática a la que nos estamos enfrentando. La tarea contenida en este proyecto es clave en el reconocimiento de locutor en la actualidad por lo que se pretenden reconocer las limitaciones y posibles mejoras para poder continuar la investigación iniciada en este proyecto.

1.3. Metodología y plan de trabajo.

Para el correcto desarrollo y consecución de los objetivos marcados en el presente Proyecto Fin de Carrera se ha seguido el plan de trabajo que se muestra en el diagrama temporal de la Figura 1.1 y que se detalla a continuación.

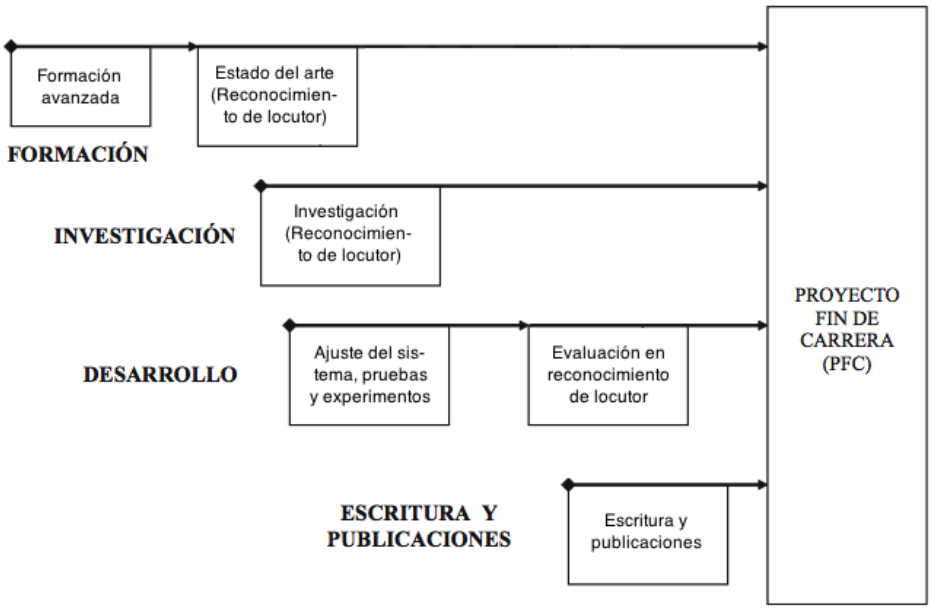

Figura 1.1: Diagrama del plan de trabajo seguido.

- **Estudio del estado del arte.** Todo inicio fundamental de un proyecto pasa por una etapa de formación en la que se obtienen los conocimientos necesarios para su desarrollo. Para este proyecto en concreto se han estudiado en primer lugar las características básicas del reconocimiento biométrico seguido del estado del arte en reconocimiento de locutor utilizando la bibliografía del estado del arte adecuada para sistemas basados en FA.

- **Estudio del software** En la segunda fase del proyecto el alumno se ha familiarizado con el software desarrollado por el grupo ATVS y las herramientas *Matlab* necesarias para el desarrollo de experimentos.

- **Experimentos y desarrollo de software** Posteriormente, se han realizado experimentos siguiendo los protocolos de las evaluaciones *Speaker Recognition Evaluation (SRE)* organizadas por el NIST con el objetivo de analizar diferentes técnicas así como su adaptación a entornos de corta duración. Además se ha analizado de qué forma afectan los distintos parámetros de las técnicas evaluadas a los resultados sobre una tarea cuyas locuciones son de corta duración. Todo el código desarrollado se ha organizado para su uso posterior.

- **Evaluación de resultados y elaboración de la memoria.** Se ha realizado un análisis de los resultados obtenidos en las pruebas llevadas a cabo así como una comparativa entre las diferentes técnicas utilizadas según los mismos. Estos análisis, junto con la revisión del estado del arte y un estudio completo del proyecto llevado a cabo, servirán para elaborar la memoria que pone punto y final al presente Proyecto Fin de Carrera.

1.4. Organización de la memoria.

El presente trabajo se estructura en cinco capítulos:

- **Capítulo 1: Introducción.** Este capítulo presenta la motivación para el desarrollo de este proyecto, así como, los objetivos a cumplir durante la ejecución del proyecto, la estructura de este documento y las contribuciones del mismo.
- **Capítulo 2: Estado del arte en reconocimiento biométrico de locutor.** En este capítulo se presenta el estado del arte actual en reconocimiento de locutor independiente de texto. En primer lugar se describen los distintos modos de operación de los sistemas biométricos y se destacan las características de la señal de voz como rasgo biométrico. Posteriormente se describe la extracción de características así como los tipos de parametrización utilizadas en el reconocimiento de locutor. A continuación se presentarán las herramientas y métodos que se utilicen para medir de forma cuantitativa el rendimiento de los sistemas de reconocimiento de locutor. Por último se describen las diferentes técnicas empleadas en el estado del arte del reconocimiento de locutor independiente de texto. Se hará un especial hincapié en aquellas que modelan las características acústicas de la señal de voz mediante modelos de mezclas de Gaussianas ya que los sistemas objeto de estudio en este proyecto se basan en éstas.
- **Capítulo 3: Marco Experimental.** En este capítulo se presenta en primer lugar las bases de datos que se utilizarán durante el proyecto así como el protocolo de evaluación que se seguirá en la ejecución de los experimentos. En segundo lugar, se describirá de forma más concreta los sistemas que se utilizarán en la fase experimental desde el pre-procesado de la señal hasta la normalización de puntuaciones pasando por las diferentes técnicas empleadas en el reconocimiento de locutor.
- **Capítulo 4: Experimentos realizados y resultados.** En este capítulo se presentan los resultados obtenidos a lo largo de la evaluación de los sistemas analizados. Asimismo, se detalla la secuencia de experimentos realizados para dar con la configuración de parámetros que mejor se ajusta al problema de las duraciones cortas.
- **Capítulo 5: Conclusiones y trabajo futuro.** En este capítulo se presenta las conclusiones extraídas del proyecto realizado, así como las futuras líneas a seguir en este ámbito.

Al final del presente documento se añaden una serie de apéndices con información adicional para completar la descripción del proyecto.

1.5. Contribuciones.

El presente Proyecto Fin de Carrera ha contribuido con el grupo de reconocimiento biométrico ATVS y la comunidad científica en los siguientes aspectos:

- Se ha realizado un estudio detallado y actualizado del estado del arte de los sistemas de reconocimiento automático de locutor independiente de texto.
- Se ha estudiado el software de reconocimiento de locutor del ATVS actualizando y estructurando el mismo.

- Se ha realizado un análisis de las diferentes técnicas de compensación de la variabilidad de sesión.
- Se ha realizado un análisis del impacto de los parámetros del algoritmo a los resultados de la tarea comprendida.
- Se ha realizado un análisis y extracción de conclusiones de la problemática de las duraciones cortas.
- Los trabajos realizados durante la ejecución de este PFC han derivado además en la publicación de un artículo aceptado en un congreso científico de carácter internacional.

2

Estado del arte.

2.1. Introducción

En este capítulo se presenta el estado del arte en los sistemas de reconocimiento biométrico, dedicando mayor atención a aquellos basados en la voz. En la sección 2.2 se repasan las características comunes a todos los sistemas biométricos. En la sección 2.3 se analiza el bloque de extracción de características de los sistemas de reconocimiento de locutor. En la sección 2.4 se presentan las herramientas con que se evalúa el sistema analizado y, por último, en la sección 2.5 se estudian las técnicas de reconocimiento de locutor utilizadas en este trabajo así como aquellas anteriores que han supuesto un hito importante para el desarrollo de ellas.

2.2. Sistemas de reconocimiento biométrico

El ser humano se desarrolla en una sociedad cada vez más interconectada y global, en la que el reconocimiento de los individuos representa, cada vez más, un papel fundamental. Desde el control de fronteras hasta el pago mediante dispositivos electrónicos, la identificación fiable del individuo es un requisito imprescindible. Las dos modalidades más usadas en la actualidad se basan en algo que el individuo **sabe** (ej. una contraseña), o en algo que **tiene** (ej. una tarjeta). Un sistema de reconocimiento biométrico se basa en el reconocimiento de patrones, rasgos biométricos extraídos a partir de rasgos intrínsecos de una persona. El reconocimiento biométrico permite que éstas aplicaciones se basen en algo que el individuo **es** (ej. su voz)[4]. Frente a los otros dos tipos de reconocimiento, presenta la ventaja de que un rasgo biométrico no puede ser olvidado, robado o perdido.

2.2.1. Características de los rasgos biométricos

Se pueden utilizar diferentes rasgos para identificar al usuario. Estos se pueden clasificar asimismo en patrones morfológicos o anatómicos (p.e. voz, huella, iris, cara, etc) y patrones de comportamiento (p.e. firma, escritura, forma de andar, etc). La Fig. 2.1 contiene ejemplos de

rasgos biométricos. La conveniencia de un rasgo para una aplicación está determinada en base a una serie de características que todo rasgo debe cumplir en mayor o menor medida [7]:

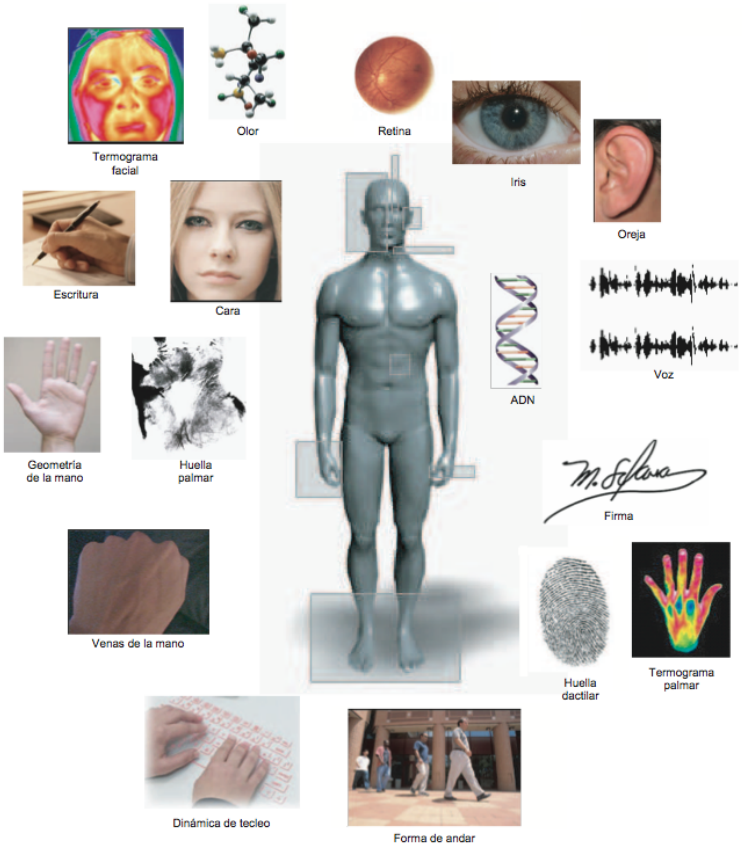

Figura 2.1: Rasgos biométricos humanos.

- **Universalidad:** existencia del rasgo en todos los usuarios.
- **Unicidad:** capacidad discriminativa del rasgo (personas distintas deben poseer rasgos distintos).
- **Permanencia o Estabilidad:** invariabilidad del rasgo en el tiempo.
- **Mensurabilidad o Evaluabilidad:** capacidad para caracterizar el rasgo cuantitativamente, es decir, para ser medido.

En los sistemas de reconocimiento biométrico deben cumplirse también, en mayor o menor medida, los siguientes requisitos:

- **Aceptabilidad:** grado de aceptación personal y social.
- **Rendimiento:** precisión y rapidez en la identificación.
- **Seguridad o Evitabilidad:** resistencia a ser eludido o burlado.

Rasgo biométrico	Universalidad	Unicidad	Permanencia	Mensurabilidad	Rendimiento	Aceptabilidad	Evitabilidad
ADN	A	A	A	B	A	B	B
Oreja	M	M	A	M	M	A	M
Cara	A	B	M	A	B	A	A
Termograma facial	A	A	B	A	M	A	B
Venas de la mano	M	M	M	M	M	M	B
Huella dactilar	M	A	A	M	A	M	M
Forma de andar	M	B	B	A	B	A	M
Geometría de la mano	M	M	M	A	M	M	M
Iris	A	A	A	M	A	B	B
Huella palmar	M	A	A	M	A	M	M
Olor	A	A	A	B	B	M	B
Retina	A	A	M	B	A	B	B
Firma	B	B	B	A	B	A	A
Forma de teclear	B	B	B	M	B	M	M
Voz	M	B	B	M	B	A	A
Escritura	B	B	B	A	B	A	A

Tabla 2.1: Comparación cualitativa de las características de los distintos rasgos biométricos (A=Alto, M=Medio, B=Bajo). Fuente: [4].

Desafortunadamente ningún rasgo biométrico cumple todos los atributos con éxito, presentando cada uno ventajas y desventajas. En la tabla 2.1 podemos ver el grado de cumplimiento de las características para distintos rasgos.

En función del tipo de rasgo empleado por el sistema biométrico, existen dos grandes categorías:

- **Biometría estática**: Engloba las medidas de características corporales o físicas del individuo. En este primer grupo se encuentran, por ejemplo, la huella dactilar, el ADN y el iris, entre otras.
- **Biometría dinámica**: Engloba las medidas de características conductuales del individuo. En este segundo grupo se encuentran, por ejemplo, la firma manuscrita, la forma de andar y la dinámica de tecleo, entre otras.

En el caso de la voz, los sistemas de reconocimiento hacen uso principalmente de características estáticas (contenido espectral de la voz, o características acústicas). Sin embargo, también hacen uso de la evolución temporal de estas características, e incluso otras determinadas por la forma de hablar, como cambios de entonación o uso de las pausas; dependientes del comportamiento del individuo. Podríamos por tanto decir que los sistemas biométricos que utilizan la voz como rasgo engloban, al mismo tiempo, biometría estática y biometría dinámica.

En la tabla 2.1 podemos ver el grado de cumplimiento de la voz como rasgo biométrico. Cabe destacar su alta **aceptabilidad**, debido a que no requiere un método de adquisición intrusivo.

Aunque no se trate de uno de los rasgos más distintivos como puede ser el ADN o el iris, proporciona un alto grado de seguridad en aplicaciones de verificación y presenta varias ventajas en sistemas de identificación menos exigentes. Como principal desventaja debemos destacar la baja permanencia o **estabilidad**, ya que la voz puede variar en función de diversos factores como podrían ser la salud, la edad o el estado de ánimo.

2.2.2. Funcionamiento de los sistemas biométricos

Los sistemas de reconocimiento automático basan su funcionamiento en el reconocimiento de patrones. Poseen una estructura funcional común formada por varias fases cuya forma de proceder depende de la naturaleza del patrón a reconocer como podemos ver en la figura 2.2. Los módulos o etapas con línea continua son las entidades hardware o software básicas del sistema, y los módulos marcados con línea discontinua se corresponden con etapas de procesado opcionales, dependiendo del tipo de sistema, que puede ser *identificación* o *verificación* como veremos en la Sec. 2.2.3. La línea más oscura marca la frontera entre la interfaz de usuario y el sistema.

El usuario tiene acceso al **sensor** que captura el rasgo biométrico y, en caso de un sistema de *verificación*, el módulo **identidad** permitirá al usuario identificarse mediante un PIN o tarjeta de identificación.

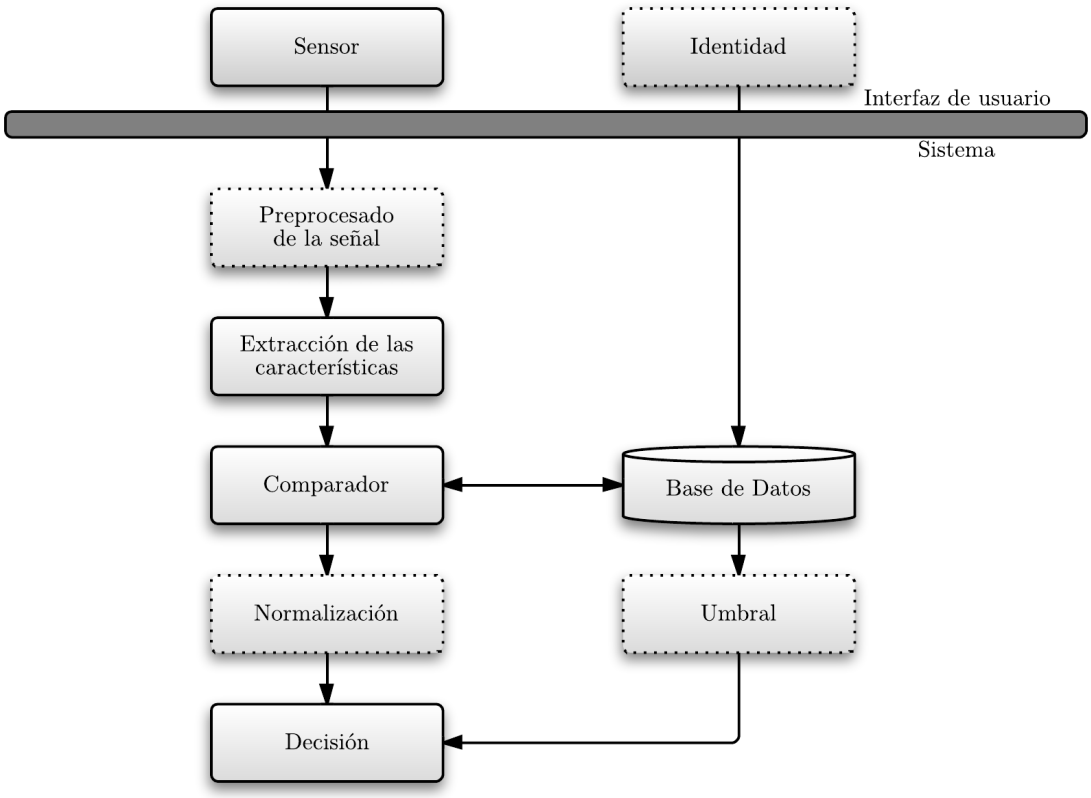

Figura 2.2: Esquema de funcionamiento de un sistema de reconocimiento biométrico.

A continuación se presenta la etapa de **pre-procesado**. Esta etapa se considera opcional porque su existencia dependerá de la degradación de la señal capturada así como de las carac-

terísticas del rasgo a utilizar. En este módulo la señal es pre-procesada para compensar posibles degradaciones o facilitar la **extracción de características**. Posteriormente, en el siguiente módulo se extraerán las características de la señal capturada. Estas características pretenden ser las más distintivas de la identidad para ese rasgo biométrico.

El siguiente bloque funcional se corresponde con la etapa de **comparación**. En este módulo las características previamente extraídas son comparadas con los patrones de referencia que provienen de la base de datos. Para cada enfrentamiento o comparación (***trial***) realizada el sistema obtiene un valor de similitud, que también es llamado ***puntuación*** o ***score***, entre las características y el patrón. En la siguiente etapa opcional, esta puntuación obtenida puede ser sometida a una **normalización** que permite la transformación de las puntuaciones a un rango de valores en que es más fácil de identificar si dicho ***score*** pertenece a un impostor o a un usuario genuino.

Por último, el sistema comparará el ***score*** con un **umbral** para tomar una decisión: en caso de un sistema de ***verificación*** el sistema deberá decidir si las características obtenidas en el sensor pertenecen a la misma persona que las características previamente registradas de dicho individuo en la base de datos; para un sistema de ***identificación*** el sistema deberá decidir si las características corresponden a alguna de las identidades registradas previamente en el sistema. Algunas aplicaciones como la biometría forense carecen de esta etapa final, siendo la salida del mismo un ratio de similitud entre las características obtenidas y las registradas previamente, que será después tenido en cuenta junto con otras evidencias para tomar la decisión final.

2.2.3. Modos de operación

En los sistemas de reconocimiento biométrico pueden distinguirse tres modos de operación [4]. El primer modo, modo ***registro*** se considera una fase previa al funcionamiento del sistema. Tras este modo podemos reconocer el modo ***identificación*** y el modo ***verificación***. En las figuras 2.3, 2.5 y 2.4 podemos ver un esquema específico para la señal de voz de cada uno de ellos.

Durante el modo **registro**, ver Fig. 2.3, los usuarios son dados de alta en el sistema. Los rasgos biométricos del usuario son adquiridos por el sensor. Tras una etapa opcional de pre-procesado procedemos a la extracción de las características identificativas del usuario. En algunos sistemas estas características conforman el patrón de referencia, en otros sistemas estas características son utilizadas para *entrenar* un modelo estocástico de la identidad a reconocer, por lo que este modo es también conocido como **fase de entrenamiento**.

Una vez los usuarios han sido registrados en la base de datos el sistema puede entrar en funcionamiento en dos modos diferentes:

En el modo de **verificación**, ver Fig. 2.4, el usuario proporciona la identidad mediante una tarjeta identificativa o un número PIN. A continuación, el sistema obtiene el rasgo biométrico del usuario y realiza una comparación uno-a-uno entre este y el patrón almacenado en la base de datos del usuario obteniendo una identificación positiva, en caso de que el sistema detecte que los rasgos coinciden con una similitud suficiente o negativa, en caso contrario.

En el modo de **identificación**, ver Fig. 2.5, el sistema realiza una comparación del patrón obtenido mediante el sensor con todos los patrones almacenados en la base de datos. Es, por tanto, una comparación uno-a-muchos en la que el sistema no necesita ningún tipo de identidad. El sistema tratará de encontrar la identidad a la que pertenece el rasgo extraído, en caso de que ésta se encuentre en la base de datos. Este modo es más demandante computacionalmente

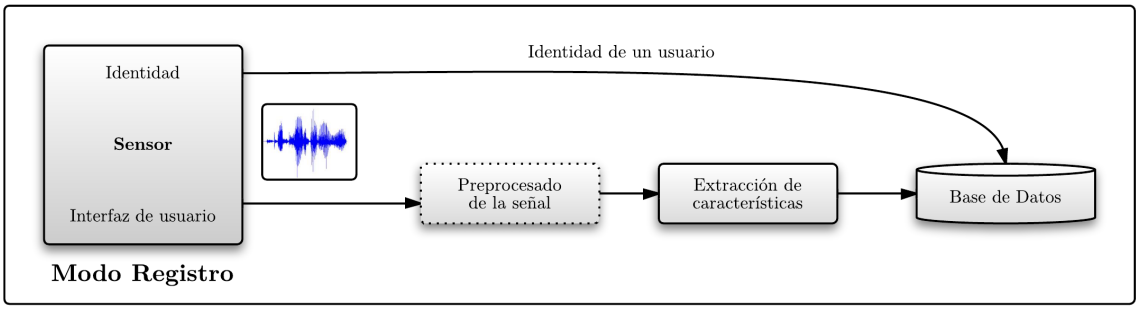

Figura 2.3: Esquema de funcionamiento de un sistema de reconocimiento biométrico en modo registro.

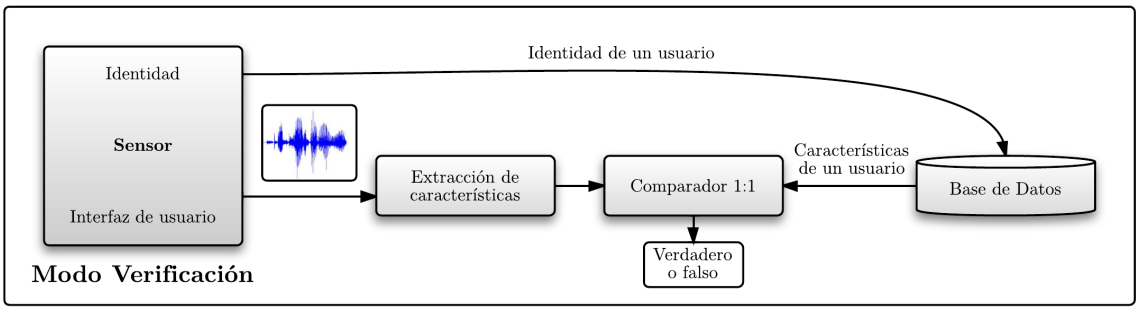

Figura 2.4: Esquema de funcionamiento de un sistema de reconocimiento biométrico en modo verificación.

y esta demanda aumenta de forma lineal con el número de entradas de la base de datos. La salida de este modo será la identidad del usuario al que pertenece el rasgo recibido o un mensaje indicando que dicha identidad no se encuentra en la base de datos.

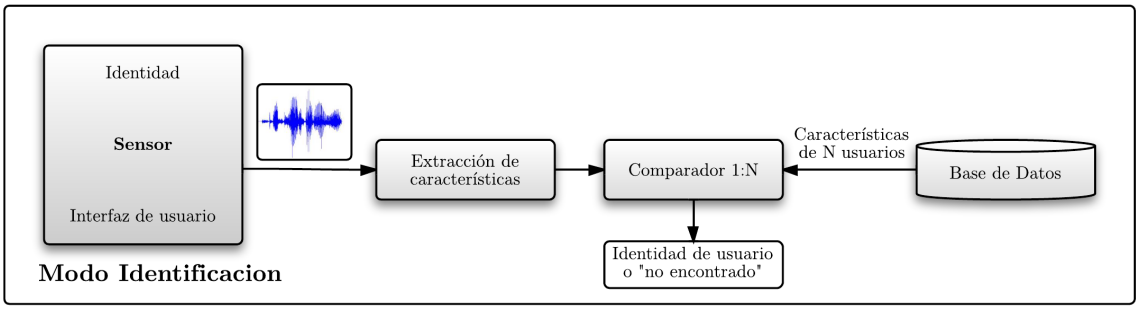

Figura 2.5: Esquema de funcionamiento de un sistema de reconocimiento biométrico en modo identificación.

2.3. Extracción de características a partir de la voz.

El primer paso para utilizar la voz como rasgo biométrico consiste en la extracción de las características identificativas, también conocido como *parametrización*. Idealmente estas características deberían cumplir con los siguientes requisitos [8] [9]:

- Baja variabilidad para un mismo locutor (*intra-locutor*) combinada con una alta variabilidad entre distintos locutores (*inter-locutor*).
- Dificultad para ser suplantadas o imitadas.
- Ocurrir de forma continuada y frecuente durante cualquier locución.
- Invariabilidad frente a la salud, estado de ánimo u otras variaciones a largo plazo en la voz.
- Robustez frente a ruido y distorsión.
- Facilidad de ser medidas a partir de la señal de voz.

La información de locutor que podemos extraer de la señal de voz puede dividirse en cuatro grandes grupos o niveles que, desde el más alto hasta al más bajo, son: el nivel *lingüístico*, el nivel *fonético*, el nivel *prosódico* y el nivel *acústico*. Los niveles más altos se corresponden con características aprendidas mientras que los niveles más bajos dependen más de cuestiones fisiológicas.

El **nivel lingüístico** se corresponde con la información de más alto nivel y recoge las características idiolectales. Este nivel describe la manera en que el locutor hace uso del sistema lingüístico y se ve afectado por aspectos tales como la cultura, educación, origen o condiciones sociológicas del locutor. Los sistemas que utilizan este nivel se concentran en sucesos como la frecuencia o las secuencias de palabras.

En el **nivel fonético** se encuentran las características relacionadas con los fonemas y sus secuencias, también conocido como características *fonotácticas* [10]. Cada locutor utiliza estas unidades léxicas con un patrón distintivo que los sistemas intentan modelar para la fase de reconocimiento.

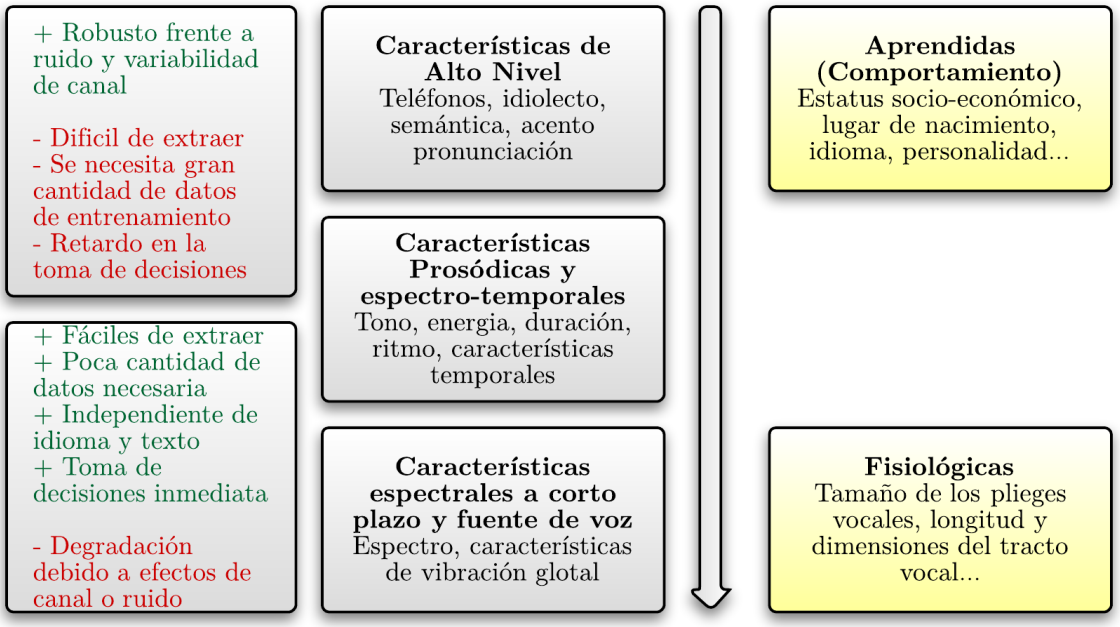

Figura 2.6: Niveles de información en la señal de voz (adaptado de [2]).

La combinación de energía, tono y duración de los fonemas conforma el **nivel prosódico**. La prosodia tiene elementos comunes a todos los locutores, es lo que nos permite distinguir entre un mensaje interrogativo de otro imperativo. Aun así, cada locutor emplea dichos elementos de forma distinta pudiéndose implementar un sistema de reconocimiento de locutor en base a estas características.

Por último el nivel más bajo de la clasificación se corresponde con el **nivel acústico**. Este nivel se centra en las características espectrales a corto plazo de la señal de su voz y cómo varían a lo largo del tiempo. Estas características dependen directamente de la configuración fisiológica del aparato fonatorio, de la forma en que se produce cada sonido y de las acciones articulatorias de cada individuo.

Históricamente el *nivel acústico* ha sido el más utilizado debido a que es, con diferencia, el nivel que permite la mayor precisión en el reconocimiento de locutor. Sin embargo existen muchas líneas de investigación actuales que pretenden combinar información de diferentes niveles para mejorar el rendimiento [11]. Para este proyecto se ha utilizado un sistema basado únicamente en el nivel acústico. Una de las ventajas con las que cuentan los sistemas que trabajan con información de este nivel es que tan sólo necesitan la voz para su funcionamiento. Los sistemas con niveles más elevados de información necesitan, además, contar con las transcripciones fonéticas o de palabra de las locuciones, por lo que es necesaria la intervención de una persona o apoyarse en sistemas de reconocimiento de habla (*Automátic Speech Recognition, ASR*), que obtienen estas etiquetas de forma automática pero introducen otra fuente de posibles errores al sistema.

En la figura 2.6 podemos ver los principales niveles de información de identidad que podemos encontrar en la voz junto a sus principales ventajas y desventajas.

2.3.1. Coeficientes MFCC (Mel-Frecuency Cepstral Coefficients)

Los coeficientes MFCC propuestos en [12] se introdujeron originalmente para el reconocimiento de habla y fueron posteriormente adaptados para sistemas automáticos de reconocimiento de locutor de nivel espectral. Actualmente es la *parametrización* más extendida para dichos sistemas. Este método utiliza un banco de filtros *mel* que pretende imitar el comportamiento del oído humano dando mayor importancia a las frecuencias bajas frente a las altas. Su obtención requiere una serie de etapas que describiremos a continuación.

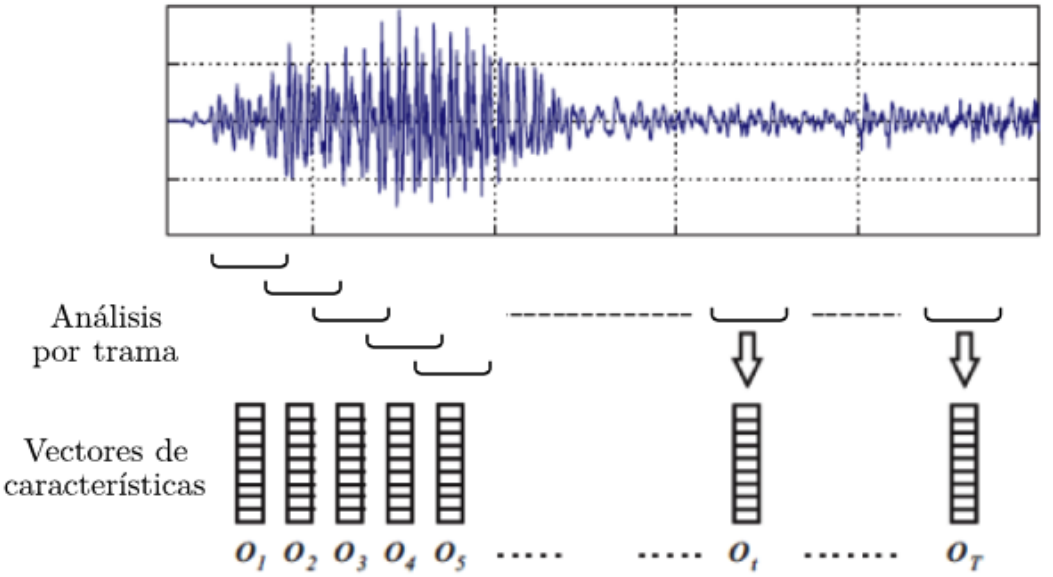

Figura 2.7: Segmentado de la señal de voz en tramas para la obtención de las características.

En primer lugar se lleva a cabo el proceso conocido como ***enventanado***. La señal de voz es dividida en segmentos de 20 milisegundos de duración con un solape del cincuenta por ciento que serán después analizados uno a uno, como podemos ver en la figura 2.7. Esta segmentación se lleva a cabo con el objetivo de poder analizar cada uno de los segmentos individualmente considerando la señal contenida en ellos estacionaria. Estos segmentos se filtran con una ventana de tipo *Hamming* procurando perder la menor información posible. Esta ventana atenúa la señal en los extremos de la ventana evitando que aparezcan componentes de alta frecuencia, véase la figura 2.8.

A continuación puede aplicarse un preénfasis a las tramas que resalta las altas frecuencias del espectro. Tras este paso opcional se procede al cálculo del análisis espectral a través de la **transformada discreta de fourier (*DFT*)** [13] o en su defecto a través de la transformada rápida de fourier (*Fast Fourier Transform FFT*), que es la que se utiliza habitualmente debido a su eficiencia.

El siguiente paso consiste en el análisis del espectro de potencia mediante la utilización de los filtros mel. En primer lugar se aplica una transformación según la fórmula

$$f_m = 2595 * log(1 + f/700) \tag{2.1}$$

donde f es la frecuencia lineal. Tras esta transformación el banco de filtros *mel*, véase la figura 2.9, es aplicado de forma que obtenemos un vector por cada uno de los filtros. A partir de las

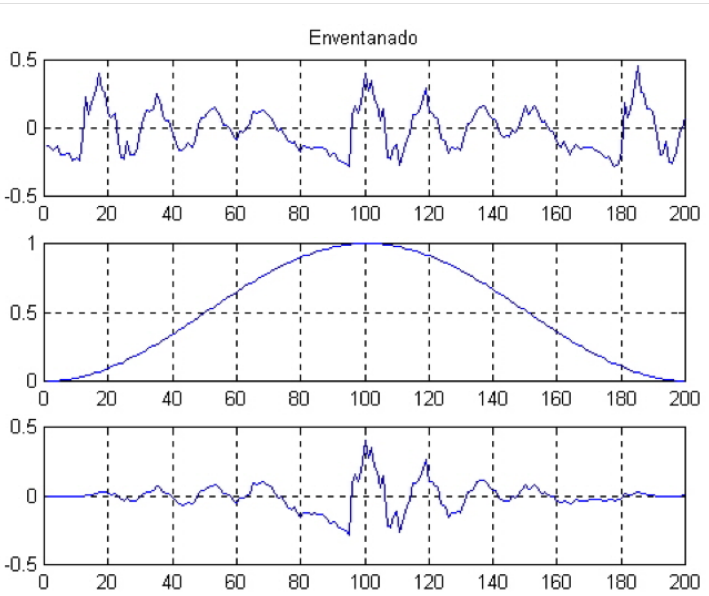

Figura 2.8: Enventanado de la señal con ventana tipo Hamming [3].

salidas de dichos filtros, denotadas mediante Y(m), m = 1, ..., M, los coeficientes *MFCC* son obtenidos siguiendo la siguiente ecuación

$$C_n = \sum_{m=1}^{M} (\ln(Y(m)) \cos(\frac{\pi * n}{M})(m - 1/2)) \tag{2.2}$$

donde n es el índice del coeficiente cepstral. El vector final de características se forma con los 12 o 20 primeros coeficientes C_n.

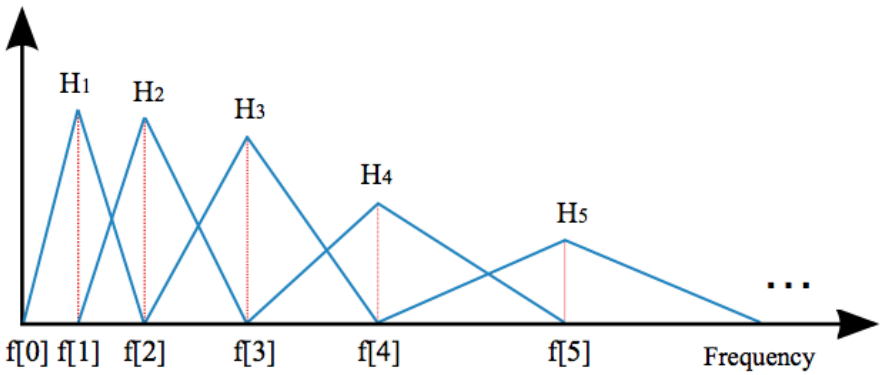

Figura 2.9: Banco de filtros Mel utilizado para la extracción de los coeficientes *MFCC*.

A partir de los coeficientes *cepstral* también es posible incorporar información dinámica en los vectores, de forma que podamos saber cómo varían en el tiempo. Eso se consigue con los coeficientes δ y $\delta\delta$, que son aproximaciones polinomiales de las derivadas de primer y segundo orden y dependen de la velocidad y aceleración con las que varían los coeficientes *cepstral*.

Los coeficientes *MFCC* son ortogonales. Esto permite trabajar con matrices de covarianza diagonales debido a la independencia entre las distintas dimensiones. Por otra parte, la influencia del canal en el dominio *cepstral* se convierte en una componente aditiva, de forma que con técnicas como la normalización con respecto a la media central (*Central Mean Normalization, CMN*), o el filtrado *RASTA (RelAtiveSpecTrAl)* es posible reducir dicha componente. Ambas técnicas serán detalladas en la sección 2.3.2.

2.3.2. Normalización de características

La señal de voz que se utiliza para el reconocimiento de locutor puede tener ruido o distorsiones provenientes del canal u otras fuentes. Una de las formas de paliar esta variabilidad consiste en utilizar técnicas supresoras de ruido para mejorar la calidad de esta señal, sin embargo, este pre-procesado implica un aumento de la carga computacional del sistema. Para evitar esto en la medida de lo posible, se utilizan **extractores de características robustos** y/o se utilizan **técnicas de compensación sobre características**, la carga computacional se reduce considerablemente debido a que la información de locutor incluida en las características es más compacta y no tenemos que tratar la señal completa. En esta sección se presentarán distintas técnicas que se utilizan para compensar efectos perturbadores en la señal sobre las características.

2.3.2.1. Normalización por media cepstral (CMN, Cepstral Mean Normalization)

La señal de entrada de los sistemas puede modelarse como el producto, en el dominio espectral, de la señal de voz, S(z), y la función de transferencia del canal, G(z). [14]. En el dominio *cepstral*, como se vió en la sección 2.3.1, se utiliza el logaritmo de las componentes espectrales de forma que el efecto del canal se convierte en aditivo.

$$\begin{aligned} T(z) &= S(z) * G(z) \\ DFT^{-1}(\log(|T(z)|) &= DFT^{-1}(\log(|S(z)|) + DFT^{-1}(\log(|G(z)|) \end{aligned} \tag{2.3}$$

Si asumimos que el canal es invariante y que la media de los coeficientes cepstrales de una señal ideal es nula, el canal puede ser estimado como la media temporal de la señal de entrada T(z). Sustrayendo la media temporal de los coeficientes centrales el efecto aditivo de canal será, en cierta medida, compensado mediante la siguiente ecuación.

$$y[n] = t[n] - \frac{1}{N} * \sum_{n=1}^{N} t[n] \tag{2.4}$$

donde $t[n]$ representa la señal de entrada en el instante n (afectada por el canal) e $y[n]$ representa el vector de características compensado mediante *CMN*.

2.3.2.2. Filtrado RASTA (RelAtiveSpecTrAl)

Debido a las características del aparato fonador la voz tiene un rango de velocidad de variación determinado. El filtrado *RASTA* [15] consiste en un **filtrado paso banda** sobre la trayectoria temporal de cada característica o dimensión del vector en el dominio *cepstral*. De

esta forma cualquier componente que varíe demasiado rápido o demasiado lento será filtrada, no considerándose habla.

A diferencia de la normalización *CMN* el filtrado *RASTA* es independiente de la señal de entrada por lo que no se considera una normalización adaptativa.

2.3.2.3. Feature warping

Según se desarrolla en [16] la distorsión de canal modifica la distribución real de los coeficientes cepstrales en cortos periodos de tiempo. Mediante esta técnica se pretende modificar la distribución de las características a corto plazo de manera que se ajusten a una **distribución final Gaussiana de media nula y varianza unidad**.

2.3.2.4. Feature mapping

En [17] se propone modelar, mediante *GMM-UBM*, la influencia de los **distintos canales por separado**. De esta forma la compensación que se aplicará dependerá de las características de cada canal de forma independiente.

2.4. Rendimiento de los sistemas de reconocimiento de locutor

Para el diseño e implementación de sistemas de reconocimiento de locutor es necesario contar con herramientas que permitan comprobar las bondades de un sistema, evaluar las mejoras llevadas a cabo y comparar distintos sistemas entre sí. En esta sección se describe el método y herramientas utilizadas para evaluar el sistema de verificación de locutor analizado.

2.4.1. Evaluación del rendimiento

La evaluación del sistema pretende comprobar las capacidades y servir como herramienta para ayudar a la mejora del mismo. Para ello, se utilizan un conjunto de valores y gráficas que forman las pruebas de reconocimiento para evaluar las diferentes técnicas empleadas. Estas pruebas deben realizarse en un entorno que reproduzca en lo posible las condiciones reales de funcionamiento del sistema de forma objetiva, podremos así extrapolar estos valores a un funcionamiento real del sistema.

Existen diferencias entre dos muestras tomadas de un mismo rasgo biométrico debido a imperfecciones en las condiciones de captura de las características, variabilidad en el rasgo biométrico, ruido y/o interacción del usuario con el sensor, entre otros. La respuesta de un comparador de un sistema biométrico consiste en un valor de similitud, una puntuación o *score* que mide, de forma cuantitativa, la similitud entre la entrada y el patrón o rasgo de la base de datos con el que se está comparando. La puntuación devuelta por el comparador será más alta cuanto mayor sea la similitud entre las muestras, es decir, mayor sera el apoyo a la hipótesis de que la identidad de ambos coincida.

La decisión final del sistema se toma históricamente en base a un umbral; aquellos pares de muestras cuyas puntuaciones sean mayores que el umbral serán etiquetados como pertenecientes a la misma persona mientras que aquellos cuya puntuación sea menor que este se asumirán de personas diferentes. Este tipo de funcionamiento provoca que el sistema genere la misma salida cuando dos patrones están justo por encima del umbral (mayor probabilidad de error) y cuando los patrones están muy por encima de este. Existen otros sistemas que en lugar de generar una decisión devuelven una relación de similitud, como veremos en la sección 2.4.2.

En los sistemas de verificación cuya salida es la decisión final se pueden dar dos tipos de errores; el **error de falso rechazo** se produce cuando el sistema rechaza el rasgo biométrico de un usuario genuino, es decir, rechaza a un usuario que reclama su verdadera identidad; el **error de falsa aceptación** se produce cuando el sistema acepta a un usuario que reclama una falsa identidad, es decir, acepta el rasgo de un *impostor*.

En base a estos tipos de error se obtienen dos tasas de error, la **tasa de falsa aceptación** *(FAR, False Acceptance Ratio)* y la **tasa de falso rechazo** *(FRR, False Rejection Ratio)*. Las tasas de error se definen como el ratio entre el número de error producidos y el número de intentos de acceso al sistema.

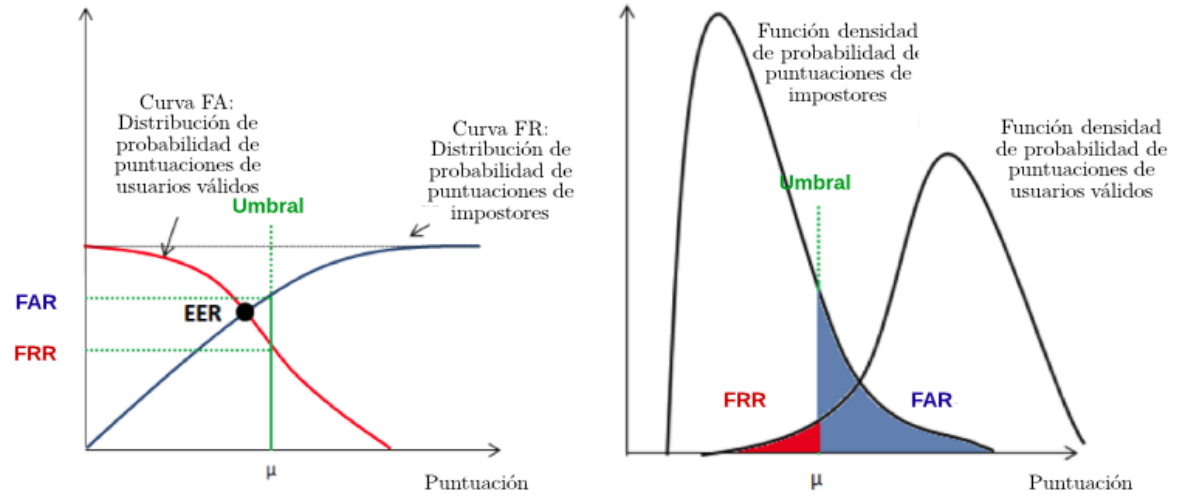

Figura 2.10: Densidades y distribuciones de probabilidad de usuarios e impostores.

En la figura 2.10 podemos ver dos formas de representar las tasas de falso rechazo y falsa aceptación. En la gráfica de la derecha podemos ver la función de distribución de probabilidad de puntuaciones de usuarios válidos e impostores. La tasa de falso rechazo *(FRR, False Rejection Ratio)* puede interpretarse como la probabilidad de que un usuario válido sea rechazado mientras que el valor de la tasa de falsa aceptación *(FAR, False Acceptance Ratio)* puede interpretarse como la probabilidad de que un usuario impostor sea aceptado. El valor de ambas tasas coincide con el valor del eje de ordenadas en el punto en que el umbral coincide las funciones correspondientes en el eje de abscisas. La intersección entre ambas curvas marca el valor de la **tasa de error igual** *(EER, Equal Error Rate)*, que determina el punto donde la tasa de falsa aceptación y la tasa de falso rechazo coinciden. Como podemos ver en la gráfica, un umbral bajo permitiría a un mayor número de impostores ser aceptados como válidos pero disminuiría el número de usuarios genuinos rechazados; para un valor muy alto del mismo el sistema difícilmente aceptará a un impostor pero aumentará el número de usuarios válidos que son rechazados. De esta forma el umbral debe ser fijado manualmente dependiendo de la aplicación concreta definiendo así el punto de trabajo deseado.

En la gráfica de la izquierda de la figura 2.10 se representan, en rojo y azul respectivamente, las funciones de densidad de probabilidad de las puntuaciones obtenidas de usuarios válidos e impostores. Definido el umbral, la tasa de falsa aceptación *(FAR)* coincide con el área bajo la curva de densidad de probabilidad de puntuaciones de impostores (a la derecha del umbral); la tasa de falso rechazo *(FRR)* es igual al área bajo la curva de densidad de probabilidad de puntuaciones de usuarios válidos (a la izquierda del umbral).

2.4.2. Relación de Verosimilitud *(LR, Likelihood Ratio)*

La relación de verosimilitud o *LR* se utiliza en la acústica forense y se define en base a las hipótesis del fiscal y de la defensa. La **Hipótesis del fiscal** (H_p) defiende que la dubitada proviene del sospechoso. La **Hipótesis de la defensa** (H_d) defiende que dicha locución pertenece a otro individuo. La relación de verosimilitud o ratio se define como:

$$LR = \frac{P(E|H_p, I)}{P(E|H_d, I)} \tag{2.5}$$

donde **I** es la información relevante para el caso y **E** es la evidencia disponible, compuesta por una locución de origen desconocido y una locución del sospechoso obtenida en un entorno controlado. Utilizando el teorema de Bayes podemos obtener las probabilidades a posteriori del ratio de hipótesis utilizando el *LR* y la información a priori de la siguiente forma:

$$\frac{P(H_p|E, I)}{P(H_d|E, I)} = LR\frac{P(H_p|I)}{P(H_d|I)} = \frac{P(E|H_p, I)P(H_p|I)}{P(E|H_d, I)P(H_d|I)} \tag{2.6}$$

En un sistema de reconocimiento de locutor se carece de información previa sobre el caso. El objetivo del sistema es aportar un *LR* que indicará qué hipótesis apoya la evidencia (la locución dubitada) y con qué grado de apoyo o verosimilitud. Después esta información se combinará con el resto de la información del caso para obtener un veredicto final. La relación de verosimilitud para un sistema de este tipo se define, dado un locutor **S** y una locución **Y**, como:

$$LR = \frac{P(Y|H_0)}{P(Y|H_1)} \tag{2.7}$$

Donde H_0 es la hipótesis que defiende que la locución **Y** pertenece al locutor **S** y H_1 es la hipótesis que defiende lo contrario. Cuanto más alto sea el valor del *LR* el sistema apoyará la hipótesis H_0 con mayor seguridad.

En la figura 2.11 podemos ver el esquema de funcionamiento de un sistema de verificación de locutor basado en relación de verosimilitud. En el pre-procesado se extraen las características a partir de la señal de voz. Estas características se comparan con el modelo de locutor hipotético **S** y con el modelo de locutor genérico (en el contexto de los sistemas *GMM-UBM* el modelo genérico será el *UBM*) para calcular las funciones de verosimilitud de las hipótesis H_0 y H_1 respectivamente. La relación de verosimilitud se calcula finalmente mediante la resta de ambas funciones de verosimilitud, debido a que el ratio se encuentra dentro de una función logarítmica.

2.4.3. Calibración

Los sistemas de reconocimiento de locutor generan, dada una locución y un modelo de locutor, una puntuación o *score* que será mayor cuanto mayor sea el parecido entre ambos. A diferencia

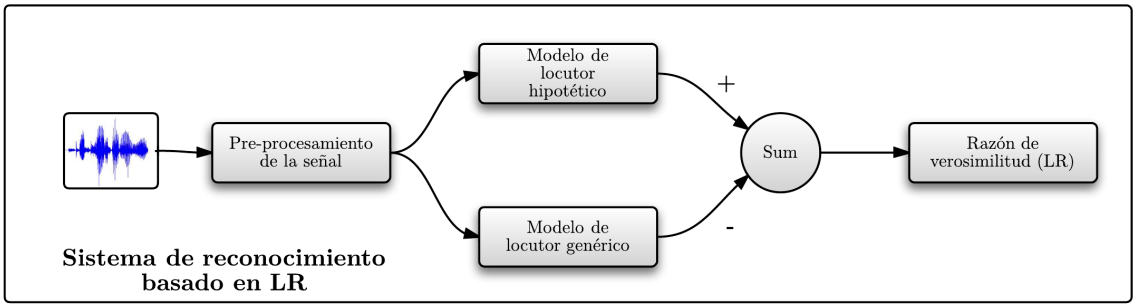

Figura 2.11: Sistema de verificación de locutor basado en relación de verosimilitud.

del *LR* este *score* no tiene una interpretación probabilística y por tanto carece de información sin la existencia de un umbral o rango de valores. En la ecuación 2.6 podemos ver que el *LR* es probabilísticamente interpretable.

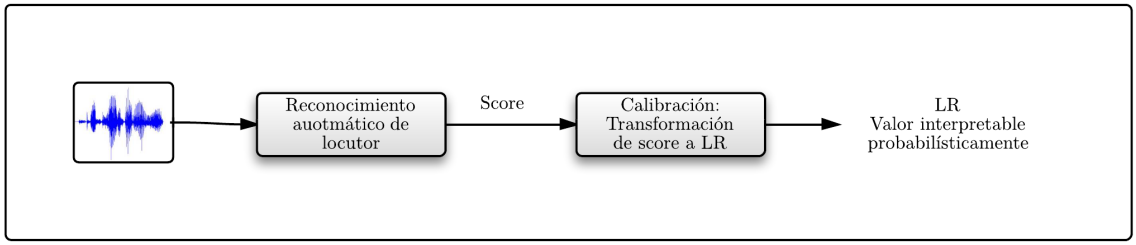

Figura 2.12: Esquema de la transformación de un score a un *LR*.

La función del científico o forense es la de aportar este *LR* para saber el grado de apoyo a una hipótesis, sin valorar la información a priori o el resto de información del caso. La **calibración** es el proceso mediante el cual se calcula el valor del *LR* a partir del *score*, en la figura 2.12 podemos ver este proceso. Existen muchos métodos para la calibración pero el más utilizado consiste en la transformación lineal de scores [18] mediante regresión logística.

2.4.4. Curvas *DET* (*Detection Error Tradeoff*)

Las curvas ***DET*** [19] son la forma visual más utilizada para representar el rendimiento de los sistemas biométricos así como cualquier sistema de clasificación binaria. Principalmente las curvas *DET* representan el error de falso rechazo *FR* frente al error de falsa aceptación *FA*.

En la figura 2.13 podemos ver un ejemplo de curva *DET*. Gracias a este tipo de gráficas resulta sencillo apreciar el compromiso entre los dos tipos de errores (*FA* y *FR*). El valor de la tasa de error igual, *EER*, coincide con la intersección de la curva *DET* y la bisectriz de los ejes de la gráfica.

2.4.5. Función de detección de coste *DCF (Detection Cost Function)*

Para la evaluación de nuestro sistema se utilizará, además de la tasa de error igual, *EER*, una función de coste que sigue el protocolo de la evaluación *NIST* de reconocimiento de locutor

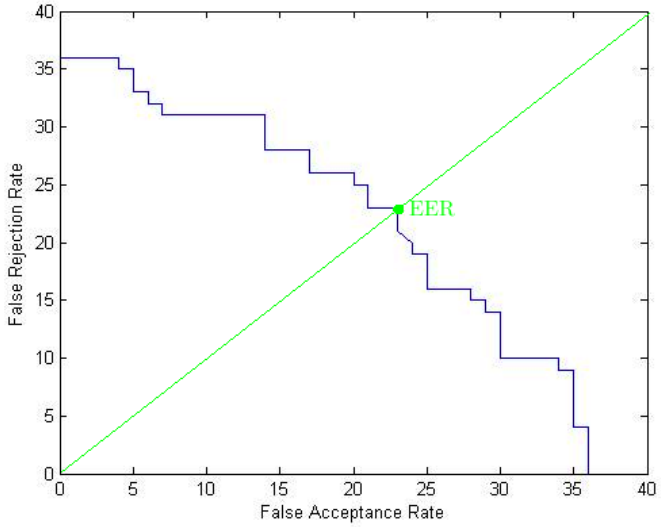

Figura 2.13: Ejemplo de curva *DET*. El sistema muestra los distintos puntos de trabajo en función de los errores *FR* y *FA* posibles.

de 2010 [20].

La función de detección de coste *DCF* se define como:

$$C_{DFC} = C_{FR} * P_{FR} * P_{Tar} + C_{FA} * P_{FA} * (1 - P_{Tar}) \tag{2.8}$$

Donde C_{FR} y C_{FA} son los costes relativos de los errores de detección y P_{Tar} es la probabilidad a priori de que un intento de acceso corresponda a un usuario genuino. Los valores P_{FR} y P_{FA} se obtienen a partir de las funciones de densidad de probabilidad de usuarios e impostores (ver figura 2.10).

En la evaluación de reconocimiento de locutor *NIST SRE 2010* [20] el coste de ambos errores es equivalente mientras que la probabilidad de que un intento de acceso corresponda a un usuario genuino es de 0.001 quedando así definida la función de coste que utilizaremos, conjuntamente con el *EER*, a lo largo de este trabajo.

2.4.6. Normalización de *scores*

En la sección 2.2.2 ya vimos que una de las etapas opcionales de un sistema de reconocimiento biométrico es la normalización de *scores*. El objetivo de la normalización es que los *scores* generados a partir de diferentes locutores estén contenidos en un rango similar. Se pretende así paliar el posible desalineamiento existente entre los *scores* de diferentes locutores para poder utilizar un único umbral de decisión.

La normalización de *scores* más extendida tiene la siguiente forma:

$$\hat{s} = \frac{s - \mu_{imp}}{\sigma_{imp}} \tag{2.9}$$

donde asumimos que la distribución de usuarios impostores es normal con media μ_{imp} y desviación estándar σ_{imp}. El *score* $\hat{s}$ se obtiene normalizando el *score* generado por el sistema utilizando

de los parámetros de dicha distribución. Los parámetros se obtienen a partir de una cohorte de impostores. La idea básica consiste en modificar la distribución de los *scores* de impostor adquiriendo media cero y varianza unidad, de forma que *scores* de diferentes locutores queden alineados.

Existen diferentes técnicas de normalización que dependen del método utilizado para estimar μ_{imp} y σ_{imp}. Las tres técnicas más utilizadas son:

- **z-norm** o *zero normalization* [21]. En *z-norm* una cohorte de puntuaciones de test de impostor se enfrenta con cada uno de los modelos de locutor, de donde se obtienen los parámetros μ_{z-norm} y σ_{z-norm}, dependientes de locutor, que se aplicarán siguiendo la siguiente ecuación:

$$s_{z-norm} = \frac{s_{raw} - \mu_{z-norm}}{\sigma_{z-norm}} \tag{2.10}$$

  donde s_{raw} es el *score* generado originalmente por el sistema. Esta distribución de puntuaciones depende del modelo; al ser aplicada a todos los modelos de forma separada se alinean todas las distribuciones de impostor del sistema.

- **t-norm** o *test normalization* [22]. A diferencia de *z-norm*, en *t-norm* una cohorte de modelos de impostor es enfrentada a cada uno de los test para generar la distribución de *scores* de impostor, obteniendo μ_{t-norm} y σ_{t-norm}. La normalización sigue la fórmula

$$s_{t-norm} = \frac{s_{raw} - \mu_{t-norm}}{\sigma_{t-norm}} \tag{2.11}$$

  Esta distribución de puntuaciones depende de cada fichero de test; al ser aplicada a todos los ficheros de test que actúan como entrada del sistema se alinearán las distribuciones de impostor de todas las locuciones de test.

- **zt-norm** o *zero and test normalization*. En esta técnica *z-norm* y *t-norm* son aplicadas de forma conjunta siguiendo la siguiente fórmula

$$s_{zt-norm} = \frac{\frac{s_{raw} - \mu_{z-norm}}{\sigma_{z-norm}} - \mu_{t-norm}}{\sigma_{t-norm}} \tag{2.12}$$

  donde a los *scores* obtenidos mediante *z-norm* se les aplica después *t-norm*. Para mantener la consistencia se debe aplicar *z-norm* a las distribuciones de *score* de impostor que se utilizarán para realizar la normalización *t-norm*.

2.4.7. Fusión de sistemas

En la sección 2.3 se presentan los diferentes niveles de información presentes en la señal de voz. Partiendo de que los sistemas específicos suelen estar diseñados para explotar un determinado nivel de información se ha demostrado que la fusión de diferentes sistemas tiene un rendimiento mayor que los sistemas individuales [23] [24]. La fusión de sistemas será más efectiva cuanto más incorrelados esté la información que estos utilizan, aunque se ha probado que la fusión de sistemas similares puede mejorar el comportamiento global [23]. Los dos tipos de fusión más extendidas son:

- **Fusión basada en reglas fijas**, también conocida como fusión a nivel de *scores*. Consiste en la combinación de las puntuaciones mediante un operador simple como puede ser la suma, el máximo o el mínimo entre otros. La ventaja de este método es que permite combinar

sistemas con arquitecturas, modelos o características muy diferentes de una forma sencilla. Para poder realizar esta combinación se requiere que las puntuaciones se encuentren en un rango controlado. Los métodos más utilizados son la normalización *min-max*, que transforma las puntuaciones de forma lineal al intervalo [0,1], y la normalización *z-score*, que transforma la distribución de *scores* en una distribución con media 0 y varianza 1.

- **Fusión basada en reglas entrenadas** o *back-end*. Consiste en la utilización de las salidas producidas por los diferentes sistemas como patrones de entrada a un nuevo sistema. La fusión consiste en un problema clásico de clasificación de patrones que puede ser resuelto con técnicas como redes neuronales, *SVMs* [25] o regresión logística [26].

2.5. Reconocimiento de locutor independiente de texto

2.5.1. Introducción

Existen dos tipos de sistemas de reconocimiento de locutor en función del trato que se hace del contenido léxico. En los sistemas **dependientes de texto** el contenido léxico de las locuciones utilizadas tanto para entrenar el sistema como para el test es conocido, es decir, sabemos lo que el locutor está diciendo. Los sistemas **independientes de texto**, en cambio, no cuentan con la información léxica de las locuciones. Sin embargo, han dominado el reconocimiento de locutor debido a que a pesar de ser más difícil y ambicioso es más flexible y natural. Resultan además idóneos para el reconocimiento de un usuario que no quiere colaborar (necesitamos sólo su voz, no que diga algo concreto).

En este documento vamos a centrarnos en el estudio de los sistemas *independientes de texto*. Más concretamente estudiaremos aquellos basados en información espectral a corto plazo, que son actualmente los sistemas que cuentan con el mejor rendimiento global. Los distintos avances producidos en el estado del arte tratan de paliar la variabilidad producida por el canal o el ruido entre otros factores, sin embargo éstas mejoras se ven rápidamente decrementadas en presencia de locuciones de corta duración [1].

A lo largo de esta sección se presentarán las diferentes técnicas empleadas para los sistemas de reconocimiento de locutor independiente de texto basados en información acústica, haciendo especial hincapié en aquellos basados en *GMM-UBM* [27].

2.5.2. Cuantificación vectorial (*Vector Quantization VQ*)

La técnica de cuantificación vectorial o modelo *centroide* fue originalmente utilizada para la compresión de datos [28]. Comienza a utilizarse en reconocimiento de locutor a partir de los años 80 [29] [30]. Es un método para el reconocimiento de locutor independiente de texto muy simple y, aunque no se considera que tenga un gran rendimiento comparado con otros sistemas más complejos, se puede utilizar para acelerar el proceso computacional [31] [32] [33] o para la implementación de prácticas ligeras [34].

La Teoría de la distorsión y del régimen binario de *Shannon* demuestra que es **siempre posible mejorar el rendimiento de un sistema que utiliza escalares para la codificación utilizando vectores**. Los sistemas basados en cuantificación vectorial tratan de, empleando esta idea, cuantificar una señal de entrada que puede tomar infinitos valores asignándola a un vector representativo de entre un conjunto finito posible. Una de las opciones es representar un

conjunto de vectores de características próximos entre sí por su vector promedio, de esta forma dividimos cualquier espacio de características en un número de regiones determinado, cada una con su vector representativo o centroide (*codeword*). Llamaremos libro de códigos (*codebook*) al conjunto de todos los vectores representativos o *codewords*. Esta división del espacio de características se lleva a cabo mediante algoritmos de agrupamiento o *clustering* entre los que destaca por su sencillez *K-means* [35] o *binary splitting*.

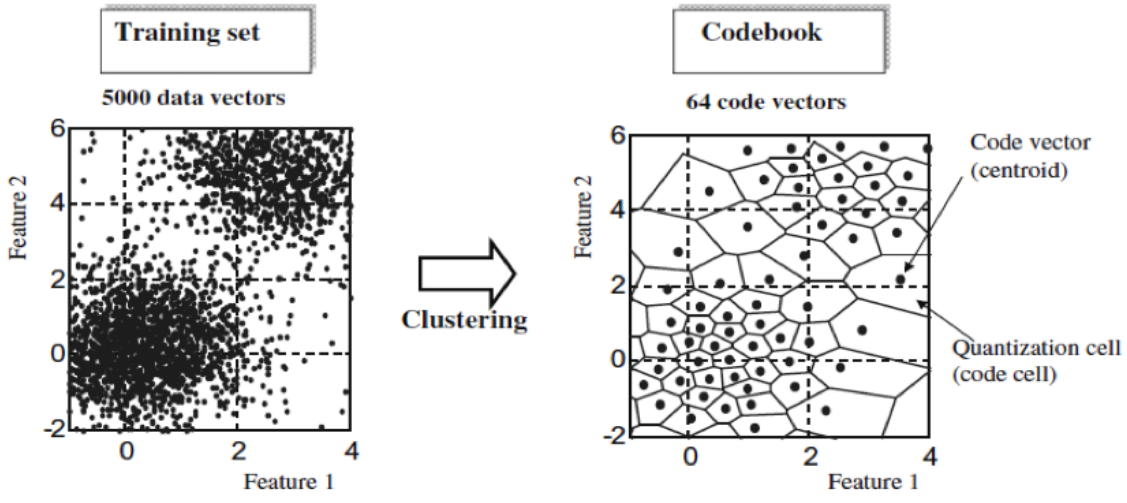

Figura 2.14: Ejemplo de cuantificación vectorial utilizando el algoritmo *k-means* extraído de [2].

En un sistema de reconocimiento de locutor basado en *Vector Quantization* la identidad de locutor queda representada mediante su *codebook*, que puede extraerse utilizando el algoritmo *k-means* como podemos ver en la figura 2.14; por tanto, en esta técnica, el *codebook* de cada locutor constituye el modelo de plantilla. Para facilitar la representación de los centroides en notación binaria se divide el espacio de características en un número de regiones potencia de 2. Un *codebook* de b bits tendrá N = 2b centroides. Toda la información que necesitamos almacenar para cada modelo de locutor está contenida en el *codebook*, lo que supone una reducción drástica de la información espectral.

Cuando un sistema basado en *VQ* recibe una locución de test se procederá, en primer lugar, a la extracción de sus vectores de características O = o1, o2, o3, ..., oT. En segundo lugar se realizará la comparación, mediante la *distorsión de cuantificación promedio* que se define en la fórmula 2.13, entre dicho vector de características y el *codebook*, vector representativo de la identidad de locutor (conjunto de centroides), R = r1, r2, r3, ..., rN.

$$D_Q(O,R) = \frac{1}{T}\sum_{t=1}^{T} min\, d(o_t, r_n) \quad 1 \leq n \leq N \tag{2.13}$$

donde $d(o_t, r_n)$ representa la distancia entre el centroide n y el vector de características t. Se pueden utilizar distintas medidas de distancia, siendo la euclídea la más popular. La probabilidad de que el conjunto de vectores O de la locución de test pertenezcan al locutor representado por R será mayor cuanto menor sea la distorsión de cuantificación promedio $D_Q(O,R)$.

En la figura 2.15 se ilustra el esquema de funcionamiento de un sistema de reconocimiento de locutor basado en cuantificación vectorial.

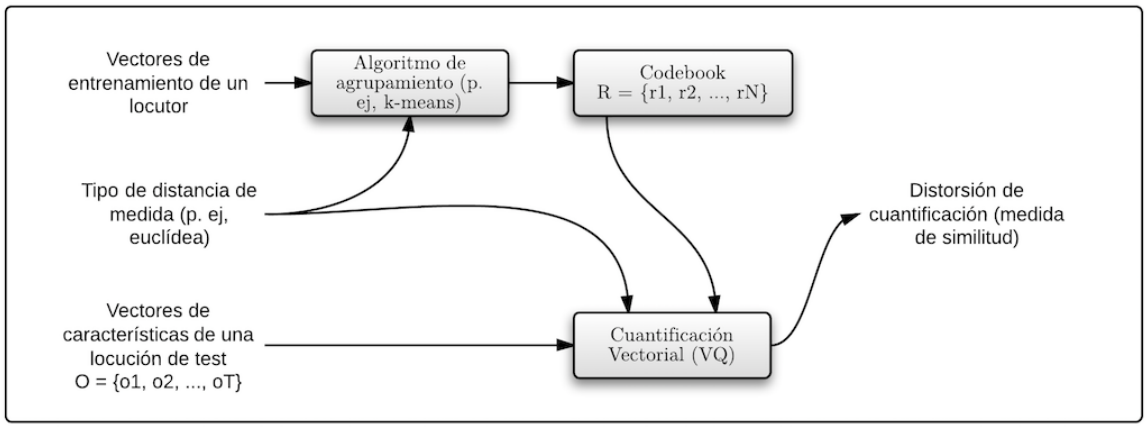

Figura 2.15: Proceso de comparación mediante cuantificación vectorial de un fichero de test y un modelo de locutor.

2.5.3. Modelos de Mezclas Gaussianas *(GMM)*

Los modelos de mezclas de Gaussianas o *GMMs* [36] [37] por sus siglas en inglés, *Gaussian Mixture Models*, son modelos estocásticos, es decir, cada locutor se modela como una función probabilística con una función de probabilidad desconocida pero fija. Durante muchos años el método basado en *GMM* ha sido la técnica de referencia por excelencia en el ámbito del reconocimiento de locutor independiente de texto. Podría considerarse el modelo *GMM* como una extensión del modelo *VQ* en el que las regiones del espacio de características se solapan entre sí; de esta forma un vector de características tiene una probabilidad no nula de pertenecer a cualquiera de las regiones (en este caso definidas por Gaussianas) en lugar de ser asignado de forma discreta a un centroide.

En el reconocimiento de locutor independiente de texto no se tiene conocimiento a priori acerca del contenido de la locución por lo que las características son continuas. Por este motivo, para implementar un detector de la razón de verosimilitud (ver sec. 2.4.2) se requiere una función de verosimilitud $p(X|\lambda)$, siendo en este caso un *GMM*.

Teóricamente sabemos **se puede reproducir cualquier distribución de probabilidad, sin importar cuán compleja sea, mediante la combinación de Gaussianas**. Apoyándose en esto como base un GMM consiste en la combinación de un conjunto finito de Gaussianas multivariadas, de forma que el espacio de características queda definido por ellas. El *GMM*, denotado por λ, queda definido mediante su función de densidad de probabilidad:

$$p(x|\lambda) = \sum_{k=1}^{K} w_k N(x|\mu_k, \Sigma_k) \tag{2.14}$$

donde w_k es la probabilidad a priori o peso de la mezcla k, K es el número de Gaussianas del modelo, que se define previamente y

$$N(x|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{-D/2}|\Sigma_k|^{-1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma (x-\mu_k)_k^{-1}\right) \tag{2.15}$$

es la función de densidad de la Gaussiana de D dimensiones (D-variada) k, formada por una combinación de densidades de Gaussianas uni-modales con vector de medias μ_k y matriz de

covarianzas Σ_k. Las probabilidades a priori quedan representadas mediante los pesos que están restringidos a $\sum_{k=1}^{K} w_k = 1$. Existen D Gaussianas uni-modales por lo que el vector de medias es de tamaño $Dx1$. Las matrices de covarianza de los *GMM* son matrices de dimensión DxD y suelen ser diagonales debido a la carga computacional, esto restringe los ejes de las curvas Gaussianas a coincidir en dirección con los ejes de coordenadas (Gaussianas con igual probabilidad en cualquier circunferencia y centradas en el eje de la misma). Además la naturaleza ortogonal de los coeficientes cepstrales *MFCC* hace que, debido a la alta independencia entre dimensiones, el error introducido al forzar matrices de covarianza diagonales sea pequeño.

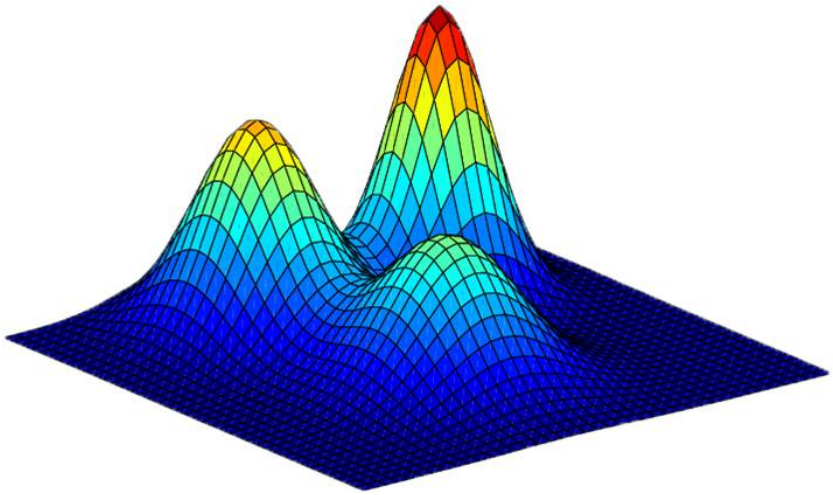

Figura 2.16: Función de densidad de probabilidad de un GMM de 3 Gaussianas sobre un espacio bidimensional.

El primer paso para utilizar *GMMs* en reconocimiento de locutor independiente de texto consiste en entrenar el modelo. El objetivo es utilizar los datos de entrenamiento $X = (x1, ..., xT)$ para estimar los parámetros del modelo $\lambda(w_k, \mu_k, \Sigma_k)_{k=1}^{K}$ de forma que la distribución del *GMM* se ajuste a los vectores de características de entrenamiento. Utilizando el método de estimación de máxima verosimilitud o *ML* por sus siglas en inglés, *Maximum Likelihood*, se pretende estimar los parámetros que maximicen la similitud del *GMM* a partir de los datos de entrenamiento. Para los vectores de características de entrenamiento $X = (x1, ..., xT)$ y asumiendo independencia entre los mismos (para obtener las matrices diagonales) tenemos la siguiente función de verosimilitud

$$p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda) \tag{2.16}$$

La ecuación 2.16 no es lineal respecto de los parámetros del modelo λ por lo que no podemos aplicar directamente el algoritmo ML, ya que la función tendrá normalmente varios mínimos locales. Sin embargo, podemos realizar esta aproximación mediante el algoritmo *Expectation Maximization* (*EM*). El algoritmo *EM* es un algoritmo iterativo con dos etapas en cada iteración: en la primera etapa el algoritmo asigna los pesos de las Gaussianas y en la segunda se reestiman los parámetros del modelo.

De forma más formal el algoritmo comienza con un modelo inicial λ y estima un nuevo modelo $\bar{\lambda} = (w_k, \mu_k, \Sigma_k)k = 1^K$ de modo que $p(X|\bar{\lambda}) \geq p(X|\lambda)$. El nuevo modelo actúa entonces como modelo inicial de la siguiente iteración repitiéndose el proceso hasta que el valor de similitud converge o se llega al número de iteraciones máximo. Durante la primera fase de cada iteración se calculan los pesos de las Gaussianas siguiendo la fórmula

$$\bar{w_k} = \frac{1}{T}\sum_{t=1}^{T} P_r(k|x_t, \lambda) \tag{2.17}$$

A continuación se fijan los pesos y se maximiza la verosimilitud actualizando el vector de medias, la matriz de varianzas y la probabilidad a posteriori respectivamente siguiendo las siguientes ecuaciones

$$\bar{\mu_k} = \frac{\sum_{t=1}^{T} P_r(k|x_t, \lambda)x_t}{\sum_{t=1}^{T} P_r(k|x_t, \lambda)} \tag{2.18}$$

$$\bar{\sigma^2} = \frac{\sum_{t=1}^{T} P_r(k|x_t, \lambda)x_t^2}{\sum_{t=1}^{T} P_r(k|x_t, \lambda)} - \bar{\mu_k^2} \tag{2.19}$$

$$P_r(k|x_t, \lambda) = \frac{\bar{w_k} N(x_t|\mu_k, \Sigma_k)}{\sum_{i=1}^{K} w_i N(x_t|\mu_i, \Sigma_i)} \tag{2.20}$$

Para reducir el número de iteraciones que el algoritmo necesitará podemos utilizar el método k-means [35] para estimar el modelo inicial λ. Para ello se realizará un procedimiento similar al explicado en la sección 2.13 obteniendo un modelo *VQ*; el vector de medias del *GMM* coincidiría con los centroides calculados; las matrices de covarianza pueden extraerse a partir de la covarianza de los vectores del conjunto de vectores de entrenamiento asignados a cada centroide; los pesos de cada Gaussiana serían proporcionales al número de vectores del conjunto de vectores de entrenamiento asignados a cada centroide sumando el total de los pesos la unidad.

Una vez los modelos han sido entrenados, siguiendo el mismo procedimiento podemos calcular la probabilidad de que el conjunto de vectores de una nueva locución pertenezca a uno de estos modelos, es decir, la probabilidad de que un modelo concreto haya generado los vectores de características de una locución de test.

2.5.3.1. GMM-UBM

Habitualmente la cantidad de datos que tenemos de cada locutor no son suficientes para generar un *GMM* robusto mediante el procedimiento descrito en la sección 2.5.3. El modelo *GMM-UBM* propuesto en [27] surge para solventar este problema.

En los sistemas basados en esta técnica se entrena primero un modelo universal (*Universal Background Model, UBM*) por medio del algoritmo *Expectation Maximization*. Este modelo representa la distribución de los vectores de características utilizando datos de muchos locutores, es decir, modela las características comunes a todos los locutores. Se utiliza una gran cantidad de audio procedente de diferentes locutores y condiciones acústicas para generar el modelo universal, de esta forma el modelo universal será más robusto frente a variabilidades como el canal ya que cuenta con datos muy diversos. Cuando se quiere registrar un nuevo locutor en el sistema se adaptan los parámetros del modelo universal a la distribución de características de dicho locutor. La idea es adaptar el *UBM* o modelo universal a los datos dependiente de locutor para generar el modelo de locutor. Solventa así dos problemas clásicos de la técnica *GMM* convencional:

- Al entrenar un modelo de locutor adaptando un modelo universal éste recogerá variabilidad acústica inexistente en sus datos de entrenamiento. Esto deriva en mayor robustez de un

modelo de locutor en presencia de **escasez de datos**, limitación principal del modelo *GMM* clásico.

- Al poder comparar un locutor con un modelo universal proporciona un mecanismo para **calcular cuán representativa o única es una identidad**. De esta forma podemos ponderar o normalizar la puntuación de una locución en función de lo representativas que sean las características presentes en la misma.

2.5.3.2. Adaptación MAP

En la sección 2.5.3 hemos visto que los parámetros que podemos adaptar de un *UBM* son pesos, vectores de medias y matrices de covarianza. Como se demuestra en [27], se pueden conseguir muy buenos resultados adaptando tan sólo los vectores de medias, de forma que tanto la cantidad de datos necesaria para adaptar el modelo *UBM* como la carga computacional se ven considerablemente reducidas. Dado un modelo *UBM* λ y el conjunto de vectores de un locutor para entrenar un modelo $X = (x1, ..., xT)$ el método **maximum a posteriori**, *MAP*, establece un compromiso entre el vector de medias del modelo universal μ_k y los nuevos datos de locutor para calcular los nuevos vectores de medias adaptados μ'_k siguiendo la siguiente ecuación

$$\mu'_k = \alpha_k \frac{1}{n_k} f_k + (1 - \alpha_k)\mu_k \tag{2.21}$$

donde

$$\alpha_k = \frac{n_k}{n_k + \tau} \tag{2.22}$$

$$n_k = \sum_t P(k|x_t) \tag{2.23}$$

$$f_k = \sum_t P(k|x_t)x_t \tag{2.24}$$

$$P(k|x_t) = \frac{w_k N(x_t|\mu_k, \Sigma_k)}{\sum_{m=1}^{K} w_m N(x_t|\mu_m, \Sigma_m)} \tag{2.25}$$

donde $P(k|x_t)$ es la probabilidad a posteriori de ocupación de la Gaussiana y n_k y f_k son los estadísticos suficientes de orden cero y primer orden respectivamente. Como podemos ver en la ecuación 2.22 los parámetros τ, factor de relevancia, y α_k, coeficiente de adaptación, controlan cuánto influyen los datos de entrenamiento en comparación a los datos originales del *UBM* sobre el modelo adaptado del locutor.

En caso de adaptar sólo los vectores de media las matrices de covarianza y el vector de pesos se mantendrán invariables por lo que las Gaussianas tan sólo se moverán o centrarán más cerca de los datos de entrenamiento sin cambiar su forma. Un modelo adaptado utilizando esta técnica se denomina *GMM-MAP*, en la figura 2.17 podemos ver el resultado de adaptar un modelo de mezcla de Gaussianas *UBM* a datos de entrenamiento dependientes de locutor.

Cuando se enfrenta una locución a un modelo se calcula la probabilidad de que los datos de test hayan sido generados a partir del modelo *GMM-MAP* del locutor adaptado λ_{target}. Por otro

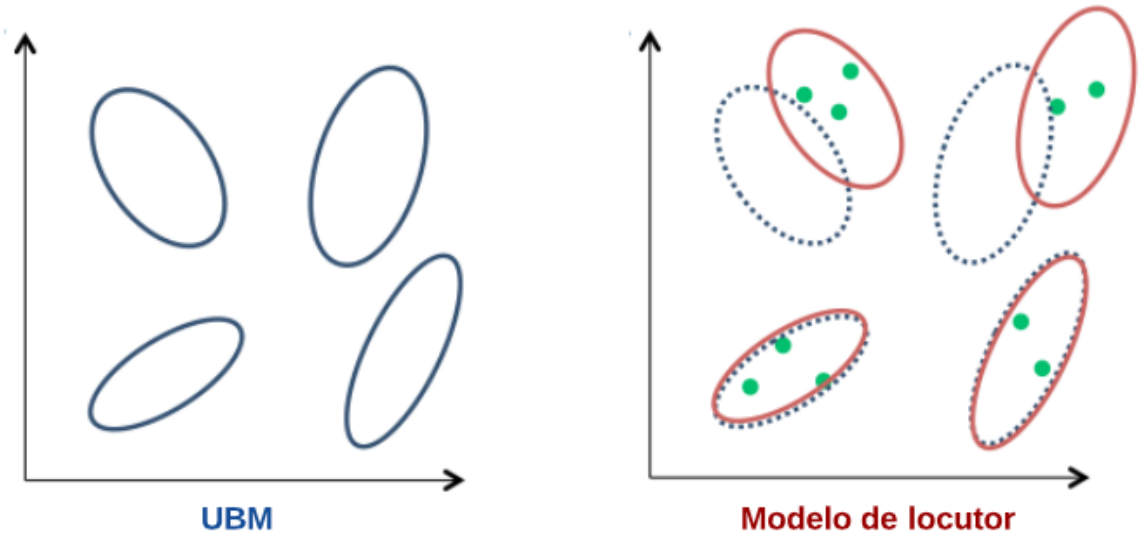

Figura 2.17: Representación del proceso de adaptación *GMM-MAP* donde un modelo de locutor (derecha) es adaptado a partir de un *UBM* (izquierda) utilizando los nuevos datos de locutor (puntos verdes).

lado se calcula también la probabilidad de que los datos de test hayan sido generados a partir del modelo *UBM*, λ_{UBM}. Esta segunda medida podría verse como la rareza de estos datos y se utiliza en la puntuación o valor de verosimilitud para ponderar el *score* siguiendo la siguiente ecuación:

$$LLR_{avg}(X, \lambda_{target}, \lambda_{UBM}) = \frac{1}{T}\sum_{t=1}^{T} logp(x_t|\lambda_{target}) - logp(x_t|\lambda_{UBM}) \tag{2.26}$$

Donde LLR_{avg} se corresponde con la razón de verosimilitud logarítmica. Cuanto mayor sea esta medida mayor será la probabilidad de que la locución de test haya sido generada por el modelo al que se enfrenta. La utilización de un mismo *UBM* común para todos los locutores supone un primer tipo de normalización ya que las puntuaciones se encontrarán en rangos directamente comparables.

2.5.3.3. Supervectores

Llamamos supervector a una forma de representar la información de un locutor o modelo presente en un modelo de mezcla de Gaussianas [38]. Este nuevo método de representación convierte la información contenida en un *GMM* a un sólo vector, lo que permite tratar el problema como un problema clásico de reconocimiento de patrones dando lugar a sistemas híbridos *GMM-SVM* (ver sec: 2.5.4.1) y a técnicas de compensación de variabilidad que veremos más adelante como *Joint Factor Analysis (JFA)*.

Un **supervector** está formado por la concatenación de los vectores de medias, de dimensión $1xD$, de las K Gaussianas de un *GMM*, en una sola fila de forma que obtenemos un vector de dimensión $1xKD$. Para que la información contenida en distintos supervectores sea comparable deben estar obtenidos a partir de *GMMs* adaptados de un mismo modelo *UBM*, ya que de esta

forma las medias corresponderán a las mismas Gaussianas y serán comparables en operaciones del espacio KD-dimensional (típicamente tiene un tamaño aproximado de 40000x1 con K=1024 y D=39).

Al representar una locución mediante un único punto en el espacio KD-dimensional de los supervectores se elimina del supervector la variabilidad no deseada. Existen varias técnicas para realizar este proceso [39] [40] [41]. Existen teorías que defienden que la variabilidad inter-locutor está contenida en un espacio distinto del que contiene la variabilidad intra-locutor. Estas técnicas se apoyan en esta teoría de forma que no necesitan conjuntos de entrenamiento que contemplen cada canal o entorno. Esto supone una gran ventaja ya que se entrena un modelo de variabilidad inter-sesión independiente de locutor que se aplicará a cualquier locutor por lo que la cantidad de datos necesaria para tener un modelo de locutor robusto se ve disminuida considerablemente. Las técnicas de *Factor Analysis* que veremos en las siguientes secciones aplican este tipo de compensación sobre sistemas basados en *GMM* obteniendo sistemas de muy alta precisión.

2.5.4. Máquina de Soporte de Vectores (*Support Vector Machines, SVMs*)

Las máquinas de soporte de vectores o *SVMs* [42], a diferencia de los *GMM*, son clasificadores discriminativos basados en técnicas de optimización que minimizan el coste. Debido a su flexibilidad y rendimiento se han utilizado en reconocimiento de locutor tanto con características espectrales [38] como con características prosódicas y de alto nivel [43].

La idea básica de un *SVM* consiste en **obtener un hiperplano que será la frontera de decisión más óptima entre dos clases**. Típicamente estas clases no serán linealmente separables pero podrán ser linealmente separables en una dimensión de tamaño mucho mayor. Esta técnica transforma los datos a este espacio de características de mayor dimensión por medio de una función *kernel* y define el hiperplano lineal en este espacio que después traerá de vuelta al espacio original convirtiéndolo en un hiperplano separador no lineal.

En un sistema de verificación una de las clases se modelará mediante los vectores de características de entrenamiento del locutor *target*, etiquetados como +1, mientras que la otra clase se modelará con los vectores de características de entrenamiento de impostor o *non-target*, etiquetados como -1. La tarea del *SVM* será encontrar el hiperplano que maximice el margen de separación entre ambas clases donde el margen se definirá en función de la distancia entre los vectores de soporte de cada clase al hiperplano como podemos ver en la figura 2.18. En dicha figura aparece representado el plano denotado por $(\hat{w}, b)$ donde w es un vector de n coeficientes que determina la orientación del plano y b es el término independiente de la ecuación paramétrica del plano.

Si volvemos a un sistema de verificación de locutor, el modelo *SVM* de locutor es el plano $\hat{w}, b$ que separa de forma óptima los vectores de características de dicho locutor y los vectores de características del resto, maximizando el margen. La función con la que obtendremos la puntuación para un vector de características $\hat{x}$ con un *SVM* $(\hat{w}, b)$ es la siguiente:

$$f(x) = w^T x + b \tag{2.27}$$

Durante el entrenamiento se busca un hiperplano que, para el conjunto de datos de entrenamiento, se cumpla que la función es mayor o igual que uno para los vectores *target* y menor o igual que menos uno para los vectores de impostor o clase *non-target*. Una vez estemos en la fase de verificación, la clasificación se basará en los valores que tome la función de puntuación $f(x)$;

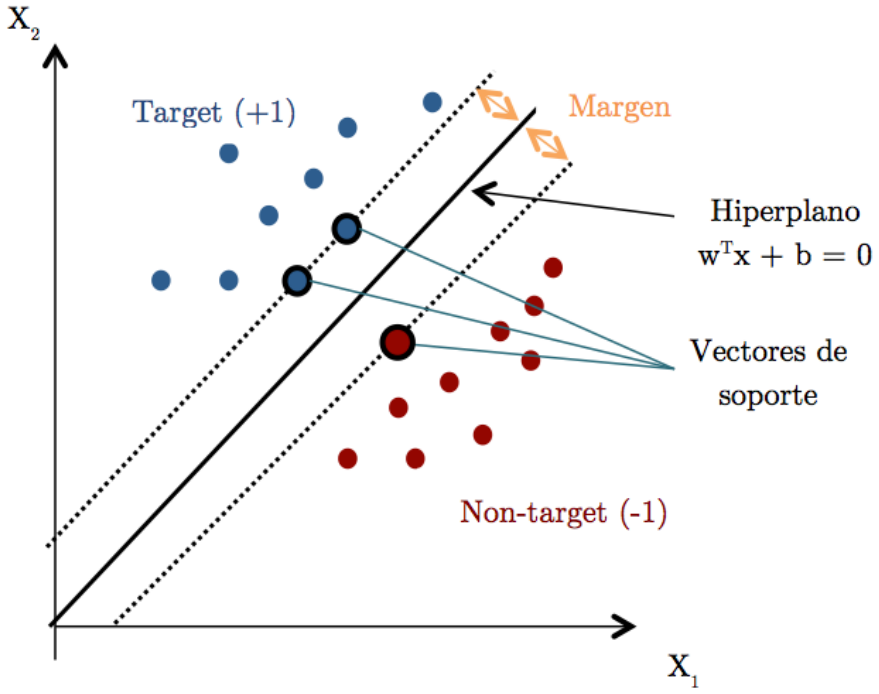

Figura 2.18: Elementos que componen un modelo *SVM*.

los vectores que cumplan $f(x) \geq 0$ pertenecerán al locutor que se pretende verificar mientras que los vectores que cumplan $f(x) < 0$ pertenecerán a la clase *non-target*

2.5.4.1. Sistemas híbridos *GMM-SVM*

La combinación de la técnica generativa *GMM* con la técnica discriminativa *SVM* da origen al concepto de *SuperVectors* [38] [44]. Esta técnica, denominada *GMM-SVM*, ha sido ampliamente utilizada para el reconocimiento de patrones ya que supone un equilibrio capaz de aprovechar los puntos fuertes de ambos sistemas: el modelado generativo de los sistemas *GMM-UBM* y el modelado discriminativo de los sistemas *SVM*.

Los sistemas *GMM-MAP*, al utilizar sólo desplazamiento de medias para adaptar un *UBM* a un locutor concreto, pueden almacenar toda la información de una locución en un supervector formado por la concatenación de dichas medias. Los sistemas *GMM-SVM* **utilizan estos supervectores generados por modelos de locutor *GMM-MAP* como vectores de entrada para las clases *target* y *non-target* de un *SVM*** convirtiendo el problema en un problema genérico de reconocimiento de patrones. De esta forma la información de cada locución, tanto de entrenamiento como de test, es procesada mediante una técnica generativa para generar un vector de características que será utilizado como entrada de un sistema discriminativo *SVM*.

2.5.5. Técnicas basadas en *Factor Analysis*

Las técnicas basadas en *Factor Analysis* (*FA*) han marcado un antes y un después en los sistemas de verificación de locutor independiente de texto **debido a la capacidad que poseen para paliar la variabilidad de sesión**. Durante los últimos años un gran número de líneas de investigación se han centrado en el estudio e implementación de estas técnicas y actualmente

estas técnicas obtienen los mejores resultados en tareas de alta exigencia como la evaluación *Speaker Recognition Evaluation, SRE* organizada cada dos años por *NIST*.

En esta sección se dará una breve descripción de las diferentes técnicas basadas en *Factor Analysis* que han supuesto un avance del estado del arte en reconocimiento de locutor independiente de texto.

2.5.5.1. Joint Factor Analysis (JFA)

La técnica *Joint Factor Analysis*, conocida como *JFA* por sus siglas en inglés, fue la primera técnica basada en *Factor Analysis* para reconocimiento de locutor [45]. Esta técnica modela de forma conjunta tanto la variabilidad intra-locutor como la variabilidad inter-locución a la que nos referiremos como variabilidad de canal.

En la sección 2.5.3 hemos visto que un modelo de locutor se puede representar con un supervector. Dado que los modelos *GMM-MAP* sólo adaptan las medias para adaptar un modelo universal a un locutor específico (todos los modelos adaptados comparten vector de pesos y matrices de covarianzas) podemos formar un supervector con toda la información concatenando las medias de los *GMM* del locutor.

Si tenemos varias locuciones de entrenamiento para un mismo locutor y obtenemos varios supervectores, podremos observar que estos difieren debido en gran parte a la variabilidad introducida por el canal de transmisión. Para compensar esta variabilidad de forma que características obtenidas en diferentes canales puedan ser comparadas correctamente con el modelo de locutor es necesario modelar esta variabilidad. El modelo *Joint Factor Analysis* asume que en un supervector que proviene de una locución h de un locutor s existe una variabilidad no deseada que está contenida en un subespacio de baja dimensionalidad y que modifica dicho supervector de la siguiente forma

$$\mu_{sh} = \mu_s + Ux_h \tag{2.28}$$

donde

- μ_{sh} es el supervector contaminado por la variabilidad de canal (que depende del locutor y de la locución)
- μ_s es el supervector ideal (dependiente sólo del locutor) con la información de locutor
- U representa el subespacio de variabilidad de sesión
- x_h representa los factores de canal o *channel factors* que dependen de la locución h.

Estos factores de locutor deben estimarse en la fase de entrenamiento y determinarán la importancia o peso de cada dirección de variabilidad del subespacio de variabilidad de sesión U. La matriz U está formada por columnas denominadas *eigenchannels* y se estiman a partir de un conjunto de datos de entrenamiento independiente de locutor con una gran variabilidad de canal.

Un supervector de locutor, μ_s, puede descomponerse de la siguiente manera:

$$\mu_s = \mu + Vy_s + Dz_s \tag{2.29}$$

donde

- μ es el vector de medias del modelo universal *UBM*
- V es una matriz rectangular cuyas columnas se denominan *eigenvoices* y define el subespacio de variabilidad del locutor
- y_s son los factores de locutor o *speaker factors*, definen el peso de las distintas direcciones de variabilidad de locutor en V y representan al locutor s en dicho espacio de variabilidad
- D es una matriz diagonal cuadrada de tamaño Kd y representa el desplazamiento del supervector debido a la adaptación *MAP*
- z_s es un vector columna de longitud Kd.

En la ecuación 2.29 podemos ver que, haciendo $y_s = 0$, tenemos que $\mu_s = \mu + Dz_s$, que coincide con el proceso de adaptación MAP. La técnica *JFA* puede, por tanto, considerarse una expansión de la técnica *MAP* que incluye el modelado de *eigenvoices* o *speaker factors* que nos ayudará en caso de escasez de datos de entrenamiento dado que restringe la adaptación de medias en el entrenamiento a un espacio V de dimensionalidad menor que el utilizado en el modelo *MAP*.

En el proceso de identificación se obtendrán los estadísticos de orden cero y primer orden de la locución de test respecto del *UBM* definidos por las ecuaciones 2.23 y 2.24 respectivamente. La probabilidad de que la locución de test haya sido generada a partir del modelo de locutor que se pretende verificar se obtiene mediante el producto escalar del supervector del modelo por el estadístico de primer orden, normalizado por la suma de los elementos de los estadísticos de orden cero.

2.5.6. Total Variability (TV)

La técnica *Total Variability* [46] representa un paso más en el modelo *JFA*. Habitualmente no se disponen de datos suficientes para poder realizar una estimación robusta de los subespacios de variabilidad de locutor y de sesión por separado. *Total Variability* pretende resolver este problema modelando conjuntamente en la variabilidad de sesión y de locutor en un solo subespacio. De una forma más formal este modelo puede definirse como:

$$\mu_s = \mu + Tw \tag{2.30}$$

donde T representa la matriz de *total variability* y w contiene los factores o variables latentes de este modelo, también conocido como *i-vector* (vector de identidad), típicamente de una dimensión entre 400x1 o 600x1. Al igual que con los supervectores una locución queda representada por un sólo vector que confina la variabilidad tanto de locutor como de sesión pero en un espacio cuya dimensionalidad es mucho menor.

Este nuevo subespacio, además de poder estimarse con una cantidad menor de datos, nos permite, gracias a su reducida dimensionalidad, la utilización de técnicas clásicas como *LDA (Linear Discriminant Analysis)* que han demostrado su utilidad en sistemas de muy alto rendimiento [47].

Una de las formas de obtener el *score* en esta técnica consiste en aplicar la distancia coseno entre el *i-vector* generado a partir de la locución de test y el *i-vector* que modela el locutor que se pretende verificar.

2.5.7. Probabilistic Linear Discriminant Analysis (PLDA)

Como hemos remarcado en la sección anterior, la técnica *Total Variability* tiene la principal ventaja de reducir una locución a una representación de pocas dimensiones y longitud fija, el *i-vector*. A partir de este punto pueden usarse métodos clásicos de clasificación como *Linear Discriminant Analysis (LDA)* directamente sobre los *i-vectors* para separar la variabilidad de locutor de la variabilidad de sesión.

Probabilistic Linear Discriminant Analysis (PLDA) da un paso más y consiste en un modelo generativo que ha sido utilizado recientemente con gran éxito para el modelado de *i-vectors* [48]. Podemos entender *PLDA* como una versión probabilística de la técnica clásica *LDA*, asumiendo que un *i-vector* i de un locutor s puede descomponerse de la siguiente forma

$$w_{si} = \mu + Fh_s + Gk_i + \epsilon_i \tag{2.31}$$

donde F y G representan los nuevos subespacios de variabilidad de locutor y de sesión respectivamente, h_s y k_i son sus variables latentes asociadas respectivamente y ϵ_i es un término de ruido residual que se asume distribuido normalmente con media cero y matriz de covarianza Σ. En la figura 2.19 podemos ver el modelo gráfico probabilístico de *PLDA*.

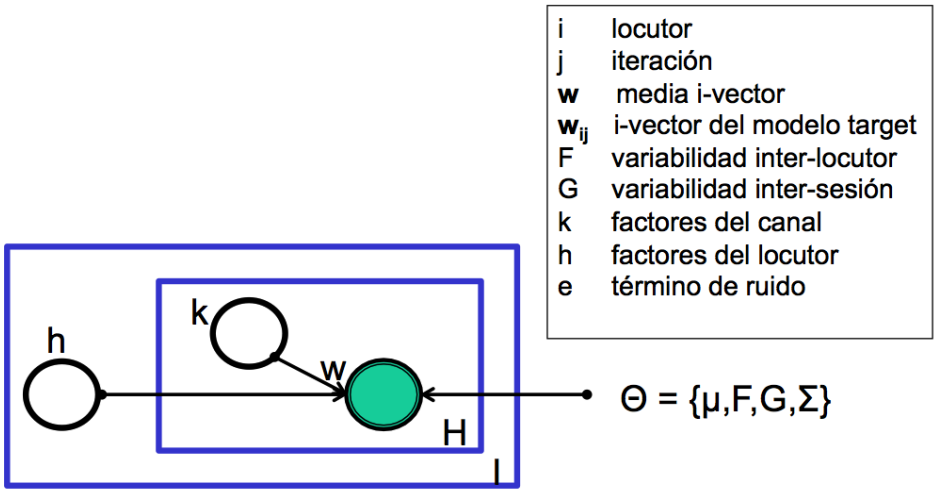

Figura 2.19: Modelo gráfico probabilístico de *PLDA*

En la fórmula 2.31 es fácil detectar una analogía entre el modelado *Joint Factor Analysis* y el modelado *PLDA*. Sin embargo, existen dos diferencias clave que debemos tener en cuenta:

- *JFA* actúa sobre supervectores (alta dimensionalidad) mientras que *PLDA* parte de *i-vectors* (baja dimensionalidad).
- *JFA* asume que los supervectores sobre los que trabaja han sido generado por una mezcla de Gaussianas mientras que *PLDA* asume que los *i-vectors* son generados a partir de una sola Gaussiana multivariada.

El *score* o puntuación entre un modelo y una locución de test representados por los *i-vectors* w_1 y w_2 puede calcularse como el ratio de dos hipótesis: H_0 defiende que ambos vectores han sido

generados por la misma identidad (mismo h_S) mientras que H_1 defiende lo contrario (diferente h_S). Este ratio puede ser expresado como

$$S_{w_1,w_2} = \frac{p(w_1, w_2|H_0)}{p(w_1|H_1)p(w_2|H_1)} = \frac{\int p(w_1, w_2|h)p(h)\,\mathrm{d}h}{\int p(w_1|h_1)p(h_1)\,\mathrm{d}h_1 \int p(w_2|h_2)p(h_2)\,\mathrm{d}h_2} \tag{2.32}$$

Si se asumen priors Gaussianos para las variables latentes las integrales de esta ecuación son tratables y el *score* S_{w_1,w_2} puede derivarse con una formula cerrada. Un análisis más detallado puede encontrarse en [49]

3

Marco experimental.

3.1. Introducción

El reconocimiento de locutor empezó en la década de los 70. Durante sus inicios, los experimentos científicos eran llevados a cabo utilizando datos recogidos expresamente para realizar dichos experimentos. En la mayoría de estos casos el proceso de recogida de datos era llevado a cabo a mano [14]. A medida que este área fue creciendo en importancia se comenzó a desarrollar un marco experimental común compuesto por bases de datos y protocolos que permiten comparar sistemas diseñados por diferentes grupos o personas justamente y facilita el aprovechamiento de contribuciones de terceros de forma que el estado del arte comenzó a desarrollarse de forma más rápida y regular.

La aparición del *American National Institute of Standards and Technology, NIST,* [50] supuso un antes y un después en la comunidad científica. Esta organización comenzó a organizar de forma regular evaluaciones de reconocimiento de locutor en 1996 (*SRE*). Estas evaluaciones han proporcionado bases de datos que se han convertido en estándares de facto facilitando la tarea de la comunidad científica. Las evaluaciones *NIST* están diseñadas con los siguientes objetivos [51]:

- Explorar nuevas ideas en el campo del reconocimiento de locutor.
- Desarrollar tecnología puntera que incorpore dichas ideas
- Poder medir de forma cuantitativa el rendimiento de estas tecnologías.

En este capítulo se presentarán las bases de datos que se han utilizado así como los protocolos de evaluación, que se corresponden con los utilizados en la evaluación *NIST SRE 2010* [20]. A continuación se detallará la descripción del sistema de verificación de locutor con el que se han realizado las pruebas, describiendo las distintas etapas: pre-procesado, parametrización, tipos de modelado y normalización de puntuaciones.

3.2. Bases de datos

Desde la aparición de las técnicas basadas en *UBMs* (ver sec: 2.5.3.1) los sistemas se ven beneficiados de la introducción de mayor cantidad de datos proveniente de diferentes locutores, canales y condiciones de forma que el modelo universal sea lo más robusto posible. Además estos datos se utilizan para tareas como entrenar las matrices de compensación de los modelos basados en *Factor Analysis* o crear las cohortes que se utilizarán para normalizar la puntuación por lo que la cantidad y calidad de los datos se convierte en un factor clave en el desarrollo de los sistemas.

Para el entrenamiento de las distintas técnicas de compensación de variabilidad se necesita que dicha variabilidad se vea reflejada en los datos con los que se entrenan los algoritmos. Por este motivo, las bases de datos que se utilizarán deberían contar con una gran población de locutores. Además se deberán combinar locuciones tomadas en diferentes canales (telefónico, microfónico, etc.), en diferentes condiciones (entornos controlados, entornos con y sin ruido de fondo, etc.), en diferentes sesiones o momentos de tiempo, etc.

A continuación se presentan las bases de datos que se han utilizado en este proyecto tanto para el entrenamiento de los distintos algoritmos como para la evaluación del rendimiento [52].

- **Switchboard 1** [53]: Esta base de datos contiene habla conversacional en inglés americano sobre línea telefónica convencional. Las locuciones tienen una duración aproximada de 2,5 minutos y contienen un total de 543 locutores diferentes. Recoge variabilidad producida por el uso de diversas líneas de teléfono y terminales telefónicos (micrófono de carbon, tipo *electrect*, etc). Fue publicada en 1997 y se utilizó en la evaluación *NIST* 2001.
- **Switchboard 2** [54]: De características similares a la primera versión de *switchboard* se diferencia principalmente por contener mayor variabilidad debida a líneas y terminales telefónicos. Fue recogida en tres fases de forma que contiene variabilidad dialectal: La primera fase contiene inglés americano de la mitad Atlántica; La segunda fase inglés americano de la mitad oeste; La tercera fase inglés americano del sur. La tercera fase ha sido empleada en las evaluaciones *NIST* de 1996 a 1999; La segunda fue utilizada junto con la primera en las evaluaciones de 2002 y 2003; En la evaluación *NIST* 2000 se utilizó la base de datos completa. Las tres fases fueron publicadas en 1998, 1999 y 2002 y contienen un total de 657, 679 y 640 locutores respectivamente.
- **Switchboard 3** [55]: También conocida como *Switchboard cellular*, se diferencia por haber sido recogida en redes móviles. Contiene inglés américano y fue grabada en dos fases: La primera fase se grabó sobre un canal de transmisión *GSM* y contiene 254 locutores; La segunda fase utiliza un canal de transmisión *CDMA* y contiene 419 locutores. Fue publicada en 2001 y se ha utilizado en las evaluaciones *NIST* de 2001 a 2003.
- **Mixer** [56]: La base de datos *Mixer* surge para satisfacer la creciente necesidad de datos más realistas que incluyan más variabilidad de forma que suponga un reto mayor para los sistemas de reconocimiento de locutor. Recoge mayor variabilidad de canal y terminales, incluyendo teléfonos inalámbricos y redes móviles. Además esta base de datos es multilenguaje incluyendo inglés americano, español, árabe, chino mandarín y ruso. Fue empleada por primera vez en la evaluación *NIST* 2004 y ha sido utilizada y ampliada desde entonces, actualmente contiene cerca de 2200 locutores distintos.

La mayor parte de estos datos han sido empleados en la fase de entrenamiento del sistema, ya sea para entrenar los modelos *UBM*, las matrices de compensación de variabilidad de sesión

utilizadas en técnicas basadas en *Factor Analysis*, o las cohortes de normalización de puntuación (*S-Norm*).

Para el entrenado de los modelos y la realización de las pruebas se han utilizado los datos masculinos telefónicos de la evaluación de *NIST SRE 2010* [20] correspondientes a la tarea *10s-10s* donde disponemos de aproximadamente 10 segundos de habla neta tanto para el entrenamiento de un modelo como para la fase de verificación.

3.3. Protocolo de evaluación

Un protocolo de evaluación consiste en un conjunto de condiciones que todo sistema debe cumplir para participar en dicha evaluación. Suele determinar aspectos como la **base de datos a utilizar** (o si se pueden utilizar bases de datos externas) para entrenar los algoritmos, los datos que se deben utilizar para entrenar los modelos de locutor así como sus características (duración, canal, idioma...). Suele determinar el número de enfrentamientos que debe realizar cada locutor *target*, la proporción de usuarios en función de género, etc. Finalmente define los **métodos para medir el rendimiento** de forma objetiva e igualitaria para todos los sistemas.

En la realización de este proyecto se ha seguido el protocolo de evaluación diseñado por el *American National Institute of Standards and Technology* para las evaluaciones de reconocimiento de locutor, más concretamente, para la evaluación de reconocimiento de locutor llevada a cabo en 2010 [20].

3.3.1. Evaluación *NIST* de reconocimiento de locutor 2010

En esta sección se resumen los puntos clave de la evaluación de reconocimiento de locutor llevada a cabo por el *NIST* en 2010, *NIST SRE 2010*. Para una descripción más detallada ver [20].

Las evaluaciones más competitivas en reconocimiento de locutor son organizadas por *NIST*. Se han llevado a cabo anualmente desde 1996 hasta 2006, cuando comenzaron a realizarse de forma bianual alternándose con evaluaciones de reconocimiento de idioma. Siguiendo los objetivos de las evaluaciones (impulsar el desarrollo tecnológico, medir el estado del arte y encontrar técnicas novedosas que hagan frente a los desafíos) las condiciones de la evaluación han ido endureciéndose incluyendo distintos canales en lugar de sólo telefónico, entrevistas en lugar sólo habla conversacional, varios idiomas, etc. Estas evaluaciones son abiertas, de forma que cualquier grupo de investigación, empresa o entidad puede participar siempre que presente el sistema desarrollado.

Desde la aparición de las evaluaciones *NIST* de reconocimiento de locutor, *SRE*, los distintos protocolos y conjuntos de datos definidos por la entidad se han convertido en un estándar para la publicación de resultados en este ámbito, favoreciendo la comparación de sistemas de forma justa así como la posibilidad de replicar pruebas de terceros lo que en definitiva se ve reflejado en una mejora más rápida del estado del arte.

De forma resumida la organización ofrece una medida de rendimiento o función de coste y los datos con los que se realizarán los enfrentamientos. Una tarea se nombra concatenando el tipo de fichero utilizado para entrenar el modelo seguido del tipo de fichero de test. Por ejemplo la tarea *10-sec/10-sec* consiste en la utilización tanto para el entrenamiento de un modelo como para el test de un fichero de tipo *10-sec*. En cuanto a los datos estos pueden dividirse en dos grupos:

- **Datos de entrenamiento:** También conocidos como datos de *train*, son utilizados para aprender al locutor, es decir, para realizar el modelo dependiente de locutor con el que se compararán los ficheros de test. En la evaluación de 2010 se utilizaron 4 tipos de ficheros de entrenamiento:

  1. ***10-sec***: Consiste en un fichero telefónico de habla conversacional que contiene 10 segundos de habla neta del locutor *target*.
  2. ***core***: Consiste en un segmento de una conversación telefónica o entrevista. El contenido neto de habla del locutor *target* es apróximadamente dos minutos y medio.
  3. ***8conv***: Consiste en 8 conversaciones (de 5 minutos de duración total) telefónicas.
  4. ***8summed***: 8 conversaciones telefónicas con dos canales superpuestos en que no se especifica cuándo está hablando el locutor *target* y cuándo está hablando el otro locutor.

- **Datos de test:** Son los datos que se utilizan para verificar un locutor, se comparan con los modelos creados a partir de datos de *train* siendo un enfrentamiento cada par modelo fichero de test. En la evaluación de 2010 se utilizaron 3 tipos de ficheros de test:

  1. ***10-sec***: Consiste en un fichero telefónico de habla conversacional que contiene 10 segundos de habla neta del locutor *target*.
  2. ***core***: Consiste en un segmento de una conversación telefónica o entrevista. El contenido neto de habla del locutor *target* es apróximadamente dos minutos y medio.
  3. ***summed***: Una conversación telefónica con dos canales superpuestos en que no está especificado cuándo está hablando el locutor *target* y cuándo está hablando el otro locutor.

En la evaluación *NIST SRE 2010* se utiliza la función de coste *DCF* (ver sección 2.4.5). En cada evaluación se proporcionan el coste de falsa aceptación C_{FA} y falso rechazo C_{FR}, así como la probabilidad a priori de que una locución de test pertenezca al locutor *target* P_{Tar}. Para la evaluación 2010 se utilizaron dos puntos de trabajo diferentes:

- Para los ficheros de test *core* y *8conv/core* se utilizará $C_{FA} = 1$, $C_{FR} = 1$ y $P_{Tar} = 0{,}001$.
- Para el resto de pruebas se utilizará $C_{FA} = 10$, $C_{FR} = 1$ y $P_{Tar} = 0{,}01$.

Este proyecto tiene como objetivo estudiar y adaptar un sistema cuyo rendimiento ha sido probado en múltiples evaluaciones *NIST* con ficheros conversacionales cuando se utilizan ficheros de corta duración. Por este motivo en este proyecto nos centramos en la tarea ***male 10-sec/10-sec*** en la que contamos con 10 segundos de habla neta del locutor tanto para entrenar el modelo como para la verificación del locutor, utilizando sólo datos de locutores masculinos.

3.4. Descripción de los sistemas implementados

En esta sección se presenta una descripción de los sistemas de verificación de locutor implementados. En primer lugar se detalla la parte del proceso común a ambos sistemas: el preprocesado, el tipo de parametrización y los datos utilizados hasta la creación de los *i-vectors*. A continuación se describe el modelado para los distintos sistemas por separado: *Total Variability* y *PLDA*. Los dos sistemas están basados en la técnica *GMM-UBM* (ver sección 2.5.3.1) por lo

que se debe disponer de un modelo universal *UBM* que recoja toda la variabilidad posible como se ha explicado anteriormente.

Los experimentos llevados a cabo utilizan la parte masculina de los datos telefónicos incluidos en la tarea *10s-10s* de la evaluación *SRE 2010* [20]. De forma más específica se han evaluado un total de 10858 enfrentamientos pertenecientes a 264 y 290 modelos y ficheros de test respectivamente.

3.4.1. Pre-procesado de la señal

En la sección 2.2.2 vimos que el pre-procesado de la señal es una etapa opcional de cualquier sistema de reconocimiento biométrico. El pre-procesado de señal pretende compensar posibles degradaciones de la misma o facilitar, de cualquier manera, la posterior extracción de características. En el reconocimiento de locutor independiente de texto existen al menos dos tipos de fuentes de error que pretenden paliarse o disminuirse mediante el pre-procesado: El ruido y la existencia de silencios o fragmentos de la locución sin voz.

A continuación se presentarán brevemente las técnicas que se han utilizado en la etapa de pre-procesado en los sistemas analizados.

3.4.1.1. Filtrado *Wiener*

La primera fase de nuestro sistema consiste en filtrar todo el audio que vamos a utilizar mediante el filtrado *Wiener* para reducir el ruido [57]. Este filtro es uno de los filtros lineales óptimos más importantes. No se trata de un filtro adaptativo ya que se suponen entradas estacionarias. En la figura 3.1 podemos ver el esquema de funcionamiento en su forma más general donde x(n) es la señal de entrada, d(n) la respuesta deseada y el filtro está representado por su respuesta al impulso h(n). El filtro produce a la salida la señal y(n). La diferencia entre la señal de salida del filtro, y(n), y la señal deseada, d(n), es el error de estimación e(n).

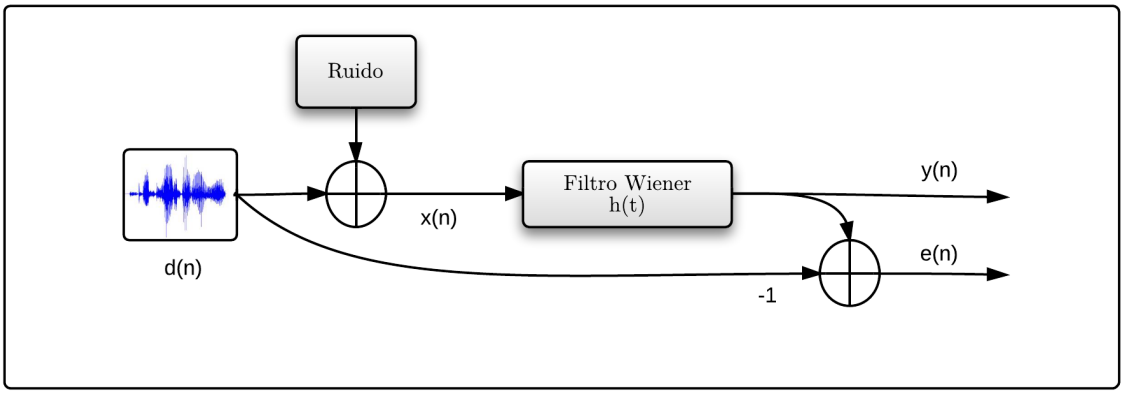

Figura 3.1: Esquema de funcionamiento básico del filtrado *Wiener*.

El objetivo del filtro de *Wiener* es determinar la respuesta al impulso h(n) de forma que el error, e(n), sea, en sentido estadístico, minimizado. El criterio por el que se rige el filtro de *Wiener* es la minimización del valor cuadrático medio del error.

Existen diversas técnicas y estructuras para realizar el filtrado de *Wiener*, en este proyecto se ha utilizado el filtro *Wiener* incluido en la herramienta *Qualcomm, ICSI, OGI (QIO) front-end software* disponible en [58] utilizando ventanas de 20 mili-segundos con un solape del 50 por ciento y ventanas *Hamming*.

3.4.1.2. Detector de actividad de voz

Cualquier sistema automático de reconocimiento de locutor se beneficia de las partes de la señal que son características del locutor objetivo o *target*. Como hemos visto en las descripciones de las bases de datos (ver sec: 3.2), las locuciones que se utilizan en nuestros experimentos provienen de conversaciones telefónicas por lo que muy posiblemente contendrán silencios y otros segmentos cuyo contenido difiere de la señal de voz. Los silencios producirían vectores de características similares para locutores diferentes. Por este motivo se utiliza una etapa de detección de actividad de voz (*Voice Activity Detection, VAD)*.

La detección de actividad de voz tiene como objetivo localizar las partes de la señal donde existe voz de forma que podamos eliminar las partes restantes, comúnmente silencios o ruidos. Existen muchos métodos diferentes para la detección de actividad de voz basados en diferentes niveles de información.

En nuestro sistema se utiliza una combinación de dos detectores de actividad de voz distintos. En primer lugar se utiliza un detector de actividad clásico, basado en **energía temporal**, que ha sido desarrollado por el grupo *ATVS*. Por otro lado se utiliza el *VAD* que proporciona la herramienta comercial *Sound eXchange*, también conocida como *SOX* [59], que está basado en una medida de la **potencia *cepstral*** de la señal y realiza varias pasadas por cada locución adaptándose a esta. Para combinar ambos detectores de voz se ha utilizado una función *AND* de forma restrictiva, es decir, se toman como voz válida aquellos segmentos que son detectados como tal por ambas herramientas.

3.4.2. Extracción de características

Para la extracción de características se ha utilizado la técnica descrita en la sección 2.3.1. Se ha utilizado una configuración de **38 coeficientes** *MFCC* donde 19 son coeficientes *cepstral* y los 19 restantes son coeficientes δ, aproximaciones lineales de las derivadas de primer orden que incluyen información dinámica de los coeficientes *cepstral*. En la figura 3.2 podemos ver el proceso de extracción de los coeficientes *cepstral*.

Para la extracción de dichos coeficientes se han utilizado ventanas de tipo *Hamming* con una duración de 20 mili-segundos espaciadas entre si 10 mili-segundos, es decir, con un solape del 50 por ciento. Los filtros *MEL* utilizados van desde los 300 hasta los 3000 hercios de forma el sistema se concentre lo más posible en el rango de posibles frecuencias de la voz.

3.4.3. Generación del modelo universal *UBM*

Las técnicas que vamos a utilizar en la fase de modelado en este estudio precisan de un modelo universal *UBM*. Además, el efecto del número de Gaussianas utilizadas en nuestros modelos sobre el rendimiento de los sistemas pretende ser también estudidado. Por este motivo se han generado múltiples *UBMs* que comprenden el rango desde 64 Gaussianas hasta 1024 incluyendo todas las potencias de 2 intermedias.

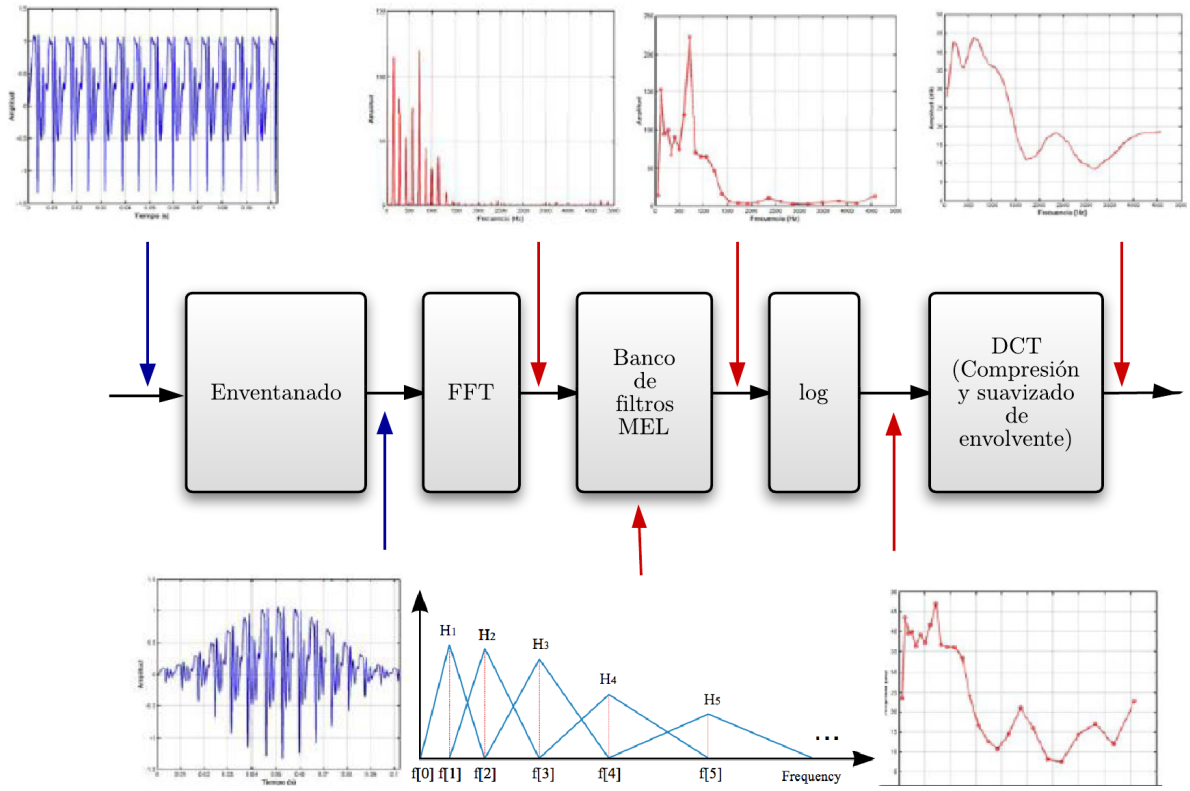

Figura 3.2: Esquema de extracción de los coeficientes *MFCC*.

Para la generación de dichos *UBMs* se utiliza el conjunto de **datos de desarrollo** también conocido con conjunto de *development* que se corresponde con las bases de datos *Switchboard I*, *Switchboard II*, *Switchboard III* (ver sección 3.2) y los datos utilizados en las evaluaciones *SRE04*, *SRE06* y *SRE08* de reconocimiento de locutor organizadas por *NIST* [50]. Los datos que se han utilizado pertenecen a la tarea *1conv/short2* por lo que contienen aproximadamente dos minutos y medio de habla neta. Los *UBMs* generados contienen aproximadamente 6 millones de vectores obtenidos agrupando vectores obtenidos de locuciones provenientes de todo el conjunto de *development*. El procedimiento que se ha seguido para ello es el explicado en la sección 2.5.3.1.

A partir de este punto los procedimientos explicados a continuación deben repetirse para los distintos *UBMs*, es decir, se replicarán todos los pasos en función del número de Gaussianas utilizadas en el modelo de mezcla de Gaussianas.

3.4.4. Generación de *i-vectors*

El siguiente paso necesario en ambos sistemas de reconocimiento de locutor es la **extracción de los *i-vectors*** a partir de los datos. Se utilizarán tres conjuntos distintos de datos, los datos de *development* detallados en la sección 3.4.3, los datos *train* que se corresponden con 264 modelos diferentes y los datos de *test* formados por un total de 290 locuciones. Los conjuntos de datos de *test* y *train* son locuciones telefónicas extraídas de la parte masculina de la tarea *10-sec/10-sec* de la evaluación *SRE10* [20] por lo que contienen aproximadamente 10 segundos de habla neta del locutor *target* y serán los datos que se utilicen para entrenamiento de los modelos y como locuciones de test respectivamente.

El proceso que se ha seguido para la generación de los *i-vectors* es el explicado en la sección 2.5.6. Para el entrenamiento de la matriz T, que confina toda la variabilidad de locutor y de sesión, se han utilizado un total de 5638 ficheros provenientes de 823 locutores diferente del conjunto de datos de *development*. El tamaño del *i-Vector* ha sido también objeto de estudio obteniéndose resultados para valores comprendidos entre 50 y 400.

3.4.5. Sistema *Total Variability*

La técnica *Total Variability*, descrita en la sección 2.5.6, confina mediante la matriz T toda la variabilidad contenida en un supervector en un subespacio de dimensionalidad mucho más reducida. De esta forma una locución queda caracterizada mediante un sólo vector de este subespacio.

Total Variability o *TV* no define ningún mecanismo para la obtención de puntuaciones de verosimilitud pero la reducción de dimensionalidad nos permite la utilización de técnicas clásicas como *LDA* para poder separar la variabilidad de locutor de la variabilidad de sesión. De ahora en adelante, en este documento denominaremos sistema *TV* a la combinación de la técnica *Total Variability* seguida de *Linear discriminant Analysis (LDA)* y *Within Class Covariance Normalization (WCCN).*

Linear discriminant analysis o *LDA* es un método clásico utilizado en reconocimiento de patrones para encontrar una combinación lineal de características que favorezca la separación de, en este caso, dos clases (variabilidad de locutor y variabilidad de sesión). En nuestro sistema se han utilizado 5214 locuciones provenientes de 611 locutores del conjunto de datos de *development* para entrenar la matriz de *LDA* que nos servirá para separar la variabilidad de locutor de la variabilidad de sesión.

Por otro lado también se han utilizado un total de 4705 locuciones provenientes de 611 locutores para entrenar la matriz descrita en el método *Within Class Covariance Normalization* o *WCCN* cuya finalidad es maximizar la variabilidad intra-locutor minimizando la variabilidad inter-locutor siguiendo el método que se describe en [60].

Utilizando estas matrices el *score* o valor de similitud entre un modelo definido por el *i-vector* w_1 y un fichero de test definido por el *i-vector* w_2 es obtenido mediante una distancia coseno siguiendo la siguiente formula:

$$S_{w_1,w_2} = \frac{(A^t w_1)W^{-1}(A^t w_2)}{\sqrt{(A^t w_1)W^{-1}(A^t w_1)}\sqrt{(A^t w_2)W^{-1}(A^t w_2)}} \tag{3.1}$$

donde A es la matriz de *LDA* y W es la matriz de covarianza intra-clase correspondiente a *WCCN.*

3.4.6. Sistema *PLDA*

El modelado *Probabilistic Linear Discriminant Analysis* o *PLDA*, descrito en la sección 2.5.7, es un modelado generativo que parte de los *i-vectors* y supone que la variabilidad tanto de locutor como de sesión están contenidas en subespacios de los mismos. Esta técnica puede verse como una versión probabilística de la técnica de modelado clásica *LDA* que se utiliza en el sistema *TV.*

Dada la reducida dimensionalidad intrínseca de los *i-vectors* así como la limitada cantidad de datos disponibles para el entrenamiento del modelo *PLDA* se ha optado por agrupar todos los términos de ruido de la ecuación 2.31 en una matriz de covarianza completa Σ. Para entrenar las matrices F y Σ se han utilizado los mismos datos que se utilizaron para entrenar la matriz T del modelo *Total Variability*, es decir, 5638 locuciones provenientes de 823 locutores.

En este sistema se ha empleado la técnica de normalización de longitud descrita en [61]. El método que se propone en dicha publicación pretende compensar el posible comportamiento no-Gaussiano de los *i-vectors* mediante una simple normalización por longitud. Esta transformación no lineal permite el uso de modelos probabilísticos que asumen datos Gaussianos, como *PLDA*, con una notable mejora del rendimiento.

Para la implementación del sistema basado en la técnica *PLDA* y el cálculo de *scores* final se ha seguido el procedimiento descrito en la sección 2.5.7.

3.4.7. Normalización simétrica *s-norm*

La última etapa de nuestros sistemas consiste en la normalización de *scores* o puntuaciones. El objetivo de esta etapa consiste en confinar los *scores* generados a partir de diferentes locutores en un rango de valores similar. Se pretende disminuir así el posible desalineamiento de los *scores* de forma que podamos utilizar un mismo umbral para todos los enfrentamientos. En la sección 2.4.6 se describen las técnicas clásicas de normalización de puntuaciones. La técnica utilizada en nuestros sistemas, *s-norm* [62], comenzó a utilizarse a partir de 2010 con muy buenos resultados para técnicas basadas en *Total Variability* (ver sec: 2.5.6).

En la técnica de normalización simétrica, *s-norm*, se define $\Lambda_{Imp} = (I_{Imp} \cup M_{Imp})$ como la unión entre una cohorte de locuciones de impostor y una cohorte de modelos de impostor. Para cada locución de test $t'_i, i = 1, ..., k$ y cada modelo $w'_j, j = 1, ..., n$ se extraen los parámetros *s-norm* formados por la media y la desviación típica de los *scores* generados al enfrentar cada locución (t'_i o w'_j) con todos los elementos de Λ_{Imp}. El *score* normalizado se obtiene de la siguiente forma:

$$s_{s-norm}(w'_s, t'_i) = \frac{s(w'_s, t'_i) - \mu_{ws}}{\sigma_{ws}} + \frac{s(w'_s, t'_i) - \mu_{ti}}{\sigma_{ti}} \tag{3.2}$$

donde μ_{ws} y σ_{ws} son los parámetros s-norm de w'_s y μ_{ti} y σ_{ti} son los parámetros s-norm de t'_i.

La normalización simétrica aprovecha el hecho de que los *i-vectors* describen las características de locutor para cualquier locución, ya sea de entrenamiento o de test, para conseguir un procedimiento universal para calcular los parámetros *s-norm* y poder llevar a cabo una normalización eficiente y simple.

4

Experimentos realizados y resultados.

4.1. Introducción

En este capítulo se presentan los experimentos más significativos llevados a cabo durante la elaboración de este Proyecto Fin de Carrera. Con ellos se ha tratado de analizar y ajustar el efecto de las técnicas de compensación de variabilidad (explicadas en los capítulos 2 y 3) aplicadas a locuciones de corta duración (en torno a 10 segundos) extraídas de las bases de datos que se detallan en la sección 3.2.

Los experimentos realizados tienen como objetivo comprobar los efectos de la compensación de variabilidad intersesión y del modelado de la variabilidad de locutor, adaptando los distintos parámetros y opciones de forma óptima en presencia de locuciones de corta duración. De esta forma variaremos parámetros como la cantidad de Gaussianas utilizadas en el modelado de mezcla de Gaussianas o el tamaño de las distintas matrices de compensación de variabilidad utilizadas.

La prueba llevada a cabo consiste en la tarea *10s-10s* de la evaluación *NIST 2010* [20] siguiendo el protocolo explicado en el capítulo 3.

Todos los experimentos llevados a cabo han sido realizados utilizando el software de ATVS en lenguajes *C++*, *Matlab* y *bash* realizando modificaciones para adaptarse a los requisitos de cada prueba.

4.2. Sistema de referencia

Como punto de partida de este análisis se ha evaluado el rendimiento de ambos sistemas, *Total Variability* y *Probabilistic Linear Discriminant Analysis*, *TV* y *PLDA* de ahora en adelante, utilizando *i-vectors* de dimensión 400. La tabla 4.1 muestra los resultados de ambos sistemas en función del número de Gaussianas utilizado para construir el modelo universal y, por tanto, para el extractor de *i-vectors*.

	# Gausianas				
Sistema	64	128	256	512	1024
TV-LDA	21.08/0.0820	21.40/0.0785	18.41/0.0710	17.53/0.0698	16.22/0.0687
PLDA	18.79/0.0766	17.27/0.0699	16.00/0.0674	16.22/0.647	**15.36/0.0614**

Tabla 4.1: *EER/DCF* para los sistemas *Total variability* y *Probabilistic Linear Discriminant Analysis* en función del número de Gaussianas utilizadas (*SRE10 10s-10s* masculino)

Como punto de referencia es interesante saber que el mismo sistema *PLDA* con 1024 Gaussianas tiene un error de 2.64/0.0149 de *EER* y *DCF* respectivamente en la condición 5 masculina de la evaluación *SRE10* en la que contamos con 2.5 minutos de conversación telefónica tanto para entrenar el modelo como para la realización del test.

A primera vista podemos realizar dos observaciones.

- En primer lugar, **el método PLDA obtiene mejores resultados que el método *LDA* seguido de *WCCN*** propuesto originalmente para separar la variabilidad de locutor y de sesión en el espacio de los *i-vectors*. Estos resultados coinciden, para locuciones de corta duración o en presencia de datos limitados, con las conclusiones extraídas en [48] y [63] que destacan la mayor capacidad de los marcos probabilísticos como los utilizados en el sistema *PLDA* para manejar la incertidumbre frente a los marcos no probabilísticos como los utilizados en el sistema *TV*.
- En segundo lugar, como puede observarse mejor en las figuras 4.1 y 4.2, **el incremento de número de Gaussianas se ve reflejado en un aumento del rendimiento** de los sistemas. El hecho de obtener un mejor rendimiento haciendo el sistema más pesado (de forma que hay muchos más parámetros que deben ser entrenados) en un problema que se basa en una pequeña cantidad de datos puede parecer contradictorio. Sin embargo, es necesario resaltar que la ventaja inherente a utilizar el marco de los *i-vectors* consiste en realizar la clasificación en un subespacio de bajas dimensiones sin importar el tamaño del extractor de *i-vectors*.

El hecho de llevar a cabo ésta etapa final en un subespacio de dimensión reducida sin importar el tamaño del extractor de *i-vectors* nos permite trabajar con sistemas muy pesados y robustos en la fase de desarrollo mientras que la etapa final de clasificación se realiza en un espacio de bajas dimensiones sin verse afectado el rendimiento final.

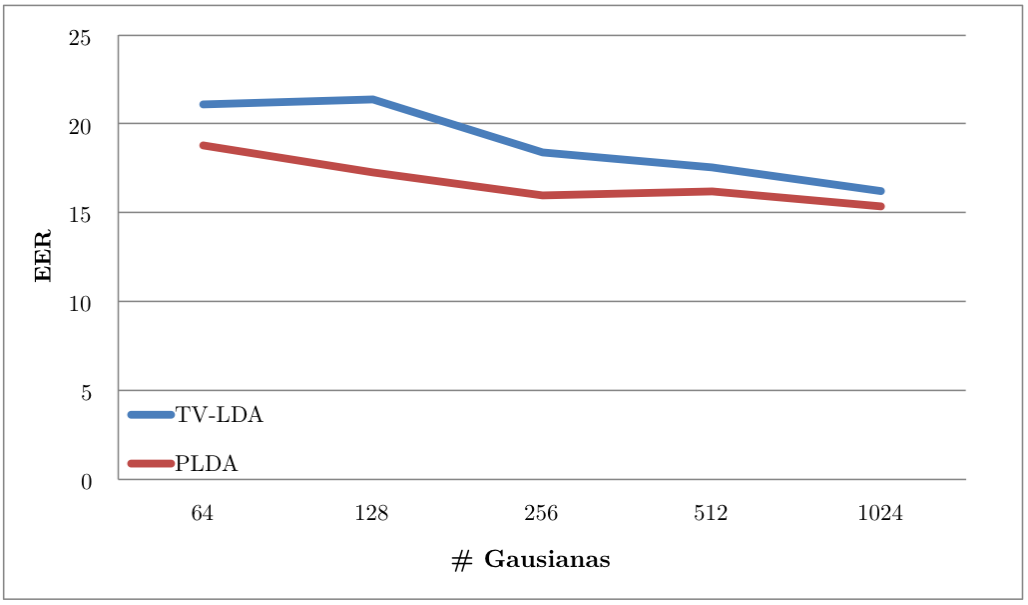

Figura 4.1: *EER* de los sistemas *TV* y *PLDA* en función del número de Gaussianas.

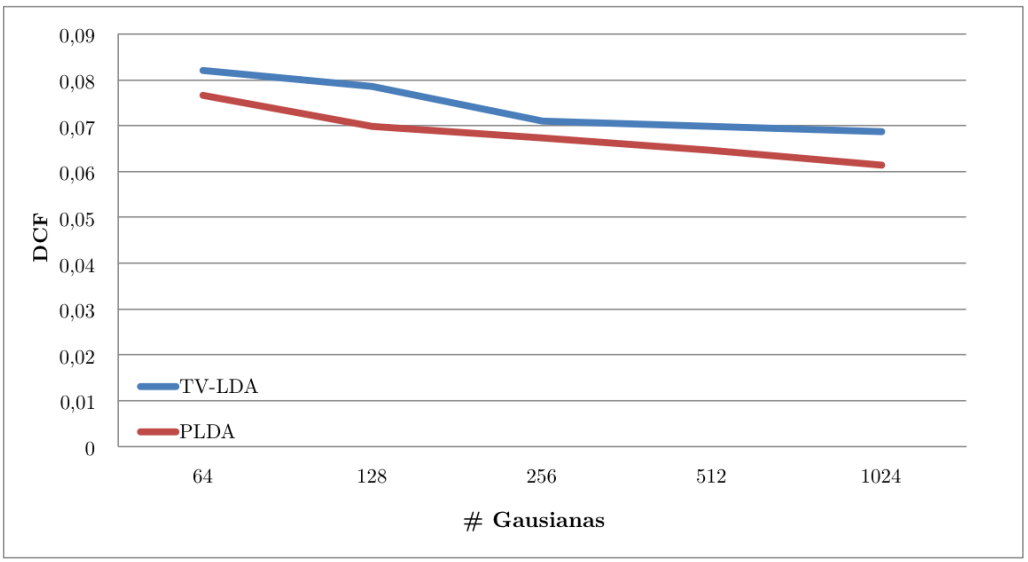

Figura 4.2: *DCF* de los sistemas *TV* y *PLDA* en función del número de Gaussianas.

4.3. Variación del tamaño del subespacio de variabilidad

Como hemos visto en las secciones 2.5.5 y 3.4 los sistemas basados en *Factor Analysis* utilizan técnicas que reducen la variabilidad a espacios de dimensiones reducidas. En el caso del sistema *Total Variability* la matriz de *LDA* transforma las características facilitando la separación entre locutor y variabilidad intersesión. En el caso del sistema *Probabilistic Linear Discriminant Analysis* es la matriz F la encargada de confinar la variabilidad de locución a un subespacio de baja dimensionalidad.

En estos experimentos vamos a variar el número de dimensiones desde 50 hasta el máximo tamaño de *i-vector* utilizado. Se obtendrá el rendimiento del sistema para cada tamaño de matriz y número de Gaussianas con que ha sido generado el modelo universal. De esta forma se pretenden buscar los parámetros óptimos así como analizar si la cantidad de componentes que debemos introducir depende de cuánta información extraiga el sistema en la fase de extracción de características.

Para evitar la duplicación de resultados las siguientes gráficas se mostrarán sólo en función del error *DCF* ya que es el error en que se mide la evaluación *NIST 2010* cuyo protocolo estamos siguiendo.

4.3.1. Variación del número de componentes de *LDA* utilizadas

En esta sección se presentan los resultados obtenidos variando el número de componentes de *LDA* utilizadas para diferentes tamaños de modelo universal, es decir, en función del número de Gaussianas utilizadas en el *UBM*.

En la figura 4.3 podemos ver cómo afecta la variación del tamaño de la matriz *LDA* cuando el sistema parte de un modelo universal que utiliza 64 mezclas. El generador de *i-vectors* de 64 mezclas es el que menos información conserva, por ese motivo podemos ver que la variación en el error es pequeña.

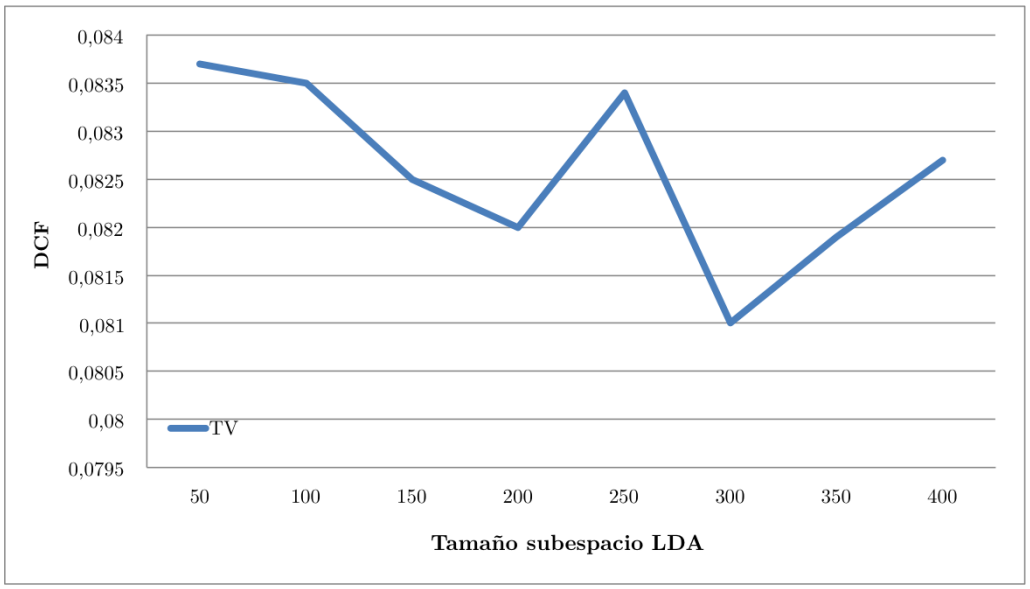

Figura 4.3: *DCF* del sistema *TV* generado a partir de 64 Gaussianas en función del tamaño de la matriz *LDA*.

En la figura 4.4 podemos ver el efecto de la misma variación en un sistema que parte de un *UBM* de 128 mezclas. Con 128 Gaussianas empezamos a apreciar una mejora del rendimiento al aumentar el número de componentes utilizadas en la matriz *LDA* hasta 150, a partir de este punto el error se mantiene aproximadamente constante salvo un aumento del mismo con tamaño 200 que no se repite en pruebas con otros generadores como veremos a continuación. Además como podemos ver en la misma Figura la variación en la escala *DCF* no es demasiado notable.

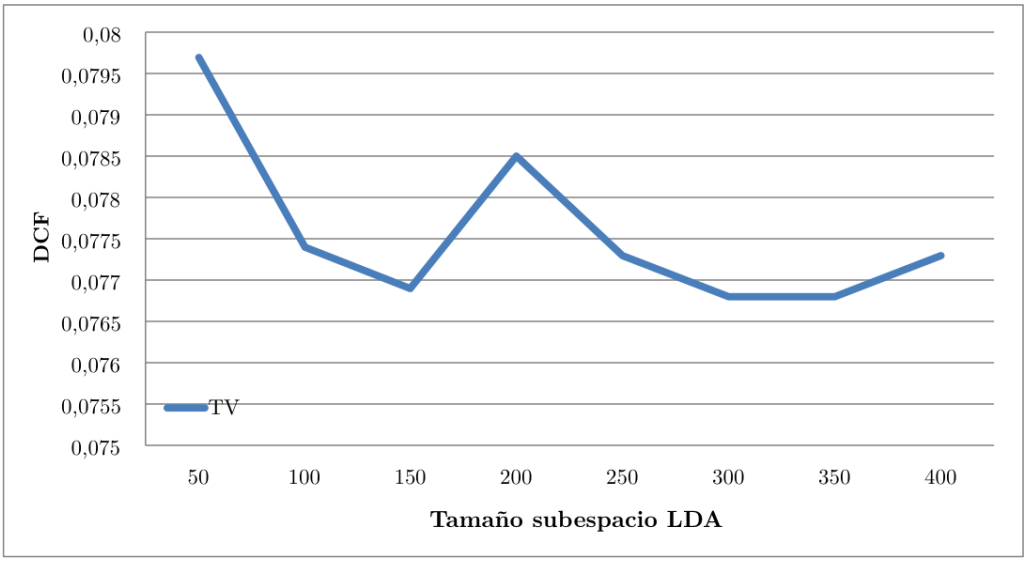

Figura 4.4: *DCF* del sistema *TV* generado a partir de 128 Gaussianas en función del tamaño de la matriz *LDA*.

En la figura 4.5 se presenta el mismo efecto para un sistema que utiliza un modelo universal *UBM* de 256 mezclas. En esta figura podemos ver cómo el rendimiento mejora hasta utilizar las 150 componentes más representativas y empeora a partir de este punto.

En las figuras 4.6 y 4.7 podemos ver cómo se ve afectado el rendimiento de los sistemas más robustos, generados con modelos de 512 y 1024 mezclas de Gaussianas. La tendencia que podemos ver en estas figuras nos indica que el rendimiento comienza mejorando a medida que aumentamos el tamaño de la matriz *LDA* para mantenerse aproximadamente constante a partir de un cierto punto. Éstas gráficas nos sugieren que las 150 primeras componentes contienen la mayor parte de información discriminatoria ya que a partir de este punto no se aprecia mejora significativa en el rendimiento. Además como podemos observar en ambas figuras con un mayor número de Gaussianas los sistemas son más estables desapareciendo los picos que podían observarse en figuras anteriores.

4.3.2. Variación del tamaño del subespacio de locutor en el sistema *PLDA* utilizadas

En esta sección se presentan los resultados obtenidos variando el tamaño de la matriz F del modelo *PLDA* para diferentes tamaños de modelo universal, es decir, en función del número de Gaussianas utilizadas en el *UBM*.

En la figura 4.8 podemos ver cómo afecta la variación del tamaño de la matriz F cuando el sistema parte de un modelo universal que utiliza 64 mezclas. Aun siendo el generador de *i-vectors* de 64 mezclas el que menos información conserva se aprecia una mejora del rendimiento hasta

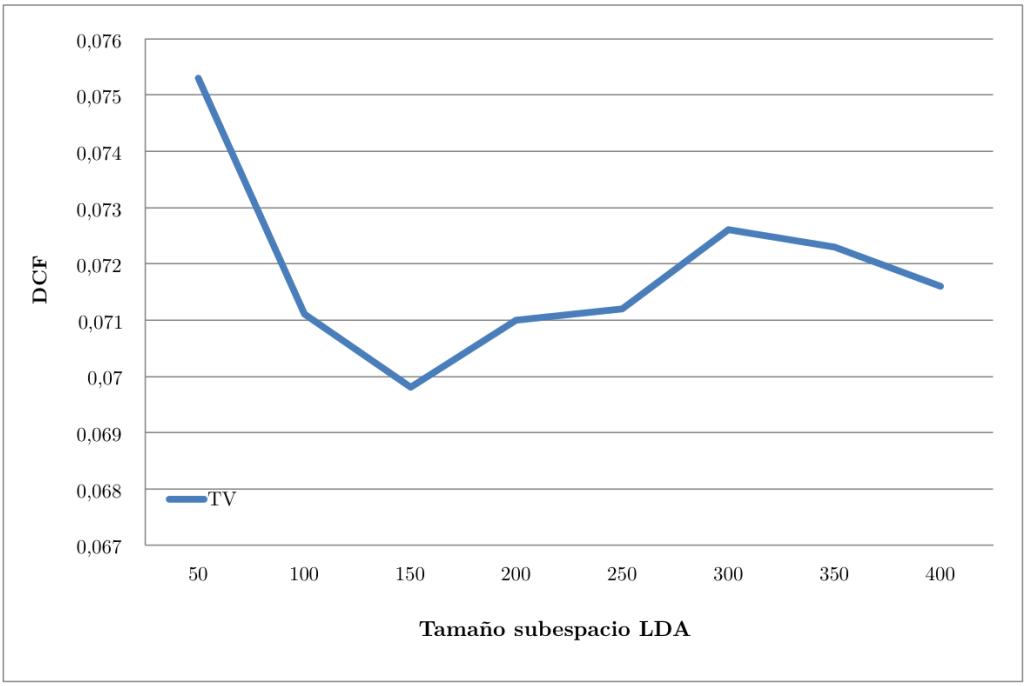

Figura 4.5: *DCF* del sistema *TV* generado a partir de 256 Gaussianas en función del tamaño de la matriz *LDA*.

aumentar el subespacio a una dimensión de 200. A partir de este punto el error se mantiene aproximadamente constante o empeora.

En las figuras 4.9 y 4.10, que se corresponden con los extractores de *i-vector* de robustez media, podemos apreciar una mejora del rendimiento del sistema *PLDA* a medida que aumentamos la dimensionalidad hasta tamaño 200. A partir de este punto el sistema no mejora su rendimiento e incluso empeora.

Las figuras 4.11 y 4.12 se corresponden con los sistemas *PLDA* cuyos modelos universales tienen 512 y 1024 mezclas respectivamente. Éstos son los sistemas más robustos, es decir, aquellos que utilizan un mayor número de mezclas en sus modelos universales por lo que recogen una mayor cantidad de información de la señal original. Se puede apreciar que el mínimo error se obtiene cuando se utiliza un subespacio de variabilidad para el locutor de dimension 200.

4.3.3. Análisis de la variabilidad del subespacio de locutor para ambos sistemas de forma conjunta

La figura 4.13 recopila los resultados de los sistemas más robustos y por tanto con mejor rendimiento tanto de la técnica *PLDA* como de la técnica *TV*. Podemos ver que mientras el error *DCF* en el sistema *TV* se mantiene aproximadamente constante a partir de 150 dimensiones, en el sistema *PLDA* encontramos el mínimo con 200 dimensiones y el sistema se degrada lentamente a partir de ese punto. Los resultados coinciden con los extraídos para locuciones de mayor duración [64] que indican que el tamaño óptimo para el subespacio de variabilidad de locutor es menor que el tamaño del subespacio de los *i-vector* utilizados, 400 en nuestro caso.

En las tablas 4.2 y 4.3 podemos encontrar con mayor detalle los resultados descritos anteriormente para ambos sistemas. Se presentan los valores de error tanto *DCF* como *EER*correspondientes

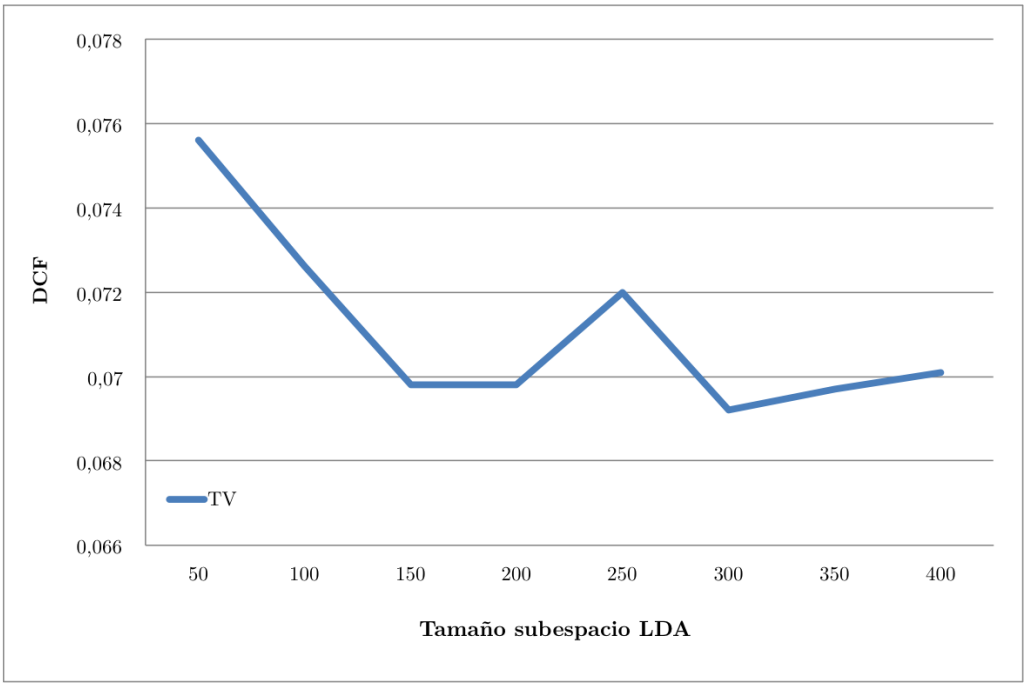

Figura 4.6: *DCF* del sistema *TV* generado a partir de 512 Gaussianas en función del tamaño de la matriz *LDA*.

a ambos sistemas en función del número de Gaussianas y el tamaño del subespacio de variabilidad de locutor, *LDA* en el sistema *TV* y *F* en el sistema *PLDA*.

	# Gausianas				
LDA dim.	64	128	256	512	1024
50	23.37/0.0837	22.23/0.0797	19.17/0.0753	18.96/0.0756	17.61/0.0731
100	22.61/0.0835	21.03/0.0774	18.89/0.0711	17.75/0.0726	16.12/0.0697
150	21.56/0.0825	21.08/0.0769	18.79/0.0698	17.52/0.0698	16.12/0.0688
200	21.08/0.820	21.40/0.0785	18.41/0.0710	17.53/0.0698	16.22/0.0687
250	21.47/0.0834	20.04/0.0773	18.26/0.0712	16.72/0.0720	16.31/0.0680
300	21.56/0.0810	20.05/0.0768	18.55/0.0726	17.27/0.0692	**16.12**/0.0677
350	21.47/0.0819	20.42/0.0768	18.70/0.0723	17.53/0.0697	16.22/**0.0673**
400	21.47/0.0827	20.27/0.0773	17.64/0.0716	17.23/0.0701	16.82/0.0675

Tabla 4.2: *EER/DCF* para el sistema *TV* en función del número de Gaussianas utilizadas y tamaño de *LDA* (*SRE10 10s-10s* masculino)

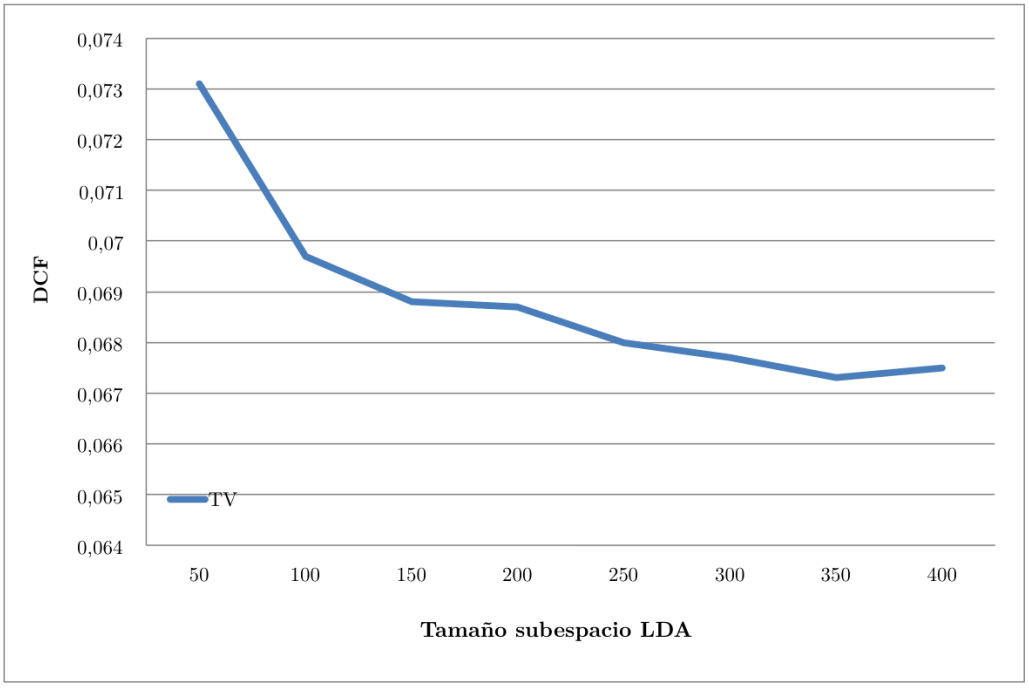

Figura 4.7: *DCF* del sistema *TV* generado a partir de 1024 Gaussianas en función del tamaño de la matriz *LDA*.

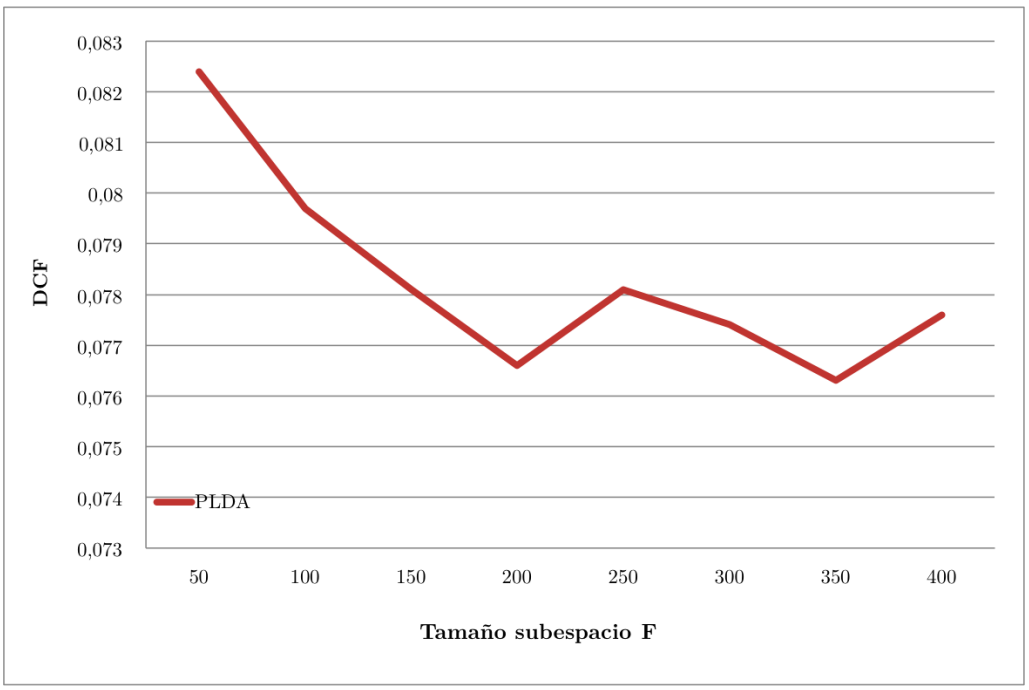

Figura 4.8: *DCF* del sistema *PLDA* generado a partir de 64 Gaussianas en función del tamaño de la matriz *F*.

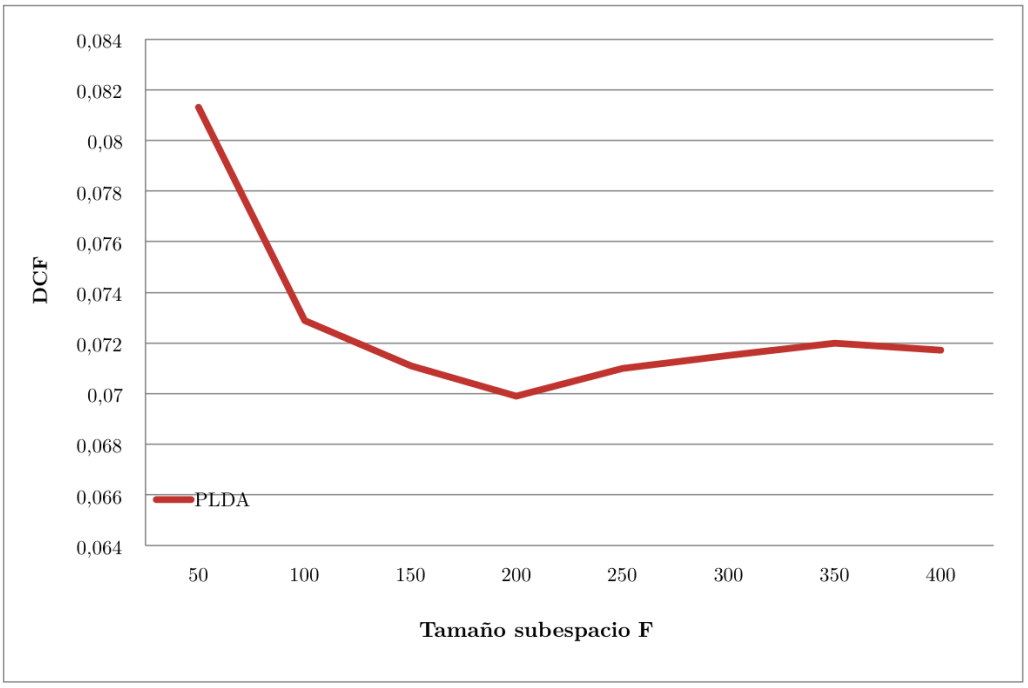

Figura 4.9: *DCF* del sistema *PLDA* generado a partir de 128 Gaussianas en función del tamaño de la matriz *F*.

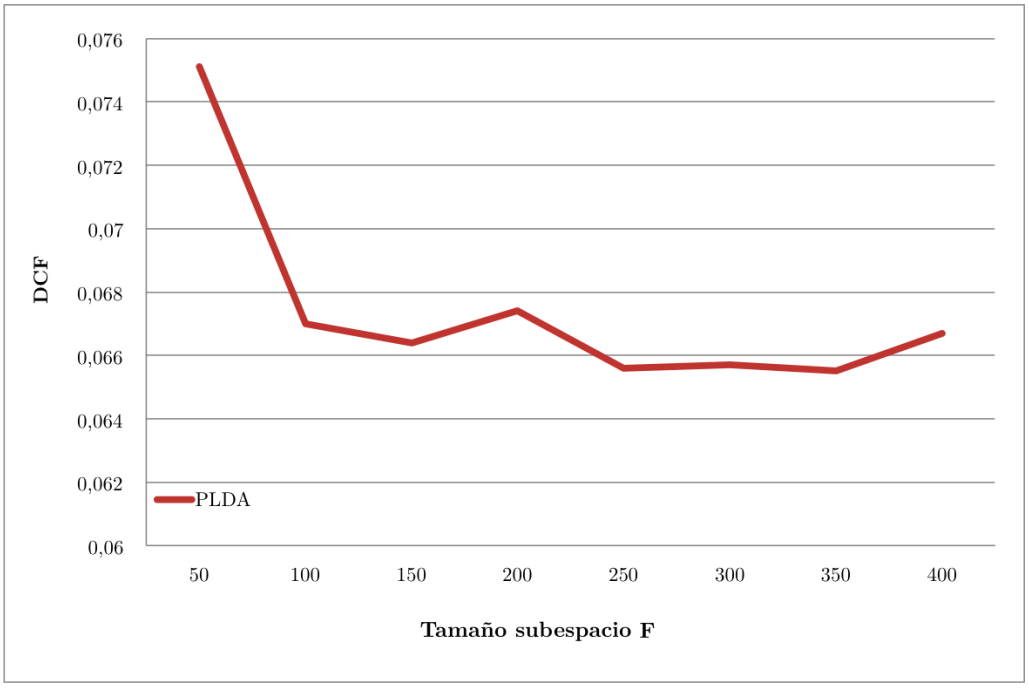

Figura 4.10: *DCF* del sistema *PLDA* generado a partir de 256 Gaussianas en función del tamaño de la matriz *F*.

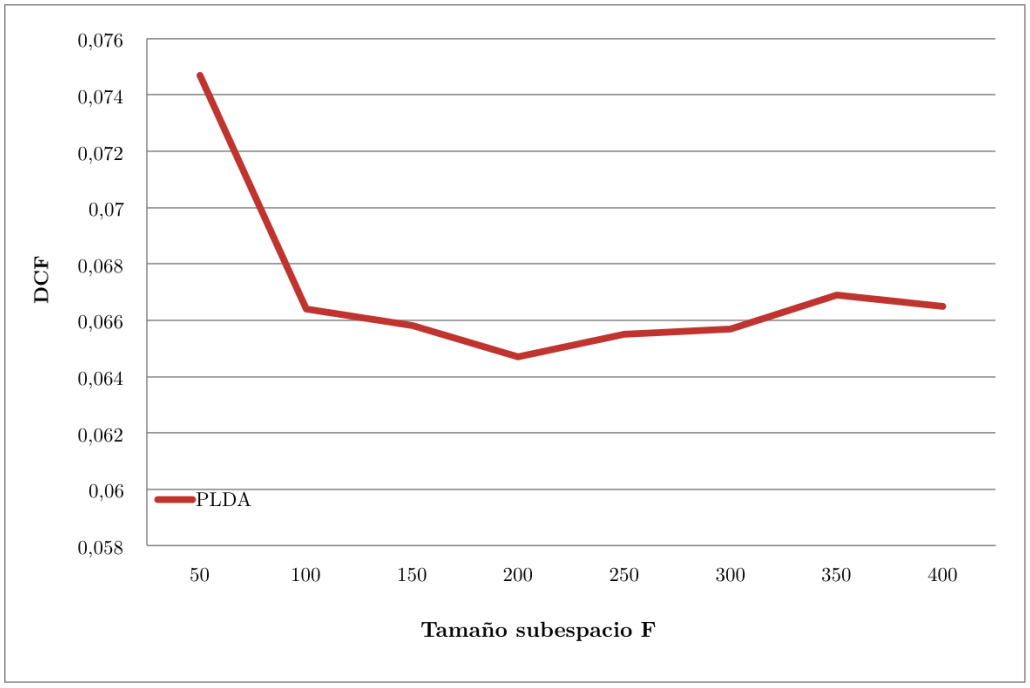

Figura 4.11: *DCF* del sistema *PLDA* generado a partir de 512 Gaussianas en función del tamaño de la matriz *F*.

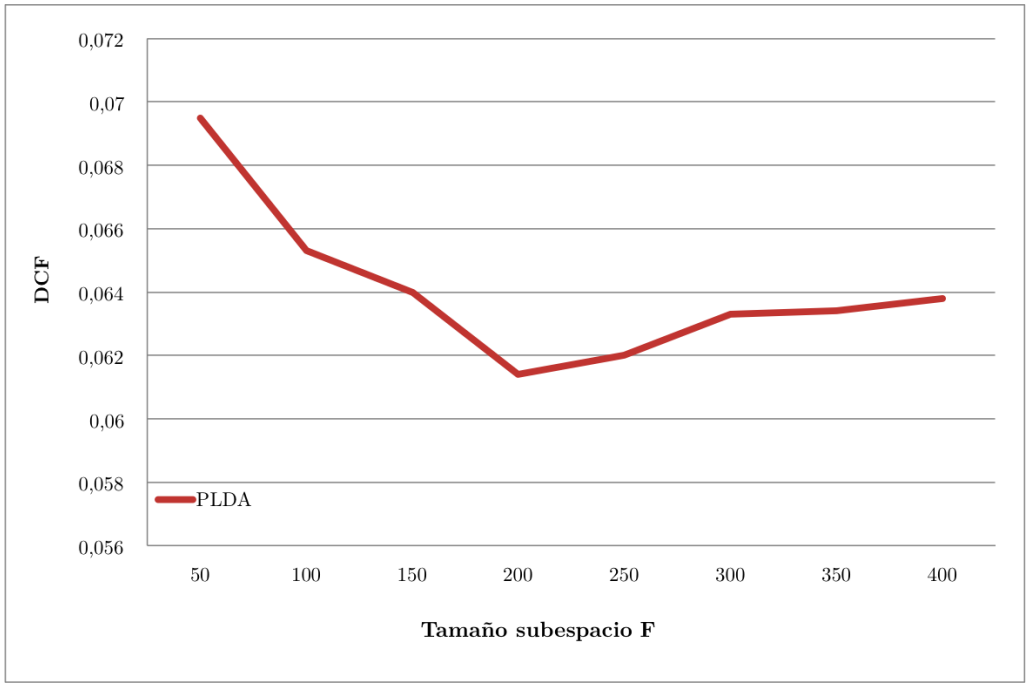

Figura 4.12: *DCF* del sistema *PLDA* generado a partir de 1024 Gaussianas en función del tamaño de la matriz *F*.

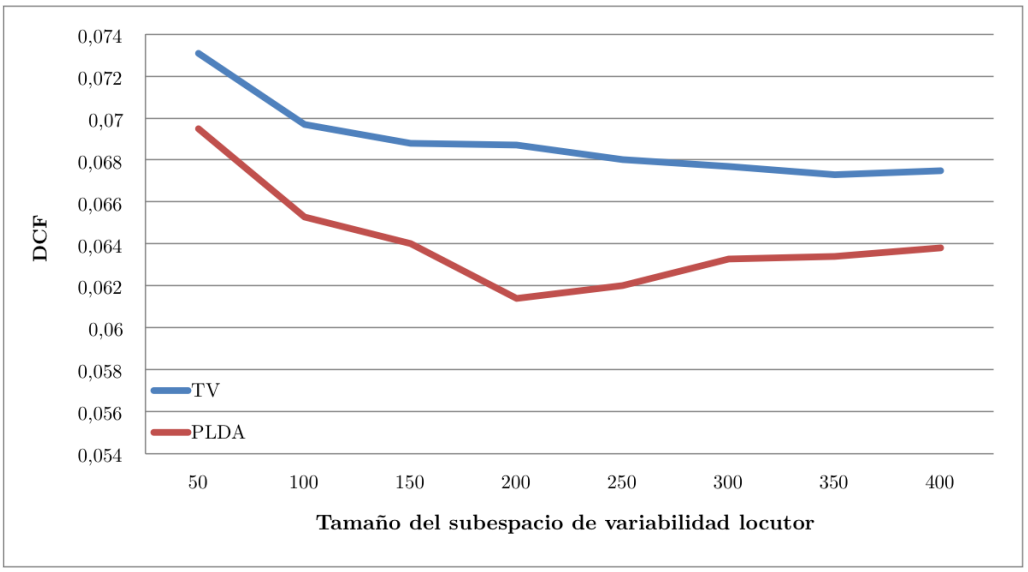

Figura 4.13: *DCF* de los sistemas *TV* y *PLDA* generados a partir de 1024 Gaussianas en función del tamaño del subespacio de variabilidad de locutor.

	# Gausianas				
***F* dim.**	64	128	256	512	1024
50	20,04/0.0824	17,27/0,0813	17,00/0,0751	16,50/0,0747	16,87/0.0695
100	19,17/0,0797	17,27/0,0729	16,89/0,0670	16,12/0,0664	15,36/0.0653
150	19,10/0,0781	16,89/0,0711	16,30/0,0664	16,12/0,0658	16,12/0.0640
200	18,79/0.0766	17.27/0.0699	16.00/0.0674	16.22/0.0647	15.36/**0.0614**
250	19,28/0,0781	16,50/0,0710	16,11/0,0656	16,50/0,0655	15,74/0.0620
300	19,28/0,0774	16,72/0,0715	**14,97**/0,0657	15,84/0,0657	15,74/0.0633
350	19,01/0,0763	16,50/0,0720	15,54/0,0655	16,12/0,0669	15,36/0.0634
400	18,41/0,0776	16,50/0,0717	15,84/0,0667	15,97/0,0665	15,26/0,0638

Tabla 4.3: *EER/DCF* para el sistema *PLDA* en función del número de Gaussianas utilizadas y tamaño de *LDA* (*SRE10 10s-10s* masculino)

4.4. Variación del tamaño del *i-vector*

El último experimento llevado a cabo pretende analizar cómo afecta el tamaño del *i-vector* al rendimiento de ambos sistemas. La figura 4.14 resumen los resultados obtenidos en términos de DCF al aumentar el tamaño del *i-vector* desde 50 a las normalmente utilizadas 400 dimensiones. Puede observarse que el mínimo se encuentra en una dimensión de 300 para el sistema *TV* mientras que la dimensión óptima para el sistema *PLDA* es 200.

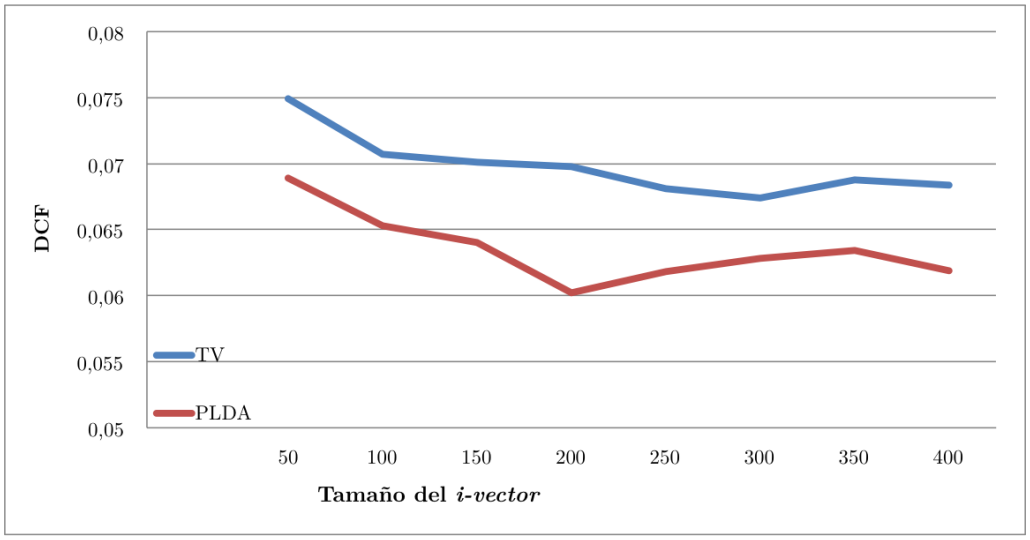

Figura 4.14: *DCF* para los sistemas *TV* y *PLDA* en función del tamaño del *i-vector* (sistemas generados con un *UBM* de 1024 Gaussianas)

En la tabla 4.4 pueden verse los resultados presentados en más detalle. Estos valores sugieren que en duraciones cortas la cantidad limitada de datos provoca que la variabilidad de locutor pueda confinarse en un *i-vector* de menor longitud mejorando el rendimiento global de los sistemas. Además cabe destacar el mejor rendimiento conseguido al utilizar *PLDA* frente a *TV* seguido de *LDA* y *WCCN* en entornos donde un *i-vector* de bajas dimensiones se ajusta mejor a las condiciones del problema. La mejora conseguida en nuestra tarea (*10s-10s SRE10*) es de un 14 por ciento al utilizar *PLDA* frente a *TV*.

	Dimension del *i-vector*							
Sistema	50	100	150	200	250	300	350	400
TV-LDA	0.0749	0.0707	0.0701	0.0698	0.0681	**0.0674**	0.0688	0.0684
PLDA	0.0698	0.0653	0.0640	**0.0602**	0.0618	0.0628	0.0634	0.0619

Tabla 4.4: *DCF* para los sistemas *TV* y *PLDA* en función del tamaño del *i-vector* (*SRE10 10s-10s* masculino)

5

Conclusiones y trabajo futuro

5.1. Conclusiones.

El presente Proyecto Final de Carrera se ha centrado en el desarrollo y análisis de los sistemas de reconocimiento automático de locutor independiente de texto utilizando en técnicas basadas en *Factor Analysis* en presencia de locuciones de corta duración. Para ello se ha llevado a cabo un amplio estudio del estado del arte de los sistemas acústicos de reconocimiento de locutor basados en el modelado *GMM-UBM* que han proporcionado históricamente los mejores rendimientos cuando la cantidad de datos disponible es mayor de forma que podamos entender la problemática de las duraciones cortas.

Mientras que técnicas basadas en *Factor Analysis* como *TV* seguido de *LDA* y *WCCN* o *PLDA* han demostrado paliar la variabilidad de forma extremadamente eficiente en presencia de una cantidad *razonable* de datos estos resultados se degradan rápidamente en presencia de locuciones de corta duración.

En este proyecto se han explorado las limitaciones de dichas técnicas así como su comportamiento en presencia de locuciones de corta duración y se han adaptado a nuestra tarea mediante un estudio del impacto que tienen diferentes parámetros de los algoritmos utilizados sobre el rendimiento final en una tarea marcada por la escasez de datos tanto para el entrenamiento del modelo como para la realización de los test.

De los análisis y experimentos llevados a cabo durante el proyecto podemos extraer las siguientes conclusiones generales:

- Los sistemas de reconocimiento de locutor que han marcado el estado del arte de los últimos años se ven significativamente degradados en presencia de locuciones de corta duración. En trabajos como [65] se ha demostrado que de forma cuantitativa la importancia de este problema, además se prueba que **este degradamiento no es lineal** lo que sugiere que podrían realizarse mayores avances para paliar este efecto.
- A pesar de la baja dimensionalidad inherente al espacio de los *i-vectors* y a la utilización de una cantidad muy limitada de datos **los sistemas se ven beneficiados de un generador**

de características más complejo o pesado. Esto nos permite trabajar con sistemas altamente robustos en la etapa de desarrollo pero llevar a cabo la etapa de clasificación en un subespacio de dimensiones mucho más reducidas sin un decremento del rendimiento global.

- Además de comportarse mejor en condiciones más favorables donde la cantidad de datos es ampliamente superior a la estudiada en este proyecto **los marcos probabilísticos como *PLDA* han demostrado una mayor capacidad para enfrentarse y manejar la incertidumbre proveniente de la escasez de datos frente a los métodos clásicos** como *LDA* y *WCCN*.

- La cantidad de datos con que contamos para la generación de los *i-vectors* es un factor muy influyente en el rendimiento de nuestros sistemas. A pesar de que el sistema *i-vector* estándar ya confina la variabilidad a un espacio de dimensiones muy reducidas **los sistemas se ven beneficiados de la utilización de *i-vectors* de longitud aún menor en presencia de locuciones de duración corta**.

- Por último, en la técnica *Total Variability* presentada en este proyecto, **los *i-vectors* se calculan como una estimación del punto *MAP* de la variable latente w**. Sin embargo, sabemos que la estimación de un punto en presencia de una cantidad de datos limitada puede derivar en el cálculo de *i-vectors* poco fiables. En la sección 5.2 se presenta una posible solución para este problema.

5.2. Trabajo futuro

A pesar de que el uso de los *i-vector* ofrece un rendimiento aceptable cuando se enfrenta a tareas relacionadas con locuciones de corta duración, especialmente cuando se utiliza la técnica de modelado *PLDA*, existen varios aspectos que deben ser estudiados y se proponen como diferentes líneas de trabajo futuro que pueden continuar y mejorar el estudio realizado en este proyecto:

- La realización de un estudio en mayor profundidad que analice las **diferencias existentes entre *i-vectors* generados a partir de locuciones de diferente duración** podría resultar de gran utilidad. Este estudio ampliaría nuestro conocimiento sobre el problema de las duraciones cortas y quizá nos guíe a un mayor entendimiento del impacto que la duración de las locuciones tiene en el reconocimiento automático de locutor independiente de texto.

- En nuestro sistema se han utilizado locuciones de corta duración para el entrenamiento del modelo y para la realización de los test pero el desarrollo del sistema así como la generación de los modelos universales y el entrenamiento de las matrices de variabilidad han sido llevados a cabo con locuciones de mayor duración. El uso de **locuciones de una duración similar a las condiciones de evaluación en la fase de desarrollo** podría ser otra línea de futuro prometedora.

- En referencia a la última conclusión extraída en la sección 5.1 existen **alternativas como el marco *Variational Bayes* utilizado previamente en *Joint Factor Analysis* [66]** que realizan una estimación Bayesiana completa teniendo en cuenta toda la distribución en lugar de una estimación del punto *MAP*. Esta forma de proceder podría ser más apropiada en escenarios donde la escasez de datos suponga un problema.

Glosario de acrónimos

- **ASR**: *Automatic Speech Recognition* (reconocimiento automático de habla).
- **CMN**: *Cepstral Mean Normalization* (normalización de la media cepstral). Técnica de compensación de los efectos del canal de transmisión sobre la señal de voz que se aplica en el dominio de los coeficientes cepstrales.
- **DCF**: *Detection Cost Function* (función de coste de detección). Función definida para la evaluación del rendimiento de los sistemas de reconocimiento de locutor
- **DCT**: *Discrete Cosine Transform* (transformada discreta del coseno). Función de transformación basada en la DFT que utiliza sólo números reales.
- **DFT**: *Discrete Fourier Transform* (transformada discreta de fourier). Función de transformación ampliamente empleada en el área de tratamiento de señal para analizar las frecuencias presentes en una señal muestreada.
- **DET**: *Detection Error Trade-off* (compensación por error de detección). La curva DET se utiliza para representar de forma gráfica el rendimiento de un sistema de reconocimiento biométrico para los distintos puntos de trabajo posibles.
- **EER**: *Equal Error Rate* (tasa de error igual). Tasa de error, en los sistemas de reconocimiento biométrico, en el punto de trabajo en que error de falsa aceptación y error de falso rechazo son iguales.
- **EM**: *Expectation Maximization*. Algoritmo iterativo que en reconocimiento de locución se utiliza con las siguientes dos etapas por iteración: en la primera etapa el algoritmo asigna muestras a diferentes Gaussianas y en la segunda se reestiman los parámetros de dichas Gaussianas.
- **FA**: *Factor Analysis* (análisis de factores). Técnica que en reconocimiento de locutor se emplea para el modelado explícito de la variabilidad inter-sesión e intra-sesión en el entrenamiento de los modelos de locutor.
- **FAR**: *False Acceptance Rate* (tasa de falsa aceptación). Porcentaje de errores, respecto al total de comparaciones realizadas, en los que se considera que el rasgo de test se corresponde con el patrón de referencia cuando en realidad corresponde a una identidad distinta.
- **FFT**: *Fourier Fast Transform* (transformada rápida de Fourier). Algoritmo para la implementación computacionalmente eficiente de la DFT
- **FRR**: *False Rejection Rate* (tasa de falso rechazo). Porcentaje de errores, respecto al total de comparaciones realizadas, en los que se considera que el rasgo de test se corresponde con el patrón de referencia cuando en realidad corresponde a una identidad distinta.

- **GMM**: *Gaussian Mixture Model* (modelo de mezcla de Gaussianas). Técnica para el modelado de la identidad de un sujeto por medio del ajuste de un conjunto de Gaussianas multivariadas a su distribución de características.
- **GMM-UBM**: Técnica de modelado basado en GMM pero entrenando un modelo universal UBM para la posterior adaptación del modelo de locutor vía adaptación MAP.
- **JFA**: *Joint Factor Analysis*. Técnica de compensación de variabilidad empleada para el modelado de la variabilidad intra-locutor como la debida al canal.
- **LDA**: *Linear Discriminant Analysis*. Método clásico utilizado en reconocimiento de patrones para encontrar una combinación lineal de características que favorezca la separación de clases.
- **MAP**: *Maximum a Posteriori*. Técnica empleada para la adaptación de modelos de locutor a partir de un UBM en los sistemas basados en GMM.
- **MFCC**: *Mel Frequency Cepstral Coefficients* (coeficientes cepstrales en escala de frecuencias Mel). Coeficientes para la representación del habla basados en la percepción auditiva humana.
- **NIST**: *National Institute of Standards and Technology* (Instituto Nacional de Estándares y Tecnología de los Estados Unidos de América).
- **PLDA**: *Probabilistic Linear Discriminant Model*. Técnica de compensación de variabilidad que separa la variabilidad de locutor de la variabilidad de canal a partir de los i-vector
- **RASTA**: *RelAtiveSpecTrAl* (espectral relativo). El filtrado RASTA es una técnica de compensación de los efectos del canal de transmisión sobre la señal de voz que se aplica en el dominio de los coeficientes cepstrales.
- **S-norm**: *Symmetric Normalization*. Técnica de normalización de puntuaciones utilizada en los sistemas i-vector.
- **Score**: Puntuación obtenida por un sistema de reconocimiento biométrico en la comparación entre un patrón de referencia y un rasgo biométrico de test.
- **SRE**: *Speaker Recognition Evaluation* (evaluación de reconocimiento de locutor). Serie de evaluaciones organizadas por el NIST para fomentar el avance en las técnicas de reconocimiento de locutor.
- **SVM**: *Support Vector Machine* (máquina de vectores soporte). Clasificador discriminativo empleado en reconocimiento de locutor independiente de texto.
- **T-norm**: *Test Normalization*. Técnica de normalización de puntuación en que se usa una cohorte de modelos de impostor.
- **Trial**: Juicio o comparación entre un rasgo de test y un patrón de referencia.
- **TV**: *Total Variability*. Técnica de compensación de variabilidad empleada para modelar toda la variabilidad contenida en una locución a un subespacio de dimensiones reducidas.
- **UBM**: *Universal Background Model* (modelo de fondo universal). GMM independiente de locutor utilizado para adaptar modelos de locutor vía MAP en sistemas de reconocimiento de locutor independiente de texto GMM-UBM.

- **VAD**: *Voice Activity Detector* (detector de actividad de voz). Herramienta que pretende detectar qué partes de una locución contienen voz.
- **VQ**: *Vector Quantization* (cuantificación vectorial). Técnica de compresión de datos utilizada como sistema de reconocimiento de locutor independiente de texto o para la reducción del conjunto de observaciones en sistemas dependientes de texto.
- **Within Class Covariance Normalization**: Técnica cuya finalidad es maximizar la variabilidad intra-locutor minimizando la variabilidad inter-locutor.
- **Z-norm**: *Zero Normalization*. Técnica de normalización de puntuación en que se usa una cohorte de puntuaciones de test.
- **ZT-Norm**: *Zero and Test Normalization*. Técnica de normalización de puntuaciones que combina una cohorte de puntuaciones de test con una cohorte de modelos de impostor.

Bibliografía

[1] Javier Gonzalez-Dominguez, Rubén Zazo, and Joaquin Gonzalez-Rodriguez. On the use of total variability and probabilistic linear discriminant analysis for speaker verification on short utterances. In Doroteo Torre Toledano, Alfonso Ortega Giménez, António J. S. Teixeira, Joaquín González Rodríguez, Luis A. Hernández Gómez, Rubén San-Segundo-Hernández, and Daniel Ramos-Castro, editors, *IberSPEECH*, volume 328 of *Communications in Computer and Information Science*, pages 11–19. Springer, 2012.

[2] Tomi Kinnunen and Haizhou Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Commun.*, 52(1):12–40, January 2010.

[3] G. Sosa y M. Rocamora E. López. Tratamiento de voz. `http://iie.fing.edu.uy/investigacion/grupos/gmm/audio/seminario/seminariosviejos/2003/charlas/charla1/voz8.htm`, 2003.

[4] Anil K. Jain, Arun Ross, and Salil Prabhakar. An introduction to biometric recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, 14:4–20, 2004.

[5] Anil K. Jain, Patrick Flynn, and Arun A. Ross. *Handbook of Biometrics*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[6] Douglas A. Reynolds. An overview of automatic speaker recognition technology. In *ICASSP*, pages 4072–4075. IEEE, 2002.

[7] Davide Maltoni, Dario Maio, Anil K. Jain, and Salil Prabhakar. *Handbook of Fingerprint Recognition*. Springer Publishing Company, Incorporated, 2nd edition, 2009.

[8] P. Rose. *Forensic Speaker Identification*. Taylor and Francis, London, 2002.

[9] J. Wolf. Efficient acoustic parameters for speaker recognition. In *The Journal of the Acoustical Society of America, 51(2)*, pages 2044–2055, 1972.

[10] Philip Carr. *Forensic Speaker Identification*. Blackwell Publishers, Malden, Mass., 1999.

[11] Javier Franco-Pedroso, Fernando Espinoza-Cuadros, and Joaquin Gonzalez-Rodriguez. Cepstral trajectories in linguistic units for text-independent speaker recognition. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 20–29. Springer, 2012.

[12] J. S. Bridle and M. D. Brown. An experimental automatic word recognition system. Technical report, Joint Speech Research Unit, Ruislip, England, 1974.

[13] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck. *Discrete-time Signal Processing (2Nd Ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1999.

[14] B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America*, 55(6):1304–1312, 1974.

[15] Hynek Hermansky and Nelson Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994.

[16] Jason Pelecanos and Sridha Sridharan. Feature warping for robust speaker verification. In *2001: A Speaker Odyssey - The Speaker Recognition Workshop*, pages 213–218, Crete, Greece, 2001. International Speech Communication Association (ISCA).

[17] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Qin Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and Bing Xiang. The SuperSID project: exploiting high-level information for high-accuracy speaker recognition. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 4, 2003.

[18] Niko Brümmer and Johan du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2):230–275, 2006.

[19] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The det curve in assessment of detection task performance. In *warning*, pages 1895–1898, 1997.

[20] National Institute of Standards and a. o. Technology. The nist year 2010 speaker recognition evaluation plan. `http://www.nist.gov/itl/iad/mig/upload/NISTSRE10evalplanr6.pdf`, 2010.

[21] K. P. Li and J. E. Porter. Normalizations and selection of speech segments for speaker recognition scoring. In *ICASSP*, pages 595–598, New York, NY, USA, 1988.

[22] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1):42–54, 2000.

[23] N. Brummer, L. Burget, J. H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim. Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006. *Trans. Audio, Speech and Lang. Proc.*, 15(7):2072–2084, September 2007.

[24] Ignacio Lopez-Moreno, Daniel Ramos, Joaquin Gonzalez-Rodriguez, and Doroteo Torre Toledano. Anchor-model fusion for language recognition. In *INTERSPEECH*, pages 727–730. ISCA, 2008.

[25] Julian Fiérrez-Aguilar, Javier Ortega-Garcia, Daniel Garcia-Romero, and Joaquin Gonzalez-Rodriguez. A comparative evaluation of fusion strategies for multimodal biometric verification. In *Audio-and Video-based Biometric Person Authentication*, pages 830–837. Springer, 2003.

[26] Niko Brümmer and Johan du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2):230–275, 2006.

[27] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1):19–41, 2000.

[28] Allen Gersho and Robert M Gray. *Vector quantization and signal compression*. Springer, 1992.

[29] D Burton. Text-dependent speaker verification using vector quantization source coding. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35(2):133–143, 1987.

[30] F. Soong, A. Rosenberg, L. Rabiner, and B.-H. Juang. A vector quantization approach to speaker recognition. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.*, volume 10, pages 387–390, Apr 1985.

[31] Jérôme Louradour and Khalid Daoudi. Svm speaker verification using a new sequence kernel. In *Proc. 13th European Conf. on Signal Processing (EUSIPCO 2005), Antalya, Turkey*, 2005.

[32] Tomi Kinnunen, Evgeny Karpov, and Pasi Franti. Real-time speaker identification and verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(1):277–288, 2006.

[33] Marie Roch. Gaussian-selection-based non-optimal search for speaker identification. *Speech communication*, 48(1):85–95, 2006.

[34] Juhani Saastamoinen, Evgeny Karpov, Ville Hautamä, Pasi Frä, et al. Accuracy of mfcc-based speaker recognition in series 60 device. *EURASIP Journal on Advances in Signal Processing*, 2005(17):2816–2827, 1900.

[35] Yoseph Linde, Andres Buzo, and Robert M Gray. An algorithm for vector quantizer design. *Communications, IEEE Transactions on*, 28(1):84–95, 1980.

[36] Douglas A Reynolds and Richard C Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on*, 3(1):72–83, 1995.

[37] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1):19–41, 2000.

[38] William M Campbell, Joseph P Campbell, Douglas A Reynolds, Elliot Singer, and Pedro A Torres-Carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech & Language*, 20(2):210–229, 2006.

[39] Lukas Burget, Pavel Matejka, Petr Schwarz, Ondrej Glembek, and Jan Cernocky. Analysis of feature extraction and channel compensation in a gmm speaker recognition system. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(7):1979–1986, 2007.

[40] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. A study of interspeaker variability in speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(5):980–988, July 2008.

[41] Robbie Vogt and Sridha Sridharan. Explicit modelling of session variability for speaker verification. *Computer Speech & Language*, 22(1):17–38, 2008.

[42] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[43] Joseph P Campbell, Hirotaka Nakasone, Christopher Cieri, David Miler, Kevin Walker, Alvin F Martin, and Mark A Przybocki. The mmsr bilingual and crosschannel corpora for speaker recognition research and evaluation. Technical report, DTIC Document, 2004.

[44] Nir Krause and Ran Gazit. Svm-based speaker classification in the gmm models space. In *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, pages 1–5. IEEE, 2006.

[45] Patrick Kenny. Joint factor analysis of speaker and session variability: Theory and algorithms. *CRIM, Montreal,(Report) CRIM-06/08-13*, 2005.

[46] Najim Dehak, Reda Dehak, Patrick Kenny, Niko Brümmer, Pierre Ouellet, and Pierre Dumouchel. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *INTERSPEECH*, volume 9, pages 1559–1562, 2009.

[47] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788–798, 2011.

[48] Patrick Kenny. Bayesian speaker verification with heavy tailed priors. In *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2010.

[49] Peng Li, Yun Fu, Umar Mohammed, James H Elder, and Simon JD Prince. Probabilistic models for inference about identity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1):144–157, 2012.

[50] Face and ocular challenge series. `http://www.nist.gov/itl/iad/ig/focs.cfm`.

[51] George R Doddington, Mark A Przybocki, Alvin F Martin, and Douglas A Reynolds. The nist speaker recognition evaluation–overview, methodology, systems, results, perspective. *Speech Communication*, 31(2):225–254, 2000.

[52] Daniel Ramos Castro. *Forensic evaluation of the evidence using automatic speaker recognition systems*. PhD thesis, Universidad autónoma de Madrid, 2007.

[53] John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE, 1992.

[54] David Graff, Alexandra Canavan, and George Zipperlen. Switchboard-2 phase 1. *Philadelphia: Linguistic Data Consortium, University of Pennsylvania*, 1998.

[55] David Graff, Kevin Walker, and D Millier. Switchboard cellular part 1 transcribed audio. *Linguistic Data Consortium*, 2001.

[56] Christopher Cieri, Walt Andrews, Joseph P Campbell, George Doddington, Jack Godfrey, Shudong Huang, Mark Liberman, Alvin Martin, Hirotaka Nakasone, Mark Przybocki, et al. The mixer and transcript reading corpora: Resources for multilingual, crosschannel speaker recognition research. In *International Conference on Language Resources and Evaluation (LREC)*, pages 22–28, 2006.

[57] Jingdong Chen, Jacob Benesty, Yiteng Huang, and Simon Doclo. New insights into the noise reduction wiener filter. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1218–1234, 2006.

[58] Qualcomm, icsi, ogi (qio) front-end software. `http://www.icsi.berkeley.edu/ftp/global/pub/speech/papers/quio/`.

[59] Sound exchange software. `sox.sourceforge.net/`.

[60] Andrew O Hatch, Sachin S Kajarekar, and Andreas Stolcke. Within-class covariance normalization for svm-based speaker recognition. In *INTERSPEECH*, 2006.

[61] Daniel Garcia-Romero and Carol Y Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Interspeech*, pages 249–252, 2011.

[62] Stephen Shum, Najim Dehak, Reda Dehak, and Jim Glass. Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification. In *Proc. Odyssey*, number 76-82, 2010.

[63] Ahilan Kanagasundaram, Robbie Vogt, David B Dean, Sridha Sridharan, and Michael W Mason. I-vector based speaker recognition on short utterances. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, pages 2341–2344. International Speech Communication Association (ISCA), 2011.

[64] Pavel Matejka, Ondrej Glembek, Fabio Castaldo, Md Jahangir Alam, Oldrich Plchot, Patrick Kenny, Lukas Burget, and Jan Cernocky. Full-covariance ubm and heavy-tailed plda in i-vector speaker verification. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4828–4831. IEEE, 2011.

[65] Ahilan Kanagasundaram, Robert J Vogt, David B Dean, and Sridha Sridharan. Plda based speaker recognition on short utterances. In *The Speaker and Language Recognition Workshop (Odyssey 2012)*. ISCA, 2012.

[66] Xianyu Zhao and Yuan Dong. Variational bayesian joint factor analysis models for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(3):1032–1042, 2012.

A

Presupuesto

1) Ejecución Material

- Compra de ordenador personal (Software incluido) 1000 €
- Material de oficina 150 €
- Total de ejecución material 1150 €

2) Gastos generales

- sobre Ejecución Material 180 €

3) Beneficio Industrial

- sobre Ejecución Material 70 €

4) Honorarios Proyecto

- 1500 horas a 15 €/ hora 22500 €

5) Material fungible

- Gastos de impresión 100 €
- Encuadernación 200 €

6) Subtotal del presupuesto

- Subtotal Presupuesto 25650 €

7) I.V.A. aplicable

- 21 % Subtotal Presupuesto 5386,5 €

8) Total presupuesto

- Total Presupuesto 31036,5 €

Madrid, Julio 2014

El Ingeniero Jefe de Proyecto

Fdo.: Rubén Zazo Candil

Ingeniero de Telecomunicación

B

Pliego de condiciones

Pliego de condiciones

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un *sistema de reconocimiento de locutor independiente de texto*. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales.

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad "Presupuesto de Ejecución de Contratаz anteriormente llamado "Presupuesto de Ejecución Material"que hoy designa otro concepto.

Condiciones particulares.

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.

C

Artículo publicado

Publicación en congreso internacional

A continuación se anexa el artículo mencionado en el resumen en que se publican los resultados presentes en este proyecto en un congreso de carácter internacional.

On the use of Total Variability and Probabilistic Linear Discriminant Analysis for Speaker Verification on Short Utterances

Javier Gonzalez-Dominguez, Ruben Zazo, and Joaquin Gonzalez-Rodriguez

Biometric Recognition Group (ATVS),
Escuela Politecnica Superior, Universidad Autonoma de Madrid
{javier.gonzalez}@uam.es
http://atvs.ii.uam.es

Abstract. This paper explores the use of state-of-the-art acoustic systems, namely Total Variability and Probabilistic Linear Discriminant Analysis for speaker verification on short utterances. While the recent advances in the field dealing with the session variability problem have proved to greatly outperform speaker verification systems on typical scenarios where a reasonable amount of speech is available, this performance rapidly degrades at the presence of limited data in both enrolment and verification stages. This paper studies the behaviour of TV and PLDA on those scenarios where a scarce amount of speech (∼10s) is available to train and testing a speaker identity. The analysis has been carried out on the well defined and standard 10s-10s task belonging to the NIST Speaker Recognition Evaluation 2010 (NIST SRE10) and it explores the multiple parameters, which define TV and PLDA in order to give some insight about their relevance in this specific scenario.

Keywords: i-vectors, Total variability, PLDA, short utterances

1 Introduction

The remarkable advances dealing with the session variability problem accomplished during last years, have led to highly reliable speaker verification systems at the presence of a reasonable amount of speech.

In this context, techniques based on Factor Analysis such as Joint Factor Analysis (JFA) [1] [2], Total Variability (TV) [3] or more recently Probabilistic Linear Discriminant Analysis (PLDA) [4] have demonstrated an outstanding behavior even when facing vast and challenging evaluation scenarios such as the NIST Speaker Recognition Evaluation, NIST SRE10 [5].

Unfortunately those excellent results rapidly degrade as long as the available amount of enrolment and verification speech decreases [6] [7]. This fact made critical the design and use of the speaker verification systems in real applications such as access control or forensics while penalizing its application in other everyday applications.

The purpose of this paper is to evaluate and analyse the state-of-the-art acoustic systems TV and PLDA on those scenarios where just a very limited amount of speech (∼10s) is available for both, enrolment and verification. This analysis is driven through the different design parameters of TV and PLDA, with the aim of discovering which of them have a greater impact dealing with scarce amount of data. This last point is of particular interest, as often, systems are presented adjusted to typical scenarios, overshadowing the actual relevance of the different design parameters in specific tasks.

The rest of this work is organized as follows. A description of the Total Variability and Probabilistic Linear Discriminant Analysis based system is given in Section 2. Section 3 is devoted to present the experimental set-up and obtained results. Finally, main conclusions and future work lines are summarized in Section 4 and Section 5 respectively.

2 Systems Description

2.1 Total variability

Total Variability [3] represents a step further on the use of Joint Factor Analysis [1] [2] where a single subspace is trained to jointly model both session and speaker variability. This subspace, the so-called *total variability* subspace, T, aims to constraint in a low dimensional space both the session and the speaker variability. Mathematically, this generative latent variable model can be formulated as

$$\overline{\mu}'_s = \overline{\mu}'_{UBM} + Tw \ . \tag{1}$$

where μ and μ_{UBM} are the speaker and UBM model supervector respectively, T is the total variability matrix and w are the the latent factors of the mode, also called *total vectors* or *i-vectors*.

Since T constrains all the variability, speaker and session, and it is shared for all the speakers models/excerpts, the i-vectors, w, can be considered enough to represent the set differences between one excerpt to each other. Now, the *disentangling* phase between the speaker information and non-desired information can be accomplished at the i-vectors domain. This phase is typically carried out via classical Linear Discriminant Analysis (LDA) and Within Class Covariance Normalization (WCCN) [8]. The use of those techniques is now guaranteed as the dimensional reduction performed allows obtaining a non-singular within-class covariance matrix. Hereafter, we refer the Total Variability system followed by the classical LDA and WCCN as simply Total Variability or TV.

Finally, in order to obtain an score, a straightforward cosine distance between the i-vector coming from the speaker modeling w_1 and a test excerpt i-vector w_2 is computed as

$$S_{w_1,w_2} = \frac{(A^t w_1)W^{-1}(A^t w_2)}{\sqrt{(A^t w_1)W^{-1}(A^t w_1)}\sqrt{(A^t w_2)W^{-1}(A^t w_2)}} \ . \tag{2}$$

where A is the LDA matrix and W is the within class covariance matrix corresponding to WCCN.

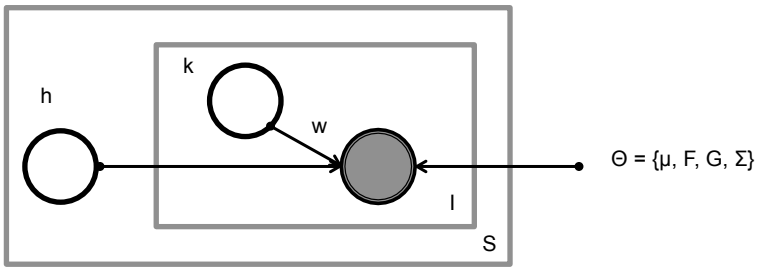

Fig. 1. Probabilistic Linear Discriminant Analysis graphical model representation for S speakers and I utterances. The observed variable w (i-vector), is explained through the identity latent factor h, the session variability hidden variable k and the set of hyperparameters Θ.

2.2 Probabilistic Linear Discriminant Analysis

As stated in the above section, the total variability framework has the main advantage of reducing a given speech utterance to a low-dimensional fixed length representation: the i-vector. From this point, i-vectors can be directly used for classification opening the door to classical methods such as Linear Discriminant Analysis (LDA) to accomplish the disentangling phase between speaker and session variability.

Probabilistic Linear Discriminant analysis (PLDA) is a generative latent variable model that has been recently used to successfully modelling i-vectors [4]. PLDA can be seen as a probabilistic version of classical LDA [9], where a specific i-vector i of a given speaker s is assumed to be decomposed as

$$w_{si} = \mu + Fh_s + Gk_i + \epsilon_i \ . \tag{3}$$

where F and G represents the new speaker and session variability subspaces respectively, h_s and k_i their respective latent variables associated and ϵ_i is a residual noisy term assumed to be normal distributed with zero mean and diagonal covariance matrix Σ. Figure 1 shows the PLDA probabilistic graphical model.

From above equation 3 the analogy between classical stated JFA and PLDA modelling approaches turns out evident. Nonetheless, two mayor important differences, in the context of speaker verification, must be taken into account

- JFA acts over speaker supervectors (high-dimensionality) while PLDA acts over i-vectors (low-dimensionality).
- JFA assumes speaker supervectors as generated by a mixture of multivariate Gaussians, while PLDA assumes i-vectors generated by a single multivariate Gaussian.

Following the PLDA model the similarity measure or score S_{w_1,w_2} between two given i-vectors, w_1 and w_2, can be computed as the ratio of the two alternative hyphotesis: H_0, both w_1 and w_2 belongs to a same identity (same h_s) and

H_1, w_1 and w_2 belongs to different identities (different h_s). This ratio can be expressed as

$$S_{w_1,w_2} = \frac{p(w_1, w_2|H_0)}{p(w_1|H_1)p(w_2|H_1)} = \frac{\int p(w_1, w_2|h)p(h)dh}{\int p(w_1|h_1)p(h_1)dh_1 \int p(w_2|h_2)p(h_2)dh_2} \ . \quad (4)$$

Assuming Gaussian priors for the latent variables in the model, it can be seen that integrals involved in above equation 4 turn out tractable and therefore the score, S_{w_1,w_2}, can be easily derived in a closed-form solution. Further details can be found in [9] [10].

3 Experiments

3.1 Experimental Setup

Experiments has been carried out on the telephone male part of the *10s-10s* NIST SRE10 task, where just ∼10s over a telephone channel are provided for both enrolment and verification stages. Specifically, a total number of 10858 trials has been evaluated belonging from 264 and 290 different models and tests segments respectively. The performance was assessed following the NIST SRE10 protocol [11] and results are presented in function of the Equal Error Rate (EER) and the minimum decision cost function (DCF).

Development data for training different Universal Background Models (UBMs) and system hyperparameters belonging to SWBI, SWBII and past NIST SRE evaluations (SRE04, SRE06 and SRE08). Utterances used with this purpose belongs to 1conv/short2 SRE tasks and therefore contains around 2.5m of speech. Specifically a total number of 5638 files from 823 speakers were used to train T, F and Σ [1]; LDA matrix was trained via 5214 files from 611 speakers while 4705 files belonging to 466 speakers were used to estimate the corresponding within class covariance matrix of WCCN method.

Symmetric score normalization (SNorm) [12] was used to finally normalize raw scores generated from the systems. A cohort of a 1000 files from the same development dataset was used for this purpose. Also, the length normalization method proposed in [10] was applied before PLDA modelling.

Regarding the feature extraction configuration, it consists of 38 MFCC coefficients ($19 + \Delta$) extracted by using a sliding Hamming window of 20ms and a 50% of overlapping. MEL filters were scaled between 300 and 3000Hz to focus as much as possible to speech voice.

[1] Given the intrinsic low-dimensionality of the i-vectors and the amount of speech for training the PLDA model available, we opted by grouping the noisy terms in equation 3 into a full-covariance matrix Σ.

	# Gaussians				
System	64	128	256	512	1024
TV-LDA	21.08/0.8202	21.40/0.0785	18.41/0.0710	17.53/0.0698	16.22/0.0687
PLDA	18.79/0.0766	17.27/0.0699	16.00/0.0674	16.22/0.0647	**15.36/0.0614**

Table 1. *EER/DCF for Total variability and Probabilistic Linear Discriminant Analysis systems depending on the number of Gaussians used (male SRE10 10s-10s).*

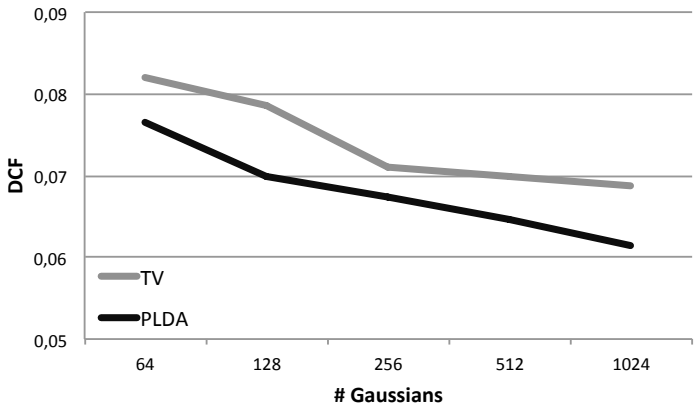

Fig. 2. DCF of TV and PLDA systems as a function of the number of Gaussians.

3.2 Results

As the starting point of this analysis, the performance obtained for both TV and PLDA systems was evaluated. Table 1 shows the results of both systems in function of the number of Gaussians used to build the Universal Background Model, and therefore of the i-vector extractor [2]. At a first glance, two observations can be done. First, PLDA method outperforms the LDA followed by WCCN method proposed originally to separate speaker and session variability on i-vectors. This result reinforces, on short utterances, the conclusions extracted in [4] [7], and highlights the mayor ability of the probabilistic framework followed in PLDA to manage uncertainty versus non-probabilistic frameworks. Second, as it can be better observed in Figure 2, increasing the number of Gaussians used in the i-vector extractor turns out in performance gains. The fact of obtaining better performance by doing the system heavier (much more free parameters to be trained) beside the nature of the faced problem where just an small amount of speech is available could seem contradictory. However, note that the inherent advantage of using the i-vector framework is that finally, regardless the *size* of

[2] As a reference performance on longer utterances, the same 1024 Gaussians PLDA system achieves a 2.64/0.0149 of EER and DFC on SRE10 task condition 5 (male part only).

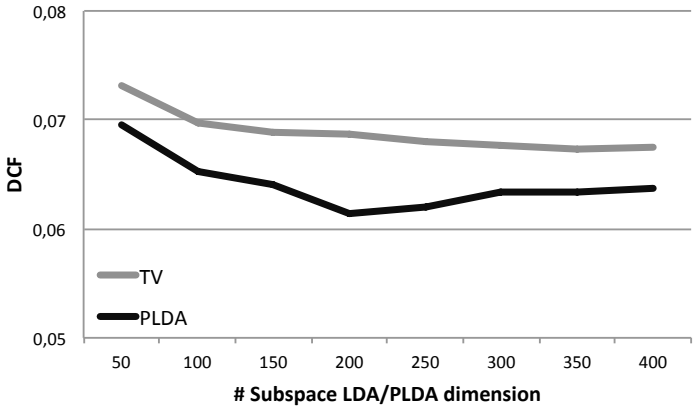

Fig. 3. DCF of TV and PLDA systems as a function of the number of kept dimensions in the LDA supspace and sepaker variability subspace F respectively.

the i-vector extractor, classification is done in a low-dimensional space. This last point allows to work with *heavier* and more robust systems at the development time to finally performing classification in a much lower-dimensional space without suffering a performance degradation.

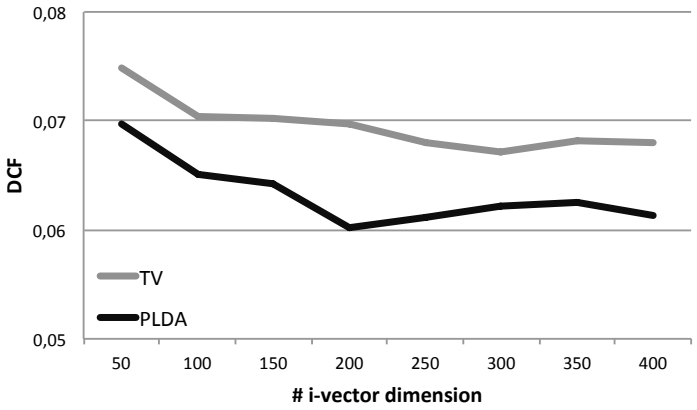

Fig. 4. DCF of TV and PLDA systems as a function of the i-vector dimension.

Another aspect explored in this study was the relevance of the LDA and the speaker variability subspace F dimensions in TV and PLDA systems respectively. Figure 3 shows a comparative of both systems by moving those dimensions

	# Gaussians				
LDA dim.	64	128	256	512	1024
50	23.37/0.0837	22.23/0.0797	19.17/0.0753	18.96/0.0756	17.61/0.0731
100	22.61/0.0835	21.03/0.0774	18.89/0.0711	17.75/0.0726	16.12/0.0697
150	21.56/0.0825	21.08/0.0769	18.79/0.0698	17.52/0.0698	16.12/0.0688
200	21.08/0.820	21.40/0.0785	18.41/0.0710	17.53/0.0698	16.22/0.0687
250	21.47/0.0834	20.04/0.0773	18.26/0.0712	16.72/0.0720	16.31/0.0680
300	21.56/0.0810	20.05/0.0768	18.55/0.0726	17.27/0.0692	**16.12**/0.0677
350	21.47/0.0819	20.42/0.0768	18.70/0.0723	17.53/0.0697	16.22/**0.0673**
400	21.47/0.0827	20.27/0.0773	17.64/0.0716	17.23/0.0701	16.82/0.0675

Table 2. *EER/DCF for Total Variability system in function of the number of Gaussians and LDA dimensions (male SRE10 10s-10s).*

	# Gaussians				
F **dim.**	64	128	256	512	1024
50	20,04/0.0824	17,27/0,0813	17,00/0,0751	16,50/0,0747	16,87/0.0695
100	19,17/0,0797	17,27/0,0729	16,89/0,0670	16,12/0,0664	15,36/0.0653
150	19,10/0,0781	16,89/0,0711	16,30/0,0664	16,12/0,0658	16,12/0.0640
200	18,79/0.0766	17.27/0.0699	16.00/0.0674	16.22/0.0647	15.36/**0.0614**
250	19,28/0,0781	16,50/0,0710	16,11/0,0656	16,50/0,0655	15,74/0.0620
300	19,28/0,0774	16,72/0,0715	**14,97**/0,0657	15,84/0,0657	15,74/0.0633
350	19,01/0,0763	16,50/0,0720	15,54/0,0655	16,12/0,0669	15,36/0.0634
400	18,41/0,0776	16,50/0,0717	15,84/0,0667	15,97/0,0665	15,26/0,0638

Table 3. *EER/DCF for Probabilistic Linear Discriminant Analysis system in function of the number of Gaussians and F subspace dimension (male SRE10 10s-10s).*

from 50 to the maximum i-vector dimension used, 400. Here, it can be seen that while the DCF in LDA kept mostly constant from 150 dimensions, PLDA find the minimum DCF at 200 dimension to slightly degrade when using higher dimensions. This result confirms, on short utterances, the studies performed for longer durations, where the optimum size of the PLDA speaker variability subspace use to be lower than the i-vector space [13]. Tables 2 and 3 complete in detail the above described results for both systems, exploring different number of Gaussians and LDA, F subspaces dimensions.

Finally the i-vector dimension used in both systems was also analysed. Figure 4 summarizes the results obtained in terms of DCF by increasing the i-vector dimension from 50 to the standard 400 dimensions. As it can be observed, a minimum at the 300 and 200 dimensions is found for the TV and PLDA systems respectively. This results suggest again that for the short utterances problem i-vector size under 400 dimensions might fit better the problem. Moreover, it encourages the use of PLDA rather than TV folllowed by LDA and WCCN when using lower dimensional of the i-vector; note that a relative improvement of 14% in DCF is achieved by using PLDA instead of TV.

4 Conclusions

A wide analysis on the use of state-of-the art acoustic approaches for speaker verification on short utterances has been carried out in this work. While Total Variability and Probabilistic Linear Discriminant Analysis methods have demonstrated to achieve outstanding results at the presence of a *reasonable* amount of data, this performance rapidly decrease when just short utterances are available for both enrolment and verification stages. This work has explored the limits of those systems when dealing with short durations. To this aim a leave-one-out analysis of the main configuration parameters of TV and PLDA system has been performed. On one hand, results show that due to the final low-dimensionality dimension of the i-vector, systems designed with complex or *heavy* i-vector extractors (high number of Gaussians, i-vector dimension) are able to obtain gains over lighter ones. On the other hand, the probabilistic framework followed by PLDA has demonstrated to better manage the implicit uncertainty of the task than the classical LDA and WCCN methods.

5 Future Work

Although the use of the i-vector framework achieves acceptable results on the challenging short utterances problem, specially by using PLDA as a modelling technique, some aspects should be explored. On one hand, a deep analysis of the differences among i-vectors extracted from different utterances durations has to be carried out. This study could give some insight, as well as turn out into a better treatment, of the duration utterance effect in speaker verification. Using development short utterances similar to the evaluation conditions, as performed in Joint Factor Analysis [6], could be a possible line in this context.

On the other hand, note that into the Total Variability framework presented, i-vectors are computed as MAP point estimates of the latent factor w. However, it is well known that the use of limited amount of data in order to get point estimates could derive in non reliable i-vectors. In this sense, alternatives as fully Bayesian frameworks as used in [14] for Joint Factor Analysis could be a more appropriate way of facing the short durations problem.

Acknowledgments. This research has been supported by the Ministerio de Ciencia e Innovacion under the proyect TEC2009-14719-C02-01 and Catedra UAM-Telefonica.

References

1. P. Kenny, G. Boulianne, P. Oullet, and P. Dumouchel, "Speaker and Session Variability in GMM-Based Speaker Verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.
2. R. Vogt and S. Sridharan, "Explicit Modeling of Session Variability for Speaker Verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.

3. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788 – 798, February 2011.
4. P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Odyssey: The Speaker and Language Recognition Workshop, Brno, Czech Republic,*, June 28 - July 1 2010.
5. N. Scheffer, L. Ferrer, M. Graciarena, S. S. Kajarekar, E. Shriberg, and A. Stolcke, "The SRI NIST 2010 Speaker Recognition Evaluation System," in *ICASSP*, 2011, pp. 5292–5295.
6. R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *INTERSPEECH*, 2008, pp. 853–856.
7. A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "I-Vector Based Speaker Recognition on Short Utterances," in *Interspeech 2011*. Firenze Fiera, Florence: International Speech Communication Association (ISCA), August 2011, pp. 2341–2344. [Online]. Available: http://eprints.qut.edu.au/46313/
8. A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *INTERSPEECH*, 2006.
9. S. Prince, P. Li, Y. Fu, U. Mohammed, and J. H. Elder, "Probabilistic models for inference about identity." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 144–157, 2012. [Online]. Available: http://dblp.uni-trier.de/db/journals/pami/pami34.html#PrinceLFME12
10. D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-Vector Length Normalization in Speaker Recognition Systems," in *INTERSPEECH*, 2011, pp. 249–252.
11. National Institute of Standards and a. o. Technology, "The NIST Year 2010 Speaker Recognition Evaluation Plan," http://www.nist.gov/itl/iad/mig/upload/NIST_SRE10_evalplanr6.pdf, 2010.
12. S. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," in *Odyssey: The Speaker and Language Recognition Workshop, Brno, Czech Republic,*, 2010.
13. P. Matejka, O. Glembek, F. Castaldo, M. J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocký, "Full-Covariance UBM and Heavy-Tailed PLDA in I-Vector Speaker Verification." in *ICASSP*. IEEE, 2011, pp. 4828–4831. [Online]. Available: http://dblp.uni-trier.de/db/conf/icassp/icassp2011.html#MatejkaGCAPKBC11
14. X. Zhao and Y. Dong, "Variational bayesian joint factor analysis models for speaker verification," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 3, pp. 1032–1042, 2012.